TI 2020-029/I
Tinbergen Institute Discussion Paper

Lie detection: A strategic analysis of the Verifiability Approach

Konstantinos Ioannidis[1,2]
Theo Offerman[1,2]
Randolph Sloof[2]

[1] CREED
[2] University of Amsterdam and Tinbergen Institute

Lie detection: A strategic analysis of the Verifiability Approach[*]

Konstantinos Ioannidis [†], Theo Offerman [‡], Randolph Sloof[§]

May 29, 2020

Abstract

The Verifiability Approach is a lie detection method based on the insight that truth-tellers provide precise details whereas liars sometimes remain vague to avoid being exposed. We provide a-game-theoretic analysis of a speaker who wants to be acquitted and an investigator who prefers to find out the truth. The investigator can verify the speakers statement at some cost; verification gets more reliable the more details are provided. If, after a falsified statement, the investigator convicts, an additional obstruction penalty is imposed. We derive all the equilibria of the game and thereby the conditions under which the investigator can infer additional information from the speaker's statement at face value. Strategic information revelation by the speaker and verification by the investigator then necessarily work in tandem. Improvements in reliability result in more valuable (strategic) information transmission, whereas a harsher obstruction penalty does not as soon as a lower limit is met.

Keywords: Lie detection, Verifiability approach, Strategic information revelation

JEL Codes: C72, D01, D82, K14

[*] The authors would like to thank participants at the 2018 CeDEx-CBESS-CREED meeting in Norwich, UK, the 2019 Annual Conference of the European Association of Law and Economics in Tel Aviv, Israel and the 2019 Annual Conference of the European Association of Psychology and Law in Santiago de Compostela, Spain for their valuable comments on the paper.

[†] CREED, University of Amsterdam and Tinbergen Institute. E-mail: ioannidis.a.konstantinos@gmail.com

[‡] CREED, University of Amsterdam and Tinbergen Institute. E-mail: t.j.s.offerman@uva.nl

[§] University of Amsterdam and Tinbergen Institute. E-mail: r.sloof@uva.nl

1 Introduction

After decades of research on lying detection, psychologists have recently made a breakthrough in revealing who is lying. The early literature focused on the idea that liars can be identified by facial microexpressions of emotions and other unintentional behaviors. In two meta-analyses, DePaulo et al. (2003) and Bond Jr and DePaulo (2006) showed that nonverbal cues of lying are weak and unreliable. A typical finding is that approximately 54% of examiners' judgments are correct, only slightly better than chance (50%).

One important reason why non-verbal cues are unreliable is that liars try to mimic the expressions of truth tellers when they become aware of which cues are used by investigators. For example, Ekman et al. (1988) have shown that truth tellers often smile when they express genuine positive feelings and that liars mimic them by also smiling. The challenge that examiners then face is that they have to distinguish between fake and genuine smiles.

The breakthrough involves recent methods of lie detection which focus on the content of what is being said. In the Verifiability Approach (VA), the examiner judges a statement based on the presence and frequency of verifiable details. VA exploits a dilemma that liars face. Liars have an incentive to include verifiable details in their statement, because detailed accounts are more likely to be believed (Bell and Loftus, 1989). At the same time, presenting specific details is risky because it makes it easier for the examiner to check a statement (Nahari et al., 2014). Truth-tellers typically do not have this dilemma and can reveal as many verifiable details as possible. The relative frequency of verifiable details in a statement may then become an informative signal of its truth. Using VA, examiners' judgments are correct in approximately 70% of the cases (Vrij, 2018).[1] Moreover, in contrast to the nonverbal cues, the accuracy of VA is enhanced when interviewees are made aware of it. Doing so results in truth tellers adding more verifiable details to their statement than liars do (Harvey et al., 2017; Nahari et al., 2014).

So far a game theoretic analysis of VA is missing. In this paper, we analyze the strategic interaction between a speaker who wants to convince an investigator that he is innocent and an investigator who pursues the truth. Applications of this type of interaction abound. A mother may want to find out if her son is using drugs; a parole officer is interested to know if an offender lives up to the agreement made; an airport officer wants to find out if a passenger is carrying dangerous items; an insurance company wants to find out whether a claim was rightly made; an employer interviews an applicant (and potentially verifies references) to learn whether he has been thorough and truthful in drafting his CV; a judge questions a suspect to assess whether he is guilty. In this paper, we use labels that correspond to the judge-suspect example for ease of illustration. A suspect is privately informed about whether he is guilty or innocent. The judge has already collected some evidence that furnishes a prior belief about whether the suspect is guilty. The suspect is asked to make a statement about what happened. He either makes a precise statement that includes verifiable details, or a vague statement. After listening to the suspect, the judge can decide to reach a verdict immediately or to check the statement at some cost. Examination of the statement yields informative but imperfect evidence. Checking

[1] Vrij (2018) provides an elaborate discussion of the state-of-the-art methods in lying detection. Besides VA, he discusses 6 prominent methods; see the next section for a brief overview. Among all these methods, VA stands out because of its success and the ease with which it is implemented.

a precise statement gives a more informative signal than checking a vague statement does. If the judge convicts the suspect after his statement was checked and falsified, an additional 'obstruction of justice' penalty is imposed on the suspect (Decker, 2004). The suspect always wants to be acquitted whereas the judge wants to reach a correct verdict. It is also assumed that she (weakly) prefers to wrongly acquit a guilty suspect over wrongly convicting an innocent one.

We derive the perfect Bayesian equilibria of the game and identify the conditions under which partial information revelation occurs in equilibrium. In a partially pooling equilibrium, the innocent suspect always chooses a precise statement, whereas the guilty suspect mixes between being precise and remaining vague. The judge now and then verifies the precise statement, whereas a vague statement leads to immediate conviction. In this equilibrium, the judge has effectively two sources of information working together to determine whether the suspect is innocent or guilty: the strategic behavior of the suspect (i.e. the statement per se) and the outcome of the verification. Pooling equilibria in which no strategic information is revealed always exist. Whenever pooling and partially pooling equilibria coexist, standard equilibrium refinements typically favor the latter.

Our equilibrium analysis points to the supporting role of the obstruction penalty. We show that the obstruction penalty should exceed a certain threshold – in particular, it should exceed the odds ratio of the verification mechanism being inaccurate – for strategic information revelation via the statements per se to occur in equilibrium. However, contrary to what one might expect, increasing the obstruction penalty further does not increase the provision of valuable information. What the judge can learn from the suspect's statement per se remains unaffected, because guilty suspects keep on lying with the same frequency. The amount of valuable information obtained via verification is actually reduced, because a higher obstruction penalty induces the judge to investigate less. Equilibrium payoffs of both the judge and the two suspect types are also unaffected.

In line with the earlier described dilemma that liars face, an improvement in the accuracy of the verification mechanism makes that more can be learned from the strategic behavior of the suspect. First, via (among others) lowering the threshold for the minimum obstruction penalty required, it enlarges the set of parameters for which both informational sources are effectively available. Second, within this set, a higher accuracy induces the guilty type to lie less often. Interestingly, although all else equal the improved accuracy would by itself also have led to more valuable information obtained via verification, in the partially pooling equilibrium it actually leads to less. The main drivers here are that precise statements are made less often by the guilty type and (therefore) are also verified less often by the judge. Hence, if the verification technology becomes more accurate, the additional benefits that come with it are purely due to the deterrence effect of the potential verification. Besides those of the judge, also the expected payoffs of the innocent type increase with an improved reliability of verification. The guilty type is effectively not harmed, as it induces him to lie less and thus to suffer the penalty for obstruction less frequently.

Finally, a decrease in the investigation costs has similar effects on the amount and source of valuable information obtained in equilibrium as an improved reliability has; driven by the deterrence effect again more information is obtained from the strategic behavior of the suspect itself and less from the actual verification of messages. However, although the judge benefits, the expected payoffs of the innocent (as well as the guilty) suspect are unaffected. The upshot of the above results is thus that especially improvements in the accuracy of the verification technology are beneficial. More valuable information is transmitted and both the judge and the innocent type of suspect gain. Apart from having to surpass a minimum threshold – which becomes easier to meet the higher the reliability of the verification technology is – a higher obstruction penalty is not helpful at all.

We extend our analysis in two ways. First, we allow the suspect to confess to receive a penalty reduction. In that case there still exists a partially pooling equilibrium with the only difference being that the guilty suspect then mixes between mimicking the innocent suspect and confessing. We show that the condition for the partially pooling equilibrium to arise can be satisfied either by increasing the obstruction penalty or by increasing the penalty reduction after confession. An obstruction penalty and a penalty reduction after confession are thus essentially two sides of the same coin; they both facilitate strategic information revelation by the suspect.[2] Second, we also consider the case in which the suspect has a 'right to silence'. In that case silence cannot be held against the suspect, effectively restricting the judge's choice of action after a vague statement. Such a right to silence may alter, but does not eliminate, strategic information revelation by the suspect and thus neither its complementary role to direct verification of statements.[3] It also makes that the obstruction penalty no longer plays a supportive role, reinforcing that especially a higher reliability is advantageous.

The remainder of this paper is organized as follows. Section 2 briefly discusses various lie detection methods that have received attention in the psychology literature and the verifiability approach in particular. Section 3 presents the setup of our baseline model. In Section 4 we first derive the set of perfect Bayesian equilibria. We subsequently discuss which of these equilibria can arguably be considered more plausible based on standard (payoff dominance or stability-based) equilibrium refinements from the literature. We end the equilibrium analysis section with discussing how the effective reliance on the different information sources varies with the characteristics of the verification technology. In Section 5 we consider two extensions of the baseline setup: the possibility of plea bargaining and accounting for a right to silence. Here we also discuss the connection with earlier game-theoretic analyses of the latter two aspects within the law and economics literature. Section 6 summarizes the paper and concludes.

2 Lie detection and the Verifiability Approach

The origins of deception detection research can be traced back to Zuckerman et al. (1981) who categorized emotion, arousal, control and cognitive processing as four different cues to

[2] As such, our paper relates to earlier game-theoretic analyses of plea bargaining; see Grossman and Katz (1983), Reinganum (1988), Baker and Mezzetti (2001), Bjerk (2007), Kim (2010) and Tsur (2017). When discussing plea bargaining in Subsection 5.1, we make this connection (as well as the differences) with our model precise.

[3] The right to silence has been analysed from a game-theoretic perspective by Seidmann and Stein (2000), Seidmann (2005), Mialon (2005) and Leshem (2010). In Subsection 5.2 we discuss the insights from these studies within the context of our model.

deception. Various methods were developed over the years which were based on the first three of these cues. The methods focused on non-verbal behavior, compared levels of arousal between liars and truth-tellers and did not intervene in the information gathering process. In a meta-analysis, DePaulo et al. (2003) showed that those methods were not reliable as the observed behaviors showed no direct links to deception. According to Vrij (2019), to overcome those issues, modern research in deception detection has made three major shifts. Modern methods focus on the content of a statement, take into account the cognitive process behind lying and have developed interview protocols to optimize the information gathering process. Some of these are already admissible as evidence in courts in countries like the United States, Germany and the Netherlands (Vrij, 2000).

Vrij (2018) provides an elaborate discussion of the state-of-the-art methods in deception detection. He compares the 7 most prominent methods in terms of how ready they are to be applied in judicial systems. The list of methods includes Criteria Based Content Analysis (CBCA), Reality Monitoring (RM), Scientific Content Analysis (SCAN), Cognitive Credibility Assessment, Strategic Use of Evidence (SUE), the Verifiability Approach (VA) and Assessment Criteria Indicative of Deception (ACID). He does so on the basis of 14 criteria, which can be grouped in two sets: academic, such as whether the method has been tested and whether it has been subjected to peer review, as well as procedural, such as whether it is easy to use and whether it provides an information gathering protocol. Five of those criteria, also known as the Daubert standard, are the minimal requirements for scientific evidence to be admissible in US courts (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 1993).[4]

Of the seven methods, only three abide to the Daubert standard, namely RM, ACID and VA. However, ACID is not easy to incorporate in interviews and RM does not provide a within-subject measure of truthfulness. Hence, our paper models VA as the investigation mechanism available to the judge.

As the name suggests, VA is based on the verifiability of details. A detail is considered verifiable if it describes an activity experienced with an identifiable person or witnessed by an identifiable person or recorded through technology (Nahari et al., 2014). Based on the finding that lying is cognitively more demanding (Vrij et al., 2017), there exist interviewing techniques which aim to magnify the cognitive task for liars.[5] On the one hand, the interviewer asks the interviewee to include as many details as possible. On the other hand, the interviewee would like to avoid mentioning details that can easily be checked by the interviewer. Balancing those orthogonal incentives, one would expect a liar to provide many non-verifiable details in a statement. The ratio of verifiable over non-verifiable details is a within-subject measure of the probability that a statement is true or fabricated. Additional benefits of VA are the fact that it is robust to countermeasures (Nahari et al., 2014) and that VA scoring could be computer-automatized as suggested by Kleinberg et al. (2016).

[4] The full list of criteria for admissibility of scientific evidence in US courts is: i) Has the technique been tested in actual field conditions (and not just in a laboratory)?; ii) Has the technique been subject to peer review and publication?; iii) What is the known or potential rate of error?; iv) Do standards exist for the control of the technique's operation?; and v) Has the technique been generally accepted within the relevant scientific community?

[5] For example, asking surprise questions or requesting a narrative in reverse chronological order have been shown to be successful in Vrij et al. (2007) and Sorochinski et al. (2014).

3 Baseline model

Although the strategic interaction that we model arguably matches various real life applications (cf. the Introduction), for concreteness we describe it in terms of one specific application, namely the interaction between a suspect (speaker) and a judge (investigator). Assume a crime has been committed and a suspect (he) is being questioned. The judge (she) can use the statement of the suspect to update her beliefs on his innocence. She can do so immediately or after conducting a costly investigation that can, with some commonly known error probability, verify or falsify the statement. The suspect wants to be acquitted and the judge wants to make the correct decision, viz. to acquit innocent suspects and to convict guilty suspects. Additionally, the judge prefers to acquit a guilty suspect over convicting an innocent one.

We immediately note that our conceptualization of the interaction between a suspect and a judge is based on a number of simplifications. In real life, a suspect can get arrested by the police, provide a statement and a prosecutor may decide whether to impose charges and bring the case to court or not. If she does so, then the suspect becomes a defendant and may provide additional testimony during the trial. All evidence is examined by a judge and/or a jury and once a verdict is reached, the judge imposes a penalty or not. In our simplified (reduced form) model, we have condensed the timing, the actors and the type of information provided. We use 'judge' as label for a representative of the judicial system with the understanding that in practice some actions described in the model might be taken by prosecutors or the jury.[6] Essentially, our model assumes that at some point during the entire judicial process, the suspect will be asked to provide some information. The untruthfulness of this information is assumed to have consequences for the sentence the suspect may be facing, if he gets convicted.

Our model corresponds to a sender-receiver game, where the sender is the suspect (S) and the receiver is the judge (J). The suspect knows his own type, that is whether he is innocent (S^I) or guilty (S^G). The type of the suspect is unknown to the judge, but we assume a commonly known prior belief of $b = Pr(S = S^I)$ that the suspect is innocent. These prior odds can be interpreted as the evidence collected by the judge before questioning the suspect, so that in principle she can convict without requesting a statement.

The suspect can choose between two actions. He can choose to answer all the questions,[7] which results in a precise statement (P), or he can choose to essentially remain silent, which results in a vague (V) statement.[8] After seeing the statement, the judge must reach a verdict to acquit (A) or convict (C) the suspect. This decision can be taken either before or after investigating (I) the statement of the suspect.

The investigation mechanism works as follows. If the judge decides to investigate statement $j \in \{V, P\}$, the investigation mechanism provides an outcome that has a probability of r_j of

[6] Assuming a unitary actor for the judicial system is an arguably reasonable simplification to the extent that the various actors within the judiciary share the same preferences and information. We briefly return to this in Subsection 5.1 where we discuss the possibility of plea bargaining and relate our strategic setup to existing models of plea bargaining in the literature.

[7] An implicit assumption in the model is that when answering questions, an innocent suspect tells the truth whereas a guilty suspect lies. Allowing both of them to choose whether to answer truthfully or not is a possible extension for future research.

[8] In Appendix B we extend our model by allowing the suspect to reveal an arbitrary number of verifiable details. All the main insights of the baseline model presented in the main text remain valid in this richer model.

being correct (which means verified for the statement of the innocent type and falsified for the statement of the guilty type) and a probability of $1-r_j$ of being wrong (which means falsified for the statement of the innocent type and verified for the statement of the guilty type). Parameters r_V and r_P thus reflect the reliability of investigating the various statements. We assume that the investigation mechanism has at least some informational value, in the sense that it gets the judge closer to the truth. This assumption translates to both probabilities r_V and r_P being (weakly) larger than 0.5.[9] Aligned with the psychology literature on content-based deception detection methods, we also assume that the differences in content between the statement of the innocent and the guilty type will be more pronounced in a more detailed statement (Harvey et al., 2017). As a result, investigating a precise statement is more likely to produce a correct outcome than investigating a vague one, i.e. we assume that investigation probabilities satisfy $0.5 \leq r_V < r_P < 1$.[10]

Preferences of the two suspect types are assumed to be as follows. Both suspect types get a payoff of 1 if they get acquitted. If they get convicted, they receive a lower payoff which depends on the amount of evidence that resulted in their conviction. If they get convicted on the basis of prior evidence, which happens when the judge does not investigate the statement or when investigation verifies the statement and provides no additional evidence against them, they receive a normalized payoff of 0 (so the imposed sentence leads to a payoff reduction of 1). If they get convicted after their statement j was investigated and falsified, they receive an additional obstruction penalty of π_j and their payoff equals $-\pi_j$.[11] We assume that this obstruction penalty is only applied when a suspect is eventually convicted.[12] The size of the penalty could depend on the amount of lying by the suspect. In line with the dilemma that liars face (cf. the Introduction), we assume that the penalties π_V and π_P are such that in principle the guilty type fears investigation of a vague statement less than investigation of a precise statement, because the expected overall penalty he might get is larger for a precise statement: $r_V(1+\pi_V) < r_P(1+\pi_P)$. Moreover, we also assume that $r_V(1+\pi_V) < 1$, for otherwise the

[9] Our sender-receiver game shares with standard signaling (Spence, 1973) and cheap talk (Crawford and Sobel, 1982) games that the sender-suspect can say anything, including outright lies. However, it differs in the assumption that statements can be imperfectly verified by the receiver at a cost. Because verification is informative in unmasking the suspect's type ($r_j > \frac{1}{2}$), this drives the potential for strategic information transmission via the statement made (cf. the sorting condition in signaling games, which requires a given message to be more costly or more beneficial for one type than for the other). In so-called disclosure games (Milgrom, 2008), verification is assumed to be costless and perfect, essentially making that the sender cannot lie. This in turn makes that in equilibrium there will be full disclosure, as any piece of information that is not revealed is hold against the sender by a skeptical receiver. In our setup, imperfect and costly verification stands in the way of equilibrium full disclosure.

[10] To the extent that the vague statement literally corresponds to remaining silent, there is obviously little to verify. Verification of a vague statement could therefore equally well be interpreted as additional independent investigation by the judge not inspired by the (empty) statement made. With a more precise statement, the judge potentially gets better clues exactly what to look for, allowing her to steer her investigation in a more promising direction.

[11] This penalty can be interpreted in various ways. If the lying was under oath, then the defendant may be charged with perjury (US Sentencing Commission, 2018, 2J1.3.). If the lying significantly impeded official investigation, then the defendant may be charged with obstruction of justice (US Sentencing Commission, 2018, 3C1.1.). The sentencing guidelines also recommend a reduction of penalty if a defendant provided substantial assistance in the investigation, for example by giving a truthful, complete and reliable testimony (see 5K1.1.). In this case, the penalty can be interpreted as the difference between the full and the reduced sentence.

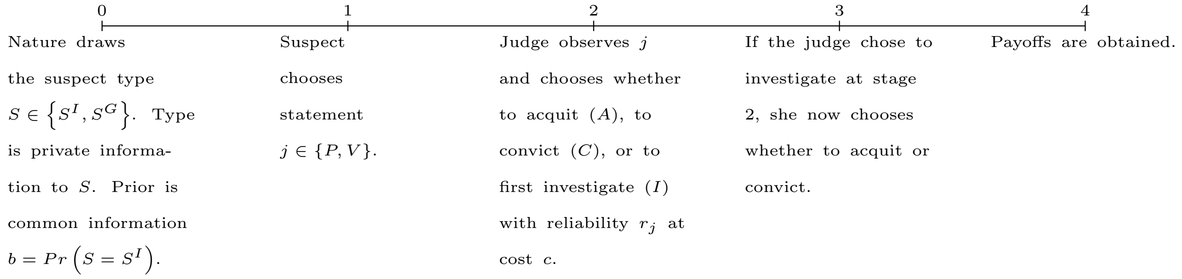
[12] A prosecutor will very often drop a criminal charge if it is determined that the evidence against the accused is not strong enough, see Cohen (1992).

guilty type might want to confess immediately and obtain 0 as to avoid a negative expected payoff from investigation of a vague statement.

The preferences of the judge are modelled in the following way. The judge gets 1 for reaching a correct verdict, that is acquit an innocent suspect and convict a guilty suspect. In case the judge makes a mistake, she receives a lower payoff that depends on the type of mistake made. We normalize the payoff of acquitting a guilty suspect to 0 and set the payoff of convicting an innocent suspect to $-\alpha$. The assumption that $\alpha \geq 0$ captures the notion that the judge (weakly) prefers to let go a guilty suspect over sending an innocent suspect to jail. Higher values of α result in a tighter threshold on the judge's belief for her to prefer conviction over acquittance; it thereby essentially quantifies exactly what is meant by "beyond any reasonable doubt". In particular, with these payoffs the (updated) belief that the suspect is innocent should exceed the tipping point of $\frac{1}{2+\alpha}$ for the judge to acquit.

We finally assume that the judge has to pay a positive cost $c > 0$ to investigate statement j. These costs not only reflect that investigating the truthfulness of statements is costly in terms of the resources needed (time and detectives), but may also capture other, more indirect types of costs. For instance, in criminal cases of high importance that receive widespread public attention, society often really disapproves cases that last for years, so our cost parameter could also be seen as pressure to reach a verdict faster. Note that our assumptions regarding the judge's payoffs arguably make these largely aligned with what society would seem to require. Her expected payoffs could thus potentially serve as a first approximation to a more encompassing welfare analysis.

Figure 1 provides a succinct summary of the order of moves in the strategic interaction between the suspect and the judge and Table 1 summarizes the payoffs of all agents.



0: Nature draws the suspect type $S \in \left\{S^I, S^G\right\}$. Type is private information to S. Prior is common information $b = Pr\left(S = S^I\right)$.

1: Suspect chooses statement $j \in \{P, V\}$.

2: Judge observes j and chooses whether to acquit (A), to convict (C), or to first investigate (I) with reliability r_j at cost c.

3: If the judge chose to investigate at stage 2, she now chooses whether to acquit or convict.

4: Payoffs are obtained.

Figure 1: Timeline of the game

For the judge the main goal of the entire process is to get a better idea of whether the suspect is guilty or not. Starting from a prior belief b that the suspect is innocent, after seeing statement j the judge updates her initial belief based on the strategic behavior of the suspect. Let p^I denote the probability with which the innocent type gives a precise statement (and hence with probability $1 - p^I$ this type remains vague). Similarly so, p^G gives the probability that the guilty type makes a precise statement. Using Bayes' rule, a rational judge then updates her

	Convict		Acquit	
	Suspect	Judge	Suspect	Judge
Without verification				
Innocent	0	$-\alpha$	1	1
Guilty	0	1	1	0
With verification				
Innocent	$-\pi_j$	$-\alpha - c$	1	$1 - c$
Guilty	$-\pi_j$	$1 - c$	1	$-c$

Note: By assumption $\alpha \geq 0$, $c > 0$ and $0 < r_V\,(1 + \pi_V) \leq \min\{1, r_P\,(1 + \pi_P)\}$.

Table 1: Payoffs of suspect and judge for all type-action combinations

belief to::[13]

$$b^P \equiv Pr(\text{S is innocent}|\text{statement is } P) = \frac{bp^I}{bp^I + (1-b)p^G} \tag{1}$$

$$b^V \equiv Pr(\text{S is innocent}|\text{statement is } V) = \frac{b(1-p^I)}{b(1-p^I) + (1-b)(1-p^G)} \tag{2}$$

Having seen a statement j, the judge convicts, investigates, and acquits with respective probabilities q_C^j, q_I^j and q_A^j. In case the judge investigates, she obtains additional information that allows her to update her beliefs another time, based on the outcome of the investigation. From the given reliability of the investigation process and again Bayes' rule, we immediately obtain that these beliefs equal (for $j \in \{V, P\}$):

$$b^{j+} \equiv Pr(\text{S is innocent}|\text{Statement is } j \text{ and verified}) = \frac{b^j r_j}{b^j r_j + (1-b^j)(1-r_j)} \tag{3}$$

$$b^{j-} \equiv Pr(\text{S is innocent}|\text{Statement is } j \text{ and falsified}) = \frac{b^j(1-r_j)}{b^j(1-r_j) + (1-b^j)r_j} \tag{4}$$

From these expressions, together with $r_j > 0.5$, it follows that $b^{j-} \leq b^j \leq b^{j+}$. Falsification of the statement made by the suspect thus lowers the judge's belief that he is innocent, while a verified statement increases this belief.

4 Equilibrium analysis

4.1 Perfect Bayesian equilibria

As explained in the previous section, besides her prior belief, the judge in principle has two information sources available: investigation of the actually received statement (at cost c) and the potentially different strategies p^I and p^G the two types of suspects employ in making statements. In this section we explore the extent to which these (additional) information sources are actually

[13] Bayes' rule can only be applied for statements that are made in equilibrium, i.e. when the denominators in expressions 1 and 2 are strictly positive. If not, out-of-equilibrium beliefs have to be formed. These are discussed in the next section.

drawn upon in equilibrium and how they interact, by providing an encompassing equilibrium analysis. The equilibrium concept we use is perfect Bayesian equilibrium.

First consider the choice of the judge whether to investigate or not. Because investigation is costly to her, the judge is willing to do so only if investigation yields her valuable information. That is, the information received should be *influential*; the judge's optimal decision whether to acquit or convict should vary with the outcome of the investigation process.[14] Otherwise the judge could better immediately opt for the decision she would in the end take anyway and avoid costly investigation altogether. Influential information requires that beliefs satisfy $b^{j-} \leq \frac{1}{2+\alpha} \leq b^{j+}$, such that the judge acquits when the suspect's statement is verified and convicts when the statement is falsified. Lemma 1 details this requirement in terms of b^j.

Lemma 1. *Investigating statement j is influential if and only if: $\frac{1-r_j}{\alpha r_j+1} \leq b^j \leq \frac{r_j}{\alpha(1-r_j)+1}$. In that case the judge would acquit if statement j were to be verified and convict if statement j were to be falsified.*

Proof. Investigating statement j is influential as long as $b^{j-} \leq \frac{1}{2+\alpha} \leq b^{j+}$. Using expressions (3) and (4) for b^{j+} and b^{j-} in the previous section and rewriting immediately gives the result. □

Intuitively, investigation can be influential only if, after having just heard statement j and correctly inferring the suspect's strategic behavior (i.e. p^I and p^G), the judge is still insufficiently confident about the suspect's type. That is, she is neither sufficiently convinced that the suspect is guilty (b^j is not very low), nor sufficiently convinced that the suspect is innocent (b^j is neither very high).

Obtaining influential information is a necessary requirement for the judge to investigate, yet it is not a sufficient. The expected benefits from the influential information received should also outweigh the costs of investigation. Lemma 2 precisely characterizes this requirement and pins down the judge's optimal choice for any belief $b^j \in [0,1]$ she might have.

Lemma 2. *Define $\underline{b}(r_j) \equiv \min\left\{\frac{(1-r_j)+c}{\alpha r_j+1}, \frac{1}{2+\alpha}\right\}$ and let $\overline{b}(r_j) \equiv \max\left\{\frac{r_j-c}{\alpha(1-r_j)+1}, \frac{1}{2+\alpha}\right\}$. After statement j and based on belief b^j, the judge's optimal choice of action equals:*

(a) convict if $b^j < \underline{b}(r_j)$;

(b) investigate if $b^j \in (\underline{b}(r_j), \overline{b}(r_j))$;

(c) acquit if $b^j > \overline{b}(r_j)$.

The interval $\left(\underline{b}(r_j), \overline{b}(r_j)\right)$ is non-empty and equals $\left(\frac{(1-r_j)+c}{\alpha r_j+1}, \frac{r_j-c}{\alpha(1-r_j)+1}\right)$ iff $c < \frac{\alpha+1}{\alpha+2} \cdot (2r_j - 1)$. In that case, the judge is indifferent between convict and investigate if $b^j = \underline{b}(r_j)$, and indifferent between investigate and acquit if $b^j = \overline{b}(r_j)$. If $c > \frac{\alpha+1}{\alpha+2} \cdot (2r_j - 1)$ and thus $\underline{b}(r_j) = \overline{b}(r_j) = \frac{1}{2+\alpha}$, the judge is indifferent between convict and acquit when $b^j = \frac{1}{2+\alpha}$.

Proof. See Appendix A. □

[14] Note that the notion of the investigation being influential is stronger than that it being *informative*. The latter holds as long as long as the outcome of the investigation is more likely to be aligned with the truth, which is guaranteed by our assumption that $r_j > 0.5$ for $j = P, V$. Sobel (2020) provides an insightful discussion of the differences between the definitions of informative and influential.

Intuitively, the range of beliefs $b^j \in \left(\underline{b}(r_j), \overline{b}(r_j)\right)$ for which investigation pays off widens if the verification process becomes more reliable, i.e. when r_j increases, and when investigation becomes cheaper (lower c). If non-empty, the interval always contains the tipping point $\frac{1}{2+\alpha}$ between acquitting and convicting; the further away beliefs b^j are from this point of indifference, the more confident the judge is to solely act on the basis of the existing evidence and to skip costly investigation altogether.

With Lemma 2, we can now easily characterize the equilibria in which investigation is the only potential source of information to the judge. In these so-called pooling equilibria, the two suspect types choose to make the same type of statement, either vague or precise (i.e. $p^I = p^G \in \{0,1\}$). In that case, $b^j = b$ for the pooling statement j made and the judge's response to statement j directly follows from Lemma 2. The only thing left to specify are out-of-equilibrium beliefs b^{-j} for the opposite (off-the-equilibrium path) statement $-j$ that actually support pooling on j as equilibrium behavior. By assuming skeptical out-of-equilibrium beliefs, i.e. $b^{-j} = 0$ such that statement $-j$ would induce immediate conviction, the circumstances under which pooling on statement j can occur as equilibrium behavior are maximized. Building on this observation, Proposition 1 characterizes all pooling equilibria.

Proposition 1. *For the pooling equilibria it holds that:*

(a) A pooling equilibrium in which both types make a vague statement ($p^I = p^G = 0$) exists for any prior belief b. The judge's response to the vague statement made follows from Lemma 2, with $b^V = b$.

(b) (i) If $\pi_P < \frac{1-r_P}{r_P}$, a pooling equilibrium in which both types make a precise statement ($p^I = p^G = 1$) exists for any prior belief b. The judge's response to the precise statement made follows from Lemma 2, with $b^P = b$.

(ii) If $\pi_P > \frac{1-r_P}{r_P}$, a pooling equilibrium in which both types make a precise statement ($p^I = p^G = 1$) only exists if belief $b^P = b$ satisfies $b \notin (\underline{b}(r_P), \overline{b}(r_P))$. In that case, if $b < \underline{b}(r_P)$ the judge convicts after a precise statement and if $b > \overline{b}(r_P)$ she acquits after a precise statement.

Proof. See Appendix A. □

Besides pooling equilibria, there are potentially other types of equilibria in which the strategic behavior of the two suspect types - as reflected by their respective probabilities p^I and p^G of making a precise statement - does reveal some (influential) information. A first immediate observation is that this strategic behavior will certainly not be fully revealing though. This holds because, if the two types would always choose different statements, the one chosen by the innocent type would lead to immediate acquittance whereas the one made by the guilty type would trigger conviction. But then the guilty type would obviously prefer the former statement and mimic the innocent type. For ease of reference, we formulate this simple observation as a separate lemma.

Lemma 3. *A separating equilibrium in which $p^I \in \{0,1\}$ and $p^G = 1 - p^I$ does not exist.*

Given that the verification process is assumed to be imperfect ($r_V < r_P < 1$), a direct consequence of Lemma 3 is that there will always remain circumstances under which the judge does not know the suspect's type for sure and (in some of these instances) takes the wrong decision.

From Lemma 3 it follows that, besides pooling, the only other type of equilibria left to consider are partially pooling equilibria in which the two types employ different, but partly overlapping strategies ($p^I \neq p^G$ and $0 < p^s < 1$ for some $s \in \{I, G\}$), and that induce the judge to let her decision depend on the statement received (i.e. $(q_C^V, q_I^V, q_A^V) \neq (q_C^P, q_I^P, q_A^P)$).[15] Proposition 2 below characterizes the nature of strategic information revelation in a partially pooling equilibrium and the circumstances under which such an equilibrium exists.

Proposition 2. *For the unique partially pooling equilibrium it holds that:*

(a) The innocent type makes a precise statement for sure ($p^I = 1$); the guilty type mixes between making a vague and making a precise statement with $p^G = \frac{b}{1-b} \cdot \frac{(1-r_P)(1+\alpha)+c}{r_P - c}$.

(b) After a precise statement the judge updates her belief to $b^P = \frac{r_P - c}{\alpha(1-r_P)+1}$; she then investigates with probability $q_I^P = \frac{1}{r_P(1+\pi_P)}$ and acquits with the remaining probability. After a vague statement she always convicts.

(c) Necessary and sufficient conditions for existence are:
(i) $\pi_P > \frac{1-r_P}{r_P}$, (ii) $c < \frac{\alpha+1}{\alpha+2} \cdot (2r_P - 1)$ and (iii) $b < \frac{r_P - c}{\alpha(1-r_P)+1}$.

Proof. See Appendix A. □

In the partially pooling equilibrium of Proposition 2, the judge's two information sources complement each other. Strategic information revelation by the two suspect types makes that after a precise statement, the judge updates her belief that the suspect is innocent upwards ($b^P > b$). She now and then verifies such a precise statement just to be sure but, more importantly, essentially only to discourage the guilty type from making such a statement too often. Improvements in the investigation process – viz. an increase in reliability r_P or a decrease in investigation costs c – do not increase the likelihood that the judge verifies a precise statement (q_I^P is decreasing in r_P and independent of c). Yet there are positive spill-over effects towards the statement strategy of the guilty type, as such improvements induce him to mimic the innocent type less often; p^G is decreasing in r_P and increasing in c. The overall effect of improvements in the investigation technology is thus that less investigation takes place. This reflects the general intuition that a more credible or more effective stick works as a stronger deterrent and thus in the end needs to be used less often.

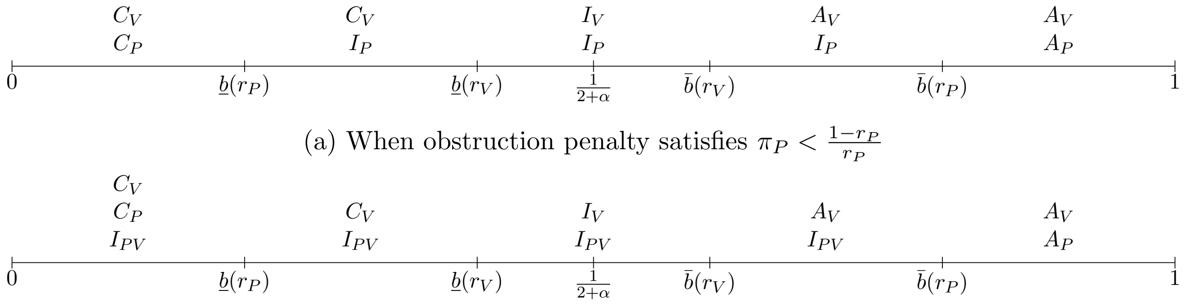
Another noteworthy feature of the partially pooling equilibrium is that the threat of investigation and the obstruction penalty work in tandem. The higher reliability of checking on

[15] Formally one could alternatively classify any equilibrium in which $p^I \neq p^G$ – and thus at least some information is revealed – as a partially pooling equilibrium. In line with the informative vs. influential information divide discussed earlier, for ease of reference we only label those equilibria in which the suspect's statement strategies yield influential information as partially pooling equilibria. Equilibria in which some information is revealed on the equilibrium path, but where this information is effectively irrelevant for the judge's choice of action (i.e. non-influential), are outcome equivalent to the pooling equilibria described in Proposition 1. For ease of presentation these are therefore not separately listed.

a precise statement (i.e. $r_P > r_V$) is by itself not a sufficient deterrent to refrain from always making a precise statement; it should be complemented with a sufficiently high additional penalty for the guilty type to be discouraged to always mimic the innocent type. In particular, the obstruction penalty should exceed the odds ratio of the verification being unreliable: $\pi_P > \frac{1-r_P}{r_P}$.

From Proposition 1 and Proposition 2 it follows that a necessary requirement for the judge to obtain some additional information in equilibrium beyond her prior belief (either through investigation or from the suspect's strategic behavior), is that investigation costs are not too high: $c < \frac{\alpha+1}{\alpha+2} \cdot (2r_P - 1)$ should hold. Otherwise her optimal choice of action – either convict or acquit – is fully determined by her prior belief. In the remainder we assume this cost condition to be satisfied. In that case, multiple equilibria exist side by side which may differ in the judge's equilibrium choice of action if $b \leq \bar{b}(r_P) = \frac{r_P - c}{\alpha(1-r_P)+1}$, i.e. essentially if the judge is a priori insufficiently convinced to always acquit were she to receive a precise statement. Especially in that case, the questions of which equilibrium is most attractive to the judge (as well as the suspect types) and which equilibrium is the most plausible one to arise become relevant. In the next subsection we turn to these questions.[16]

For ease of reference, Figure 2 summarizes the existence of the various equilibria as a function of prior belief $b \in [0, 1]$ that the suspect is innocent. This figure uses the following notation. Based on the judge's equilibrium choice (in capitals) and the suspects' equilibrium statements (in subscripts), let C_V refer to the pooling equilibrium in which both suspect types make a vague statement and – given that prior belief b is below threshold $\underline{b}(r_V)$ – the judge decides to convict. We use similar notation to refer to the other pooling equilibria, viz: C_P, A_V, A_P, I_V and I_P, respectively. The partially pooling equilibrium is labelled I_{PV}, where the subscripts thus reflect that both type of statements occur on the equilibrium path. The figure is split in two panels, each corresponding to the values of π_P for which equilibria I_P and I_{PV} exist. As these intervals are non-overlapping, equilibria I_P and I_{PV} cannot exist side by side.



(a) When obstruction penalty satisfies $\pi_P < \frac{1-r_P}{r_P}$

(b) When obstruction penalty satisfies $\pi_P > \frac{1-r_P}{r_P}$

Figure 2: Equilibria of the game as a function of prior belief b

[16] As illustrated by Proposition 1, when $b > \bar{b}(r_P)$ two different pooling equilibria exist side by side. The question which of these is more plausible to arise is essentially immaterial though, as both lead to the exact same outcome (viz. acquittance) and exact same payoffs to the parties involved.

4.2 Equilibrium selection

Depending on prior belief b, various equilibria may exist side by side (cf. Figure 2). These equilibria differ in terms of their relative attractiveness to the different parties involved. This in turn makes that arguably some equilibria are more plausible than others. Intuitively, if both suspect types and the judge all agree that one of the co-existing equilibria is best, this seems to make this (payoff dominant) equilibrium very focal. And as more information is always better for the judge, this then corresponds with the most informative equilibrium. In fact, if only the innocent type agrees with the judge that more information is better – because the judge's prior belief is such that it does not lead to immediate acquittance – it already seems most plausible that the parties coordinate on the most informative equilibrium available. In that case the innocent type would like to do anything to separate himself out as to avoid pooling with the guilty type, and off-the-equilibrium path statements should arguably be interpreted in this light. The analysis below makes these intuitions precise, by formally applying the payoff dominance (Harsanyi and Selten, 1988) and divinity (Banks and Sobel, 1987) equilibrium selection arguments to our game. To that purpose Table 2 lists the expected payoffs the parties get in the different equilibria.

Equilibrium	Judge	Innocent type	Guilty type
C_V, C_P	$1 - b(1+\alpha)$	0	0
A_V, A_P	b	1	1
I_V	$r_V - b(1-r_V)\alpha - c$	$1-(1-r_V)(1+\pi_V)$	$1 - r_V(1+\pi_V)$
I_P	$r_P - b(1-r_P)\alpha - c$	$1-(1-r_P)(1+\pi_P)$	$1 - r_P(1+\pi_P)$
I_{PV}	$1 - b \cdot \frac{(1-r_P)(1+\alpha)+c}{r_P - c}$	$1 - \frac{1-r_P}{r_P}$	0

Table 2: Expected payoffs

We first consider cases in which two informative equilibria exist side by side, viz. where either I_V and I_P, or I_V and I_{PV}, co-exist (cf. Figure 2). Intuitively, more information allows the judge to arrive at better verdicts. Hence, we immediately have that for her $I_V \prec_J I_P$, because checking a precise statement is more reliable than (but equally costly as) checking a vague statement is.[17] Similarly so, the judge prefers the partially pooling equilibrium I_{PV} over the pooling on a vague statement equilibrium I_V. In the former, the strategic behavior of the two suspect types already conveys some useful (and costless) information. In contrast to the judge, the guilty suspect prefers the less informative equilibria, essentially because he has something to hide: both $I_P \prec_{SG} I_V$ and $I_{PV} \prec_{SG} I_V$ hold. Matters are a priori ambiguous for the innocent type. He likes that investigation of a precise statement is more reliable than investigation of a vague statement is (i.e. $r_P > r_V$). Yet he dislikes that, in case the investigation results in the wrong conclusion (i.e. when the statement of the innocent type is falsified), a potentially larger obstruction penalty is imposed after a precise statement, i.e. $\pi_P > \pi_V$. Overall, with respect to the two pooling equilibria, this type prefers the more informative equilibrium I_P over I_V if and only if the gain in reliability outweighs the downside of getting the potentially higher penalty: $I_V \prec_{SI} I_P$ iff $(1-r_V)(1+\pi_V) < (1-r_P)(1+\pi_P)$. The conditions for existence of the

[17] Here we use $\prec_J$ to denote the preference relation of the judge, with $A \prec_J B$ reflecting that the judge strictly prefers B over A in terms of expected payoffs. For the two suspect types we employ a similar notation.

partially pooling equilibrium – in particular $\pi_P > \frac{1-r_P}{r_P}$ – make that this equilibrium is always preferred by her over the pooling on vague equilibrium ($I_V \prec_{S^I} I_{PV}$). The intuition here is that in equilibrium I_{PV}, investigation of the statement made by the innocent type only occurs with probability $q_I^P = \frac{1}{r_P(1+\pi_P)}$; otherwise he is immediately acquitted.

Next, consider the cases in which an informative equilibrium (either I_P or I_{PV}) exists alongside a non-informative one (either C_V, C_P or A_V) in which the judge either convicts or acquits for sure. Again, the judge always prefers the more informative equilibrium. For the suspect types this depends on the nature of the non-informative equilibrium. If in that equilibrium the judge always convicts, more information cannot harm the suspect and he is also better off in the informative equilibrium, just like the judge is. However, in comparison to equilibrium A_V in which the judge always acquits, the suspect can only lose from the judge getting more information. Here the judge's and suspect's rankings are thus opposite of each other.

Proposition 3 summarizes the cases in which the judge's and the two suspect types' rankings of co-existing equilibria coincide. Essentially this applies when all parties benefit from the judge having more information.

Proposition 3. *The players' rankings (in terms of expected payoffs) of equilibria that may exist side by side coincide in the following cases:*

(i) If $b \leq \underline{b}(r_j)$ and $\pi_P > \frac{1-r_P}{r_P}$: The non-informative equilibrium C_j (for $j = P, V$) is Pareto-dominated by informative equilibrium I_{PV}.

(ii) If $b \in [\underline{b}(r_P), \underline{b}(r_V)]$ and $\pi_P < \frac{1-r_P}{r_P}$: The non-informative equilibrium C_V is Pareto-dominated by informative equilibrium I_P.

Based on the payoff dominance equilibrium selection criterion of Harsanyi and Selten (1988), one would expect the parties to coordinate on the more informative equilibria in at least the instances listed in Proposition 3.

Arguably, coordination on a pooling equilibrium also becomes less likely if it can only be sustained with "unreasonable" out of equilibrium beliefs. In the previous subsection we assumed those off-path beliefs to be skeptical: $b^{-j} = 0$ for a statement $-j$ off-the-equilibrium path. We did so as to maximize the circumstances under which the pooling on statement j equilibrium exists. From the viewpoint of the underlying fundamental incentives that the two suspect types face, such out of equilibrium beliefs seem particularly plausible when the guilty type could potentially gain more from deviating from the equilibrium path than in principle the innocent type could. Note though that for equilibria C_j and A_j skeptical beliefs are inessential. These equilibria are equally well sustainable with $b^{-j} = b$, i.e. when out of equilibrium beliefs are equal to the prior belief.[18] Similarly so, in terms of equilibrium outcomes (i.e. the action chosen by the judge and the payoffs to the parties involved), equilibria C_P and A_P are also equal to any equilibrium in which $0 < p^I = p^G < 1$. In the latter case the two suspect types pool on the exact same mixed strategy and no out of equilibrium statements exist. In a similar spirit, also for C_V and A_V outcome equivalent equilibria can always be constructed in which

[18] One qualification here is that for equilibrium C_V this only holds when $b < \underline{b}(r_P)$. For beliefs $b \in [\underline{b}(r_P), \underline{b}(r_V)]$, any out of equilibrium beliefs $b \leq \underline{b}(r_P)$ would work.

out of equilibrium statements are absent.[19] Overall, equilibria C_j and A_j (for $j = P, V$) thus cannot be disqualified on the basis of the unreasonableness of out-of-equilibrium beliefs.

Matters are different for equilibrium I_V, however. This equilibrium crucially relies on the assumption that after an out of equilibrium precise statement, the judge's belief is updated sufficiently downwards such that she prefers to convict for sure. Given the underlying economic incentives, this is not particularly reasonable though. Recall that the payoff structure is assumed to be such that the guilty type fears investigation of a precise statement more than investigation of a vague statement: $r_P(1 + \pi_P) > r_V(1 + \pi_V)$. For the innocent type this is then necessarily less so the case. That is, under the assumptions made it holds that:[20]

$$(1 - r_P)(1 + \pi_P) - (1 - r_V)(1 + \pi_V) < r_P(1 + \pi_P) - r_V(1 + \pi_V) \tag{5}$$

The l.h.s. of (5) gives the difference in the expected penalty after investigation of a precise versus investigation of a vague statement for the innocent type, whereas the r.h.s. does so for the guilty type. Because the former difference is lower, the innocent type has relatively less to fear from making a precise statement. Given this, the judge would be inclined to conclude that an off-path precise statement is likely to come from the innocent type. If so, she would in turn acquit after a precise statement, undermining the I_V equilibrium. This way of reasoning is formalized in the divinity equilibrium refinement proposed by Banks and Sobel (1987). Proposition 4 makes this precise.

Proposition 4. *The I_V equilibrium does not survive the divinity equilibrium refinement of Banks and Sobel (1987). The unique divine equilibrium when $b \in [\underline{b}(r_V), \overline{b}(r_V)]$ is equilibrium I_P when $\pi_P < \frac{1-r_P}{r_P}$ and equilibrium I_{PV} in case $\pi_P > \frac{1-r_P}{r_P}$.*

Proof. See Appendix A. □

The upshot of Proposition 3 and Proposition 4 is that for prior beliefs $b \leq \overline{b}(r_V)$, it is most plausible that the parties coordinate on the most informative equilibrium that exists, essentially because the judge and the innocent type agree that more information is better. For initial beliefs in between $\overline{b}(r_V)$ and $\overline{b}(r_P)$ this does not hold, however, because the innocent type gets his highest possible payoff in the uninformative equilibrium A_V. An informative equilibrium then need not arise, even though it exists.

As noted in the previous subsection, in equilibrium I_{PV} the threat of investigation and the obstruction penalty are complementary to each other. Together they induce strategic information transmission by the two suspect types via the statements per se, which provides additional information on top of the information obtained from potential verification. For the judge it

[19] For equilibrium A_V the equivalence with a common mixed strategy does not hold when $b \in [\overline{b}(r_V), \overline{b}(r_P)]$, because for $b^P = b$ the judge then would want to investigate a precise statement (see Lemma 2). Yet another outcome equivalent equilibrium then arises when the suspect types employ the different mixed strategies $p^G = 0$ and $p^I = 1 - \frac{\overline{b}(r_V)}{1-\overline{b}(r_V)} \cdot \frac{1-b}{b}$. The innocent type then only now and then makes a precise statement (while the guilty type never does so), such that after a vague statement the judge is still sufficiently convinced to immediately acquit for sure ($b^V = \overline{b}(r_V)$). Again, in this equilibrium no out of equilibrium statements exist. For equilibrium C_V a similar equivalent can be constructed in which (only) the guilty type mixes.

[20] To see this, note that the inequality can be rewritten as $[2r_V - 1](1+\pi_V) < [2r_P - 1](1+\pi_P)$. From $r_V < r_P$ and $r_V(1+\pi_V) < r_P(1+\pi_P)$ then the result follows.

would be best to facilitate the existence of this partially pooling equilibrium.[21] This can be done by setting a sufficiently high obstruction penalty $\pi_P > \frac{1-r_P}{r_P}$. Perhaps somewhat surprisingly, increasing the obstruction penalty beyond this threshold serves no purpose. Although ceteris paribus this would make it less attractive for the guilty type to mimic the innocent type by making a precise statement, in equilibrium this is completely offset by a reduced willingness of the judge to investigate. The judge would be better served by an improvement in the reliability of the verification technology r_P instead. This result squares nicely with (the benefits from) the improved accuracy of the Verifiability Approach as compared to pre-existing lie detection methods that psychologists have documented. In the next subsection, we look at the potentially beneficial impact of changes in the characteristics of the verification process in more detail.

4.3 Effective reliance on information sources

In this subsection, we explore how the amount of (valuable) information revelation and the effective reliance on different information sources varies with the characteristics of the verification technology, as captured by parameters r_P, π_P and c.

The judge does not want to obtain just any information per se, but rather influential information that is instrumental to her decision. The effective value of such information can be inferred from how her payoffs are affected. From the top two rows in Table 2 it follows that the judge's payoffs in the pooling equilibria in which she either convicts or acquits for sure, equal $\min\{1-b(1+\alpha), b\}$. In these equilibria, all decision relevant information comes from the prior belief and neither the suspect's statement per se nor the potential verification thereof provides any additional influential information. This is different in the other two pooling equilibria I_V and I_P. Based on Proposition 4 we focus on I_P only (although for I_V the same comparative statics are obtained). For the case where $b > \frac{1}{2+\alpha}$, the judge's equilibrium payoffs can then be decomposed as:[22]

$$\begin{aligned} r_P - b(1-r_P)\alpha - c &= b + 0 + [(1-b)r_P - b(1-r_P)(1+\alpha)] - c \\ &= \underbrace{b}_{\text{prior}} + \underbrace{0}_{\text{statements per se}} + \underbrace{[r_P - b - b(1-r_P)\alpha]}_{\text{verification}} - c \end{aligned}$$

Based on her prior belief alone, the judge would always acquit and obtain an expected payoff b. As both suspect types always make a precise statement, observing such a statement per se does not provide any additional information and hence also does not create any additional value to her. Verification of the precise statement made has two effects. On the one hand, it improves decision making relative to deciding in the absence of verification (here acquit) when it correctly

[21] Note that $I_V \prec_J I_{PV}$ trivially follows from the fact that the judge in equilibrium I_{PV} could have ignored the information obtained via the statements per se entirely and always verify both statements instead. This would have yielded the same expected payoffs as in I_V. The fact that she does not so implies that with her best response in equilibrium I_{PV} she does better than that.

[22] For the opposite case $b < \frac{1}{2+\alpha}$ we similarly would have:

$$r_P - b(1-r_P)\alpha - c = \underbrace{(1-b) - b\alpha}_{\text{prior}} + \underbrace{0}_{\text{statements per se}} + \underbrace{[r_P - (1-b) + b\alpha - b(1-r_P)\alpha]}_{\text{verification}} - c$$

This gives the exact same comparative statics as for the case $b > \frac{1}{2+\alpha}$ discussed in the main text.

falsifies a precise statement from the guilty type. This happens with probability $(1-b)\,r_P$ and then increases the judge's payoffs by 1 (relative to acquitting for sure). On the other hand, it worsens decision making in the instances where it wrongly falsifies a precise statement coming from the innocent type. This occurs with probability $b\,(1-r_P)$ and then has a negative payoff consequence of $(1+\alpha)$, given that the judge dislikes this type of mistake by an additional amount $\alpha \geq 0$.[23] The overall net effect – reflected within square brackets – is positive and outweighs the costs of verification c. This informational value of verification increases with r_P and is independent of both π_P and c.

The judge's expected payoffs in equilibrium I_{PV} can (again for the case $b > \frac{1}{2+\alpha}$) be similarly decomposed as:[24]

$$\begin{aligned} 1-(1-b)\,p^G &= 1-(1-b)\,p^G + \sigma_P q_I^P \left\{\left[\left(1-b^P\right) r_P - b^P(1-r_P)\,(1+\alpha)\right] - c\right\} \\ &= \underbrace{b}_{\text{prior}} + \underbrace{(1-b)\left(1-p^G\right)}_{\text{statements per se}} + \underbrace{\sigma_P q_I^P\left[r_P - b^P - b^P(1-r_P)\alpha\right]}_{\text{verification}} - \sigma_P q_I^P c \end{aligned}$$

where $\sigma_P = b + (1-b)\,p^G$ denotes the overall probability that a precise statement is made in equilibrium. The term within curly brackets equals zero, reflecting that on the equilibrium path the judge is indifferent between verifying a precise statement and immediately acquitting for sure. Note that in this equilibrium, caused by the strategic behavior of the two suspect types, the statements per se now do provide additional influential information. The incremental value of this information equals $(1-b)\left(1-p^G\right)$, which increases with r_P, is independent of π_P, and decreases with c. This reflects the 'deterrence effect' of possible verification. If either verification becomes more reliable (higher r_P) or less costly (lower c), it deters the guilty type from mimicking the innocent type by making a precise statement less often (i.e. it lowers p^G). This in turn makes such a statement per se more informative. The obstruction penalty π_P has no impact on the deterrence effect, essentially because it reduces the frequency of actual verification q_I^P at the same time. Likewise, the incremental value of now and then checking on a precise statement made equals $\sigma_P q_I^P\left[r_P - b^P - b^P(1-r_P)\alpha\right]$. This value *decreases* with both r_P and π_P, and *increases* with c.

As is immediate from Table 2, the judge's overall expected payoffs in equilibrium I_{PV} are increasing in r_P. Intuitively, if the verification technology becomes more reliable, overall more valuable information is obtained. But perhaps somewhat counter-intuitively, this beneficial impact is completely driven by the incremental benefit from the statements per se. Less additional information is actually obtained from verification the higher r_P is. Effectively, the judge thus relies more on the induced changes in the strategic behavior of the suspect when r_P becomes

[23] The analysis presented in this subsection applies for any $\alpha \geq 0$, thus also for $\alpha = 0$. This effectively implies that the exact same conclusions are obtained if we just focus on the probability of taking the correct decision instead, rather than on the judge's specific payoff function (which weighs taking the correct decision potentially differently in different eventualities).

[24] For $b < \frac{1}{2+\alpha}$ we similarly get:

$$1-(1-b)\,p^G = \underbrace{(1-b)-b\alpha}_{\text{prior}} + \underbrace{b\,(1+\alpha)-(1-b)\,p^G}_{\text{statements per se}} + \underbrace{\sigma_P q_I^P\left[r_P - b^P - b^P(1-r_P)\alpha\right]}_{\text{verification}} - \sigma_P q_I^P c$$

Again this gives the exact same comparative statics as for the case $b > \frac{1}{2+\alpha}$ discussed in the main text.

higher. An increase in the obstruction penalty π_P has no overall net benefit to the judge. It does not impact the (value of) information the statements per se reveal. And because a higher π_P induces the judge to verify less often (lower q_I^P), actually *less* (valuable) information is obtained from verification. This is counterbalanced by incurring the costs of verification equally less often. The overall effect is again that (relatively) less information is obtained from verification. Finally, a decrease in c makes that also (relatively) less information is obtained from verification.

In short, if the verification process becomes more effective – either because it becomes more reliable via r_P, potentially more harmful to the suspect via π_P, or less costly through c – the additional benefits from this greater effectiveness in equilibrium I_{PV} solely come from the deterrence effect. The actual verification process itself actually yields less valuable information. Moreover, an improvement in reliability r_P is more conducive to the overall amount of information provision than an (effectively immaterial) increase in the obstruction penalty π_P is.

5 Model extensions

In this subsection we discuss two extensions that add additional realism to the model, viz. (i) incorporating the possibility of plea bargaining, and (ii) accounting for a right to silence. The overall conclusion that follows from the discussion is that these extensions leave the main insights obtained from our basic setup largely unaffected.

5.1 Plea bargaining

In practice, a very high percentage of cases – up to 95%, see US Bureau of Justice Statistics (2003) – never reach the court room and is settled through some sort of plea bargaining. In this case, the prosecutor offers a penalty reduction in exchange for the suspect pleading guilty. In the literature, plea bargaining has been studied as having (among other things) an informational role in the screening of suspect types.

To incorporate this realistic element in our setup, we allow a third option to the suspect: besides making either a vague or precise statement and the case going to court, he can also choose to Confess and immediately get a penalty of $1 - m$, where $0 < m < 1$ denotes the (exogenously given) penalty reduction offered by the prosecutor.[25] A direct implication of this added choice option is that in equilibrium both suspect types earn at least an expected payoff of $m > 0$, i.e. the payoff the suspect would obtain by accepting the plea offer. From Table 2 it follows that pooling equilibria C_V and C_P then no longer exist. These are replaced by a pooling equilibrium in which both types confess and thus both take the settlement offer (and have expected payoffs equal to $m > 0$).

[25] Although in practice the prosecutor may have some discretion in the size of the penalty reduction offered, this discretion may be considerably restricted by binding guidelines, see e.g. the 2017 ”Reduction in sentence for a guilty plea: Definitive guideline” from the sentencing council in the UK (UK Sentencing Council, 2017). Existing game theoretical models of plea bargaining typically allow the prosecutor to endogenously choose the penalty reduction; qualitatively this leads to the same conclusions with respect to amount of information revelation in equilibrium, see the discussion below.

The informative equilibria – in which the judge obtains at least some influential information – are also affected. Existence of equilibria I_i for $i = P, V$ now requires that the penalty reduction is not too large: formally $m < 1 - r_i(1 + \pi_i)$ is now needed. With respect to the partially pooling equilibria, the guilty type now mixes between Confess (thereby getting m) and the statement i made by the innocent type, which can either be $i = V$ or $i = P$ (hence now two types of partially pooling equilibria exist). Statement i induces the judge to investigate with probability ${q_I}^i = \frac{1-m}{r_i(1+\pi_i)}$ and to acquit otherwise. Existence requires $m > 1 - r_i(1 + \pi_i)$, which corresponds to $\pi_i > \frac{1-r_i-m}{r_i}$. Effectively, the penalty reduction m substitutes for the obstruction penalty π_i in facilitating a partially pooling equilibrium. For the guilty type to be willing to not always mimic the innocent type, one can either make mimicking less attractive (i.e. a higher obstruction penalty π_i), or otherwise make the alternative of not mimicking more attractive (which is essentially what the plea offer and corresponding penalty reduction does). In Section 4 we observed that in the partially pooling equilibrium the threat of investigation and the obstruction penalty work in tandem. Given that the obstruction penalty and the penalty reduction after confession are essentially two sides of the same coin, a similar observation holds for penalty reduction m.

In an early game theoretic analysis of plea bargaining, Grossman and Katz (1983) showed that – if the prosecutor could *commit* to proceed to court if the plea offer is rejected – the plea offer can be used as a screening device to fully separate the guilty types from the innocent ones. A similar observation was made by Reinganum (1988) when extending the framework of Grossman and Katz (1983) by assuming that the prosecutor has private information regarding the strength of the case. Baker and Mezzetti (2001) have challenged this equilibrium separation possibility, as the underlying commitment on which it is based ”...is inherently noncredible because any defendant that the prosecutor knows for sure is innocent will never stand trial” Baker and Mezzetti (2001, p. 151). Models that drop this possibility to fully commit to go to trial all find that plea bargaining is (at most) essentially semi-separating, with the plea offer accepted by the guilty type with some probability but rejected by the innocent type for sure.[26] In this equilibrium, the prosecutor still proceeds to trial with probability one if the plea offer is rejected (and the prosecutor did not get a strong external signal of the suspect's innocence).[27] The partially pooling equilibrium in our setup is qualitatively similar in terms of the suspect's behavior (viz. only the guilty type mixes), yet differs in the behavior of the representative of the judicial system; in our equilibrium I_{PV} the judge is using a strictly mixed strategy in equilibrium as well. The latter makes that if the obstruction penalty increases or, alternatively, the penalty reduction increases, only the judge adapts her behavior by reducing the occurrence of investigation, while leaving the behavior of the two suspect types unaffected. Like an increase

[26] See Baker and Mezzetti (2001); Bjerk (2007); Kim (2010); Tsur (2017). A remaining criticism of some of these models is that the behavior of the judge/jury is assumed to be purely exogenous and does not react to (the information revealed by) the behavior of the prosecutor and the suspect. This arguably provides another unrealistic commitment possibility, viz. to a mechanical conviction rule. Bjerk (2007) and Tsur (2017) endogenize the behavior of the judge/jury and obtain the same type of semi-separating equilibrium (though a multiplicity of these may exist). Note that our simplified setup with a unitary judiciary actor essentially corresponds to the case where different representatives of the judiciary share the same information and beliefs, and endogenously act on these; the probability of conviction is thus entirely the result of equilibrium strategies.

[27] Note that this is now based on an equilibrium best response rather than on an ex ante commitment as in the earlier papers.

in π_P, a higher penalty reduction m thus has no overall benefit to the judge in terms of getting additional valuable information (cf. Subsection 4.3).[28]

5.2 Right to silence

In our baseline model, the judge can use both the strategic behavior of the suspect as well as the outcome of the potential investigation to update her belief about the suspect's innocence and act accordingly without any restrictions. In particular, the suspect's choice of making a vague statement can in principle be fully held against him and lead to immediate conviction. Notably this occurs in the partially pooling equilibrium, in which only the guilty type now and then makes a vague statement and Bayesian updating thus induces the judge to convict after observing such a statement. Traditional common law systems, however, typically give the suspect the 'right to remain silent'; if a suspect refuses to answer any question, the verdict must solely be based on other evidence and the suspect's silence cannot be considered evidence of his guilt (*Miranda v.Arizona*, 1966). Effectively, this right thus works as a commitment to ignore some of the suspect's strategic information revelation.

In our stylized setup, prior belief b can be interpreted as the evidence that the suspect is innocent collected by the judge before any statement is made. A right to silence then can be incorporated in our model by requiring that the judge choice of action after a vague statement (cf. Lemma 2) should be guided by a restricted posterior belief $b^V_{res} = b$, rather than by Bayesian belief b^V that follows from equation (2). That is, although the judge may rationally infer that actually $b^V = 0$ (as e.g. occurs in equilibrium I_{PV}), she cannot act on that and should act on $b^V_{res} = b$ instead. This of course affects the judge's choice of action and, in turn, may affect the statement made by the suspect. With a right to silence, the pooling on $j = V$ equilibria remain unaffected, and similarly so equilibria C_P and A_P.[29] Equilibria I_P and I_{PV} are affected, however, as the choice to immediately convict after a vague statement is only in line with a right to silence in case $b < \underline{b}(r_V)$, i.e. if just the prior belief would already be sufficient for immediate conviction (cf. Lemma 2). For higher prior beliefs $b > \underline{b}(r_V)$ equilibrium I_P ceases to exist. In that case equilibrium I_{PV} is affected as follows. For $b \in (\underline{b}(r_V), \overline{b}(r_V))$, the suspect's behavior stays exactly the same, but now the judge always investigates after a vague statement. Moreover, a precise statement then leads to investigation with probability $q^P_I = \frac{r_V(1+\pi_V)}{r_P(1+\pi_P)}$, i.e. less frequently than the investigation rate of $\frac{1}{r_P(1+\pi_P)}$ as in the baseline model. This also makes that there are no further restrictions on π_P and a minimum obstruction penalty π_P is no longer required (given the other assumptions made). For $b > \overline{b}(r_V)$ equilibrium I_{PV} no longer exists.[30]

[28] Although the obstruction penalty and the penalty reduction play a similar deterrence role in incentivizing the guilty type to sometimes either implicitly (via a vague statement) or explicitly confess, their payoff implications for the suspect are quite different. Clearly both the guilty and the innocent type are better off with higher penalty reductions than with higher obstruction penalties. (In the presence of plea bargaining, the guilty type gets an expected payoff of m and the innocent type earns $1 - q^P_I(1 - r_P)(1 + \pi_P) = 1 - \frac{(1-m)(1-r_P)}{r_P}$.) From a broader social welfare perspective, however, society might dislike penalty reductions as they allow offenders to largely 'get away with it' and rather prefer penalties for obstruction (Fagan, 1981; Cohen and Doob, 1989; Herzog, 2003; Johnson, 2019). A reduced form way to incorporate such broader considerations in our model would be to let the judge's expected payoffs depend on m (and π_P) as well.

[29] This holds because, if $j = V$ is on the equilibrium path, Bayesian beliefs $b^V = b$ equal b^V_{res}. Equilibria C_P and A_P are sustainable with various out of equilibrium beliefs, including $b^V = b$ that honors a right to silence (cf. Subsection 4.2).

[30] With a right to silence, fully separating equilibria now do exist, but these are outcome-equivalent to pooling.

The shifts in the informative equilibria I_P and I_{PV} are in line with the effects of a right to silence identified by the game theoretic analyses of Seidmann (2005) and Leshem (2010) (see also Seidmann and Stein (2000) and Mialon (2005)). In particular, the innocent type benefits from such a right in two ways. A first, direct benefit is that it provides ”innocent suspects, who are otherwise compelled to speak, with the alternative of silence” (Leshem, 2010, page 400). In our setup this effect is reflected by the non-existence of the informative equilibria I_P and I_{PV} in case $b > \overline{b}(r_V)$; then only pooling equilibria in which the judge always acquits remain. With a sufficiently high prior belief that the suspect is innocent, the judge is compelled to acquit in the absence of further information (cf. Lemma 2). A right to silence then provides the innocent type a safe alternative (viz. silence) to making a precise statement, as with the latter the innocent type always runs the risk of his precise statement being wrongly falsified. A second, indirect benefit is that innocent types who choose to make a precise statement are less likely to be wrongfully convicted. This effect is exemplified by the reduced probability that the judge checks on a precise statement in the altered partially pooling equilibrium when $b \in (\underline{b}(r_V), \overline{b}(r_V))$, i.e. $q_I^P = \frac{r_V(1+\pi_V)}{r_P(1+\pi_P)} < \frac{1}{r_P(1+\pi_P)}$. Most important for our purposes, however, is the observation that strategic information revelation still plays an important role in affecting the judge choice of action and continues to be complementary to the judge now and then checking on messages; the qualitative features of the partially pooling equilibrium are robust to introducing a right to silence. Moreover, it confirms the attractiveness of improvements in reliability relative to increases in the obstruction penalty, because with a right to silence there is no longer a minimum required π_P. Overall, therefore, the verifiability approach to lie detection does not lose its bite in the presence of a right to silence.

6 Conclusion

In this paper, we analyze the strategic interaction between a speaker who wants to convince an investigator of his innocence and an investigator who wants to know the truth, i.e. whether the speaker is guilty or innocent. In our model, the investigator can check the specific details in the statement of the speaker at some cost. This yields informative, but imperfect evidence. The more detailed the speaker's statement is, the more reliable the examination of this statement becomes. This encourages innocent speakers to be forthcoming in providing many verifiable details in their statement, while guilty types would prefer to remain vague. If, on the basis of an investigation, the investigator concludes that the speaker is lying, an additional obstruction penalty is imposed on the speaker.

Our results reveal the circumstances under which a partially pooling equilibrium exists. In this equilibrium, the guilty type mixes between making a vague and making a precise statement, whereas the innocent type makes a precise statement for sure. Precise statements are now and

To illustrate, if $b > \overline{b}(r_V)$ an equilibrium exists in which the guilty type always makes a vague statement, while the innocent type always makes a precise statement. With $b^V_{res} = b > \overline{b}(r_V)$ the judge is then, given the right to silence, forced to acquit (cf. Lemma 2), whereas after a precise statement it is simply a best response to do so. This fully separating equilibrium is thus outcome equivalent to the pooling equilibria A_V and A_P. Despite the fact that the suspect's strategic behavior may be very informative, a right to silence may prohibit this information becoming influential. This also alters the perspective on what valuable information – and the source from which it is obtained – effectively entails (cf. Subsection 4.3).

then investigated by the investigator to verify their veracity. In the partially pooling equilibrium, verification and strategic information revelation by the speaker thus go hand in hand.

Our analysis allows us to understand the behavioral patterns observed for lie detection methods. It explains the shortcomings of the early approaches that were based on a speaker's micro-expression of emotional cues that do not convey sufficient reliable information. In particular, our model explains why when the observer's investigation is not sufficiently reliable as only pooling can be sustained in equilibrium. In such cases, guilty types mimic innocent types and nothing can be inferred from the strategic behavior of the speaker. For recent advances with the verifiability approach, the picture is more promising. By judging the frequency of precise verifiable details in a speaker's statement, more reliable information is acquired. In such settings, our analysis shows that a partial pooling equilibrium is most plausible. Standard equilibrium refinements typically favor this equilibrium over other, less informative (pooling) equilibria that may co-exist and in which the speaker does not engage in strategic information revelation. This equilibrium agrees with empirical observations, in which innocent types furnish their statements with precise, verifiable details, whereas guilty types face a difficult trade-off that they solve by sometimes imitating the innocent types and by remaining vague at other times.

Our analysis also offers some insights that go beyond what has been observed in the recent literature on lie detection. One insight is that partial pooling is the best that a method can accomplish. Complete separation of the types can never be sustained in equilibrium. Still, there are ways to improve on lie detection methods that allow for partial pooling. The overall amount of information provision in the partially pooling equilibrium is especially facilitated by an improved reliability of the verification technology. This renders verification more informative per se and (thus) makes the investigator more willing to investigate. Realizing this, the guilty type reduces the likelihood with which he makes a precise statement, in turn providing the investigator actually less incentives to investigate. The overall net effect is that, when reliability improves, more can be learned from the strategic behavior of the speaker and actually less is learned via actual verification. Surprisingly, not much is accomplished by enhancing the obstruction penalty. Once the obstruction penalty surpasses a critical threshold, it has no further impact on the deterrence effect of a lie detection method. An increase in the obstruction penalty also leads the investigator to investigate less, but leaves the amount of strategic information revelation unaffected. The investigator – and thus also "truth" – is better served by an improved reliability of the verification technology.

We also discuss two extensions of our model. In a first extension we enlarge the speaker's options with the possibility to confess in exchange for a penalty reduction (i.e. a form of plea bargaining). Intuitively, this penalty reduction plays a role similar to the obstruction penalty; it makes it less attractive for the guilty type to pretend being innocent by making a precise statement. Second, we discuss the implications of a right to silence in the context of our model. Such a right puts limits on the extent to which information revealed by the strategic choices of the speaker can be taken into account for the investigator's choice of action. Although such a right diminishes the role for strategic information revelation when the investigator is initially inclined to acquit, this role essentially remains the same when this is not the case. Moreover, it reinforces the attractiveness of improvements in reliability, as the obstruction penalty no longer serves a supportive role and making it harsher does not help. Even with a right to silence, lie detection via the verifiability approach thus continues to have a strong bite.

Appendix A: Proofs

This Appendix contains the proofs of Lemma 2 and of Propositions 1, 2 and 4.

Proof of Lemma 2

Proof. Assume the suspect made statement j and beliefs equal b^j. If the judge acquits without investigating, she gets an expected payoff of $1 \cdot b^j + 0 \cdot (1 - b^j) = b^j$. Similarly so, if she convicts without investigating, she gets $1 \cdot (1-b^j) - \alpha \cdot b^j = 1 - (\alpha+1)b^j$ in expectation. Comparing these two expected payoffs, the judge prefers conviction over acquittance if $b^j < \frac{1}{2+\alpha}$, and acquittance over conviction if $b^j > \frac{1}{2+\alpha}$. Given that an investigation is costly ($c > 0$), the judge is only willing to investigate if it is influential (cf. Lemma 1); it then leads to an expected payoff of $r_j - b^j(1-r_j)\alpha - c$. This exceeds the payoff of acquitting (viz. b^j) if $b^j < \frac{r_j - c}{\alpha(1-r_j)+1}$ and it exceeds the payoff of convicting (viz. $1 - (\alpha+1)b^j$) if $b^j > \frac{(1-r_j)+c}{\alpha r_j + 1}$. For the first threshold it holds that $\frac{r_j - c}{\alpha(1-r_j)+1} \geq \frac{1}{2+\alpha}$ iff $c \leq \frac{\alpha+1}{\alpha+2} \cdot (2r_j - 1)$. Similarly, for the second threshold it holds that $\frac{(1-r_j)+c}{\alpha r_j+1} \leq \frac{1}{2+\alpha}$ iff $c \leq \frac{\alpha+1}{\alpha+2} \cdot (2r_j - 1)$. Hence, if c is below this latter threshold, the interval $(\underline{b}(r_j), \overline{b}(r_j))$ is non-empty and on this interval the judge prefers (influential) investigation over both immediate conviction and immediate acquittance. □

Proof of Proposition 1

Proof. Suppose both suspect types pool on statement j. By Bayes' rule then $b^j = b$ and the judge's optimal response to j is given by Lemma 2. To support this as equilibrium behavior, statement $-j$ should yield both types weakly less than the judge's response to j. From our assumption that $r_V(1+\pi_V) < 1$, it follows that the guilty type can always secure a payoff of at least zero by choosing a vague statement; in case of influential investigation such a statement would yield the guilty type $1 - r_V(1+\pi_V) > 0$, and it would yield 0 or 1 in case of direct conviction or immediate acquittance respectively. The only instance in which a precise statement yields the guilty type less than 0 is when such a statement is investigated (cf. Lemma 2) and $1 - r_P(1+\pi_P) < 0$, i.e. when $\pi_P > \frac{1-r_P}{r_P}$. In that case there are no out-of-equilibrium beliefs b^V that support pooling on precise (because at least the guilty type would always like to deviate). In all other instances, both types get (weakly) more than zero on the purported equilibrium path, and skeptical out-of-equilibrium beliefs $b^{-j} = 0$ thus make deviating to $-j$ unattractive. □

Proof of Proposition 2

Proof. We first show – in four steps – that the general nature of the strategic information revelation in any partially pooling equilibrium follows the general description in parts (a) and (b). We subsequently derive the expressions for the stated mixing probabilities and thresholds.

(Step 1). *The judge investigates at least one of the two statements sometimes, i.e.* $q_I^j > 0$ *for some* j. Suppose the judge never chooses to investigate, i.e. $q_I^V = q_I^P = 0$. Both suspect types then will strictly prefer the statement that has the highest probability of acquittance and are only willing to mix between statements iff $q_A^V = q_A^P$, i.e. in a pooling equilibrium (recall that we label equilibria in which the judge's decision is invariant to the statement received as pooling).

(Step 2) *The innocent type does not mix in equilibrium.* Suppose to the contrary that the innocent type mixes in equilibrium. If the guilty type would not mix as well, one of the two statements would be self-revealing as coming from the innocent type and thus induce immediate acquittance. For the innocent type then to be willing to mix, so should the other statement shared with the guilty type. But this implies that the judge's behavior is invariant to the statement made, which can only occur in a pooling equilibrium. Therefore, if the innocent type were to mix in a partially pooling equilibrium, necessarily the guilty type must do so as well. In that case both types should be indifferent between the two statements. Now suppose one of the two statements ($j = m$ say) is never investigated (i.e. $q_I^m = 0$). From step 1 we know that the other statement ($j = n$ say) necessarily sometimes is ($q_I^n > 0$). By the assumed payoff structure, both suspect types get exactly the same expected payoff after the unchecked message m, while the expected payoff of the innocent type is strictly higher than the expected payoff of the guilty type after message n (the latter holds because in case of investigation, which only happens if it is influential, the innocent type gets a strictly higher expected payoff than the guilty type). So if both types are to be indifferent, both statements should be checked with positive probability.

From Lemma 2, after having received statement j, the judge will generically not mix between all three options A, C and I at the same time.[31] Moreover, the judge will also not mix between the exact same two actions (either between A and I, or between C and I) after both statements. From Lemma 2 this would imply that $b^V = b^P$, which can only happen in a pooling equilibrium. Therefore, the judge should mix between A and I after one statement and between C and I after the other statement. Suppose she mixes between A and I after a vague statement. Then the guilty type prefers vague for sure as both potential payoffs (1 after acquit and $1 - r_V(1 + \pi_V)$ after investigating a vague statement) dominate the two potential payoffs after a precise statement (viz. 0 after convict and $1 - r_P(1 + \pi_P)$ after investigating a precise statement), given our assumption that $r_V(1 + \pi_V) < \max\{1, r_P(1 + \pi_P)\}$. Hence, the only remaining relevant situation is where after a vague statement the judge mixes between C and I, and after a precise statement between A and I. For both suspect types then to be willing to mix we must have that the two statements yield equal expected payoffs to them, i.e.:

$$q_I^V[r_V - (1 - r_V)\pi_V] = q_I^P[r_P - (1 - r_P)\pi_P] + (1 - q_I^P)$$
$$q_I^V[-r_V\pi_V + 1 - r_V] = q_I^P[-r_P\pi_P + 1 - r_P] + (1 - q_I^P)$$

This system does not have a solution where $0 < q_I^V, q_I^P < 1$.[32] Hence, the innocent type will

[31] This would require both $b^j = \frac{1}{2+\alpha}$ and $c = r_j - \frac{1}{2+\alpha} \cdot [\alpha(1 - r_j) + 1]$, and thus can only happen in the knife-edge case where $c = \frac{(\alpha+1)(2r_j-1)}{(\alpha+2)}$. In our analysis we exclude such non-generic knife-edge cases.

[32] The solution would require $q_I^V = \frac{(2r_P-1)}{(2r_p-1)-(1+\pi_V)(r_P-r_V)}$, which exceeds 1.

never mix in a partially pooling equilibrium.

(Step 3) *The guilty type mixes between vague and precise, while the innocent type chooses precise for sure.* From step 2 we know that the guilty type necessarily mixes between the two statements in a partially pooling equilibrium, while the innocent type chooses one statement for sure. Suppose the innocent type chooses vague for sure. Then observing a precise statement is conclusive information that the suspect is of the guilty type, thus inducing immediate conviction. Because acquittance and (given our assumption that $r_V(1 + \pi_V) < 1$) investigation of a vague statement would yield the guilty type strictly more than conviction, for the guilty type to be willing to mix a vague statement should lead to immediate conviction as well. But this corresponds to a pooling equilibrium.

(Step 4) *A vague statement induces immediate conviction, while after a precise statement the judge mixes between acquittance and investigation.* From step 3 it follows that a vague statement induces belief $b^V = 0$ and thus immediate conviction. This gives the guilty type a payoff equal to 0. For the guilty type to be willing to mix, making a precise statement should then yield him the same. From (the proofs of) steps 1 and 2 it follows that after a precise statement the judge mixes between either C and I or between A and I. Suppose the former case applies. Then indifference of the guilty type would require $0 = q_I^P(1 - r_P - r_P\pi_P)$; this can only hold in the non-generic knife edge case $\pi_P = \frac{1-r_P}{r_P}$. Excluding this non-generic case, the judge should mix between A and I after a precise statement.

Steps 1 to 4 characterize the general nature of the unique partially pooling equilibrium. The precise mixing probability of the guilty type follows from making the judge indifferent between investigating and acquittance after having received a precise statement:

$$r_P - b^P(1 - r_P)\alpha - c = \max\left\{b^P, 1 - (\alpha + 1)b^P\right\} = b^P$$
$$\Rightarrow b^P = \frac{r_P - c}{\alpha(1 - r_P) + 1} \Rightarrow p^G = \frac{b}{1 - b} \cdot \frac{(1 - r_P)(1 + \alpha) + c}{r_P - c}$$

Here the second equality reflects the requirement that after a precise statement the judge should prefer acquit over convict (for otherwise she would not want to choose A with positive probability), i.e. $b^P > 1 - (\alpha + 1)b^P$ and thus $b^P > \frac{1}{2+\alpha}$. Similarly, q_I^P follows from the required indifference of the guilty type:

$$0 = (1 - q_I^P) + q_I^P[(1 - r_P) - r_P\pi_P] \Rightarrow q_I^P = \frac{1}{r_P(1 + \pi_P)}$$

The necessary and sufficient conditions for existence in part (c) follow from the requirements that $q_I^P < 1$, $b^P > \frac{1}{2+\alpha}$, and $b < b^P$ (or, equivalently, $p^G < 1$). These correspond respectively to $\pi_P > \frac{1-r_P}{r_P}$, $c < \frac{1+\alpha}{2+\alpha} \cdot (2r_P - 1)$, and $b < \frac{r_P - c}{\alpha(1-r_P)+1}$. □

Proof of Proposition 4

Proof. Consider equilibrium I_V. The set of potential best responses to an out-of-equilibrium precise statement is listed in Lemma 2. In case the judge would convict after a precise statement for sure, neither type would like to deviate, while in case she would acquit for sure after a precise statement, both types would want to deviate. If the judge would investigate a precise statement, or would mix between conviction and investigation (which is a best response for belief $b^P = \underline{b}(r_V)$), the guilty type would never want to deviate, while the innocent type might have an incentive to do so (viz. in case $(1-r_P)(1+\pi_P) < (1-r_V)(1+\pi_V)$). Finally, if the judge mixes between I and A after a precise statement (which is a best response when $b^P = \overline{b}(r_P)$), then the guilty type wants to deviate whenever $q_I^P \leq \frac{r_V(1+\pi_V)}{r_P(1+\pi_P)}$ and the innocent type whenever $q_I^P \leq \frac{(1-r_V)(1+\pi_V)}{(1-r_P)(1+\pi_P)}$. From $\frac{r_V}{r_P} < \frac{1-r_V}{1-r_P}$ it follows that the set of best responses for which the guilty type wants to deviate is a strict subset of the set of best responses for which the innocent type wants to deviate. Divinity thus requires $b^P = 1$ and thus acquittance after a precise statement, making that both types would like to deviate from their supposed vague statement. Hence I_V is not divine.

That equilibrium I_P is divine follows from the fact that (similar to the reasoning above) the guilty type has a stronger incentive to deviate to a vague message than the innocent type has. Finally, I_{PV} contains no out-of-equilibrium statements and thus restrictions on out-of-equilibrium beliefs have no bite. □

Appendix B: Richer set of possible statements

From a practical point of view, our binary setup where the suspect can answer either all or none of the questions appears rather restrictive. In real life suspects can always choose to answer some, but not all of the questions. A natural extension of our baseline model is thus to allow for statements with intermediate levels of precision.

Suppose the suspect can choose to answer $j \in \{0, 1, \ldots, N\}$ questions, resulting in a statement labelled j. Checking on statement j has a reliability level equal to r_j. In line with our basic setup, we assume that the reliability probabilities satisfy $0.5 \leq r_0 < r_1 < \cdots < r_N < 1$. As before, if convicted after the suspect's statement is falsified, an additional obstruction penalty π_j is imposed. These are again assumed to be such that the guilty type fears investigation of a less precise statement less than investigation of a more precise statement: $r_i(1+\pi_i) < r_k(1+\pi_k)$ for all $i < k$ weakly below N. And finally, like before, we also assume that $r_0(1+\pi_0) < 1$, such that investigation of a vague statement is better for the guilty type than immediate confession.[33]

The main driving forces that underlie the equilibrium analysis of the baseline setup continue to apply when multiple statements can be made. In particular, for the judge's behavior Lemma 1 and Lemma 2 immediately carry over to the extended setup with multiple statements. That is, investigating statement j is influential only if, based on prior information, the judge is still insufficiently convinced, i.e. if belief b^j is not too extreme (cf. Lemma 1). This makes that the judge chooses to convict when b^j is low, to acquit when b^j is high, and to investigate when belief b^j is in between. The cutoffs that define the different ranges are just as in Lemma 2. In regard to the behavior of the suspect it again holds that a separating equilibrium does not exist. Conceptually, again thus only two types of equilibria remain. In a pooling equilibrium both suspect types make the same statement j. In line with Proposition 1, such an equilibrium

[33] Obviously, our basic setup corresponds to the special case with $N = 1, r_0 = r_V, r_N = r_P, \pi_0 = \pi_V$ and $\pi_N = \pi_P$.

exists for any prior belief b if $\pi_j < \frac{1-r_j}{r_j}$. Otherwise, in case $\pi_j > \frac{1-r_j}{r_j}$ a pooling on statement j equilibrium only exists iff $b \notin (\underline{b}(r_j), \overline{b}(r_j))$. The assumption that $r_j(1+\pi_j)$ is increasing in j thus implies that, the more precise a statement is, the less likely it is that the two suspect types pool on making this particular statement. The reason is that for $r_j(1+\pi_j) > 1$, the guilty type rather gets convicted for sure than having statement j investigated; the obstruction penalty π_j together with the high reliability level r_j of the verification technology makes mimicking the innocent type unattractive.

With multiple statements, multiple partially pooling equilibria exist side by side. But these all share the same structure as described in Proposition 2. That is, the innocent type chooses some statement $i > 0$ for sure, while the guilty type mixes between statement i made by the innocent type and some other statements $g \neq i$. The latter all lead to immediate conviction, while after statement i the judge mixes between acquittance and investigation. Existence among others requires that $r_i(1+\pi_i) > 1$, as to make the (probabilistic) investigation of this statement a sufficient deterrent for the guilty type.[34] The expected payoffs players earn in these partially pooling equilibria are given by row I_{PV} in Table 2 when we replace r_P by r_i. The larger r_i, the higher the expected payoffs for the judge and the innocent type are, while the guilty type is unaffected. The partially pooling equilibrium in which the innocent type chooses the most precise statement N thus weakly Pareto dominates the other partially pooling equilibria in which the innocent type chooses some $i < N$ when these co-exist. Moreover, the circumstances under which partially pooling may occur are largest for the case where the innocent type makes the most precise statement N. In line with the reasoning in Subsection 4.2, this arguably makes this partially pooling equilibrium focal within the overall set of partially pooling equilibria.

With respect to the pooling equilibria in which the judge investigates, a similar result as in Proposition 4 holds. In particular, all equilibria I_i for $i < N$ do not survive the Divinity criterion, only equilibrium I_N does. Among the equilibria in which some information is obtained on the equilibrium path, the innocent type thus only makes the most precise statement N in the most plausible among these equilibria. As a result, the structure of (plausible) equilibria remains essentially unaffected by the inclusion of intermediate precision levels.

References

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

Banks, J. S. and Sobel, J. (1987). Equilibrium selection in signaling games. *Econometrica*, 55(3):647–661.

Bell, B. E. and Loftus, E. F. (1989). Trivial persuasion in the courtroom: The power of (a few) minor details. *Journal of Personality and Social Psychology*, 56(5):669–679.

Bjerk, D. (2007). Guilt shall not escape or innocence suffer? the limits of plea bargaining when defendant guilt is uncertain. *American Law and Economics Review*, 9(2):305–329.

[34] The exact mixing probabilities and the conditions under which a particular partially pooling equilibrium exists immediately follow from Proposition 2 by replacing r_P and π_P by r_i and π_i, respectively.

Bond Jr, C. F. and DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3):214–234.

Cohen, A. (1992). *The Living Law: A Guide to Modern Legal Research.* Rochester, N.Y.: Lawyers Cooperative.

Cohen, S. A. and Doob, A. N. (1989). Public attitudes to plea bargaining. *Criminal Law Quarterly*, 32(1):85–109.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6):1431–1451.

Daubert v. Merrell Dow Pharmaceuticals, Inc. (1993). 509 U.S. 579-601.

Decker, J. F. (2004). The varying parameters of obstruction of justice in american criminal law. *Louisiana Law Review*, 65(1):49–130.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., and Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, 129(1):74–118.

Ekman, P., Friesen, W. V., and O'sullivan, M. (1988). Smiles when lying. *Journal of Personality and Social Psychology*, 54(3):414–420.

Fagan, R. W. (1981). Public support for the courts: An examination of alternative explanations. *Journal of Criminal Justice*, 9(6):403–417.

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

Harsanyi, J. C. and Selten, R. (1988). *A general theory of equilibrium selection in games.* The MIT Press.

Harvey, A. C., Vrij, A., Nahari, G., and Ludwig, K. (2017). Applying the verifiability approach to insurance claims settings: Exploring the effect of the information protocol. *Legal and Criminological Psychology*, 22(1):47–59.

Herzog, S. (2003). The relationship between public perceptions of crime seriousness and support for plea-bargaining practices in israel: A factorial survey approach. *Journal of Criminal Law and Criminology*, 94(1):103–132.

Johnson, T. (2019). Public perceptions of plea bargaining. *American Journal of Criminal Law*, 46(1):133–156.

Kim, J.-Y. (2010). Credible plea bargaining. *European Journal of Law and Economics*, 29(3):279–293.

Kleinberg, B., Nahari, G., and Verschuere, B. (2016). Using the verifiability of details as a test of deception: A conceptual framework for the automation of the verifiability approach. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 18–25.

Leshem, S. (2010). The benefits of a right to silence for the innocent. *RAND Journal of Economics*, 41(2):398–416.

Mialon, H. M. (2005). An economic theory of the fifth amendment. *RAND Journal of Economics*, 36(4):833–848.

Milgrom, P. (2008). What the seller won't tell you: Persuasion and disclosure in markets. *Journal of Economic Perspectives*, 22(2):115–131.

Miranda v.Arizona (1966). 384 U.S. 436-545.

Nahari, G., Vrij, A., and Fisher, R. P. (2014). The verifiability approach: Countermeasures facilitate its ability to discriminate between truths and lies. *Applied Cognitive Psychology*, 28(1):122–128.

Reinganum, J. F. (1988). Plea bargaining and prosecutorial discretion. *The American Economic Review*, 78(4):713–728.

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

Seidmann, D. J. and Stein, A. (2000). The right to silence helps the innocent: A game-theoretic analysis of the fifth amendment privilege. *Harvard Law Review*, 114(2):430–510.

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

Sorochinski, M., Hartwig, M., Osborne, J., Wilkins, E., Marsh, J., Kazakov, D., and Granhag, P. A. (2014). Interviewing to detect deception: when to disclose the evidence. *Journal of Police and Criminal Psychology*, 29(2):87–94.

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, 87(3):355–374.

Tsur, Y. (2017). Bounding reasonable doubt: implications for plea bargaining. *European Journal of Law and Economics*, 44(2):197–216.

UK Sentencing Council (2017). *Reduction in sentence for a guilty plea.* United Kingdom Department of Justice. https://www.sentencingcouncil.org.uk/wp-content/uploads/Reduction-in-Sentence-for-Guilty-Plea-definitive-guideline-SC-Web.pdf.

US Bureau of Justice Statistics (2003). *Sourcebook of criminal justice statistics.* United States Department of Justice. https://www.ncjrs.gov/pdffiles1/Digitization/208756NCJRS.pdf.

US Sentencing Commission (2018). *Guidelines Manual.* United States Department of Justice. https://www.ussc.gov/sites/default/files/pdf/guidelines-manual/2018/GLMFull.pdf.

Vrij, A. (2000). *Detecting lies and deceit: The psychology of lying and implications for professional practice.* Wiley.

Vrij, A. (2018). Verbal lie detection tools from an applied perspective. In *Detecting concealed information and deception*, pages 297–327. Elsevier.

Vrij, A. (2019). Deception and truth detection when analyzing nonverbal and verbal cues. *Applied Cognitive Psychology*, 33(2):160–167.

Vrij, A., Fisher, R. P., and Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, 22(1):1–21.

Vrij, A., Mann, S., Kristen, S., and Fisher, R. P. (2007). Cues to deception and ability to detect lies as a function of police interview styles. *Law and Human Behavior*, 31(5):499–518.

Zuckerman, M., DePaulo, B. M., and Rosenthal, R. (1981). Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology*, 14(1):1–59.