Interrogation and Disclosure of Evidence

Jesse Bull*

Revised May 2026

Abstract

In recent years there has been a shift to less confrontational interrogation methods. One argument in favor of these methods is that they reduce the scope for false confessions. This paper explores this idea from an information perspective with rational agents. It studies a model that extends the plea-bargaining model of Baker and Mezzetti (2001) to one that is suitable for analyzing interrogation methods with evidence implying a continuum of probabilities that the suspect is found guilty at trial. In a benchmark model, in which police only interrogate suspects that the prosecutor is potentially willing to take to trial, the non-confrontational method weakly dominates the confrontational method. A similar result holds in a richer model, in which police may interrogate some suspects for whom the evidence will not induce the prosecutor to go to trial. The key intuition for why the non-confrontational method performs better in some cases in both settings is that the non-confrontational method is able to induce a confession from some guilty who the prosecutor would not take to trial following non-confession, and for some distributions this can result in a strictly higher expected payoff for society and the prosecutor.

1 Introduction

Viewers of television shows about police often see portrayals of police officers aggressively questioning a suspect and early on presenting incriminating evidence in a confrontational

*Florida International University. Email: bullj@fiu.edu. This paper began as a result of a related paper titled "Interrogation and Evidence Fabrication." I'm especially grateful to Scott Baker (the Editor in charge of this article) for many thoughtful and detailed comments, which stimulated significant improvements of this article. I thank two anonymous referees for their thoughtful and detailed comments, which helped to improve this article. I thank Shmuel Leshem for many helpful comments on a much earlier version. I thank Javier Muñoz for helpful comments on a slightly earlier version, and Joel Watson for many helpful comments. I also thank participants at the 2022 SEA Conference and 2023 WATE Conference.

manner. This is typically portrayed as inducing the guilty accused to confess. In North America most of the aggressive interrogation techniques have their basis in what is known as the "Reid Technique." The most recent version of this technique is presented in Inbau et al. (2013).

The Reid Technique focuses on the interrogator first determining whether she believes the accused is guilty. Once she is convinced of the guilt of the accused, she is to use "The Reid Nine Steps of Interrogation©" as described in Inbau et al. (2013). These steps are:

1. Direct, Positive Confrontation
2. Theme Development
3. Handling Denials
4. Overcoming Objections
5. Procurement and Retention of a Suspect's Attention
6. Handling the Suspect's Passive Mood
7. Presenting an Alternative Question
8. Having the Suspect Orally Relate Various Details of the Offense
9. Converting an Oral Confession into a Written Confession

As it suggests, the first step focuses on confrontation. While much of the motivation for the technique focuses on psychological aspects, three features stand out for the modeling at hand. First, police interrogate once they are adequately convinced of the guilt. The second is the focus on confrontation of the accused. Some of the discussion regarding the steps suggests presenting incriminating evidence. Third, this interrogation method's goal is a confession. Throughout Inbau et al. (2013), there are discussions that relate the interrogation exercise to that of a salesperson. So while there is scope for obtaining information from the accused, the confession of someone who is believed to be guilty is a goal.[1] The scope for false confessions is central to the comparison of these various interrogation methods. Much of the criticism of

[1]The Reid team notes that it's possible to interrogate someone and later realize they are not guilty, and recommends not pursuing a confession in that instance.

the Reid technique in the legal psychology literature highlights techniques aimed at inducing confession.[2]

Recently, in Canada, Denmark, New Zealand, and the UK, there has been a transition to a less manipulative and confrontational approach known as PEACE for Preparation and Planning, Engage and Explain, Account, Closure and Evaluate. See, for example Snook, Eastwood, Barron (2014) and Starr (2013). The approach is less manipulative and less confrontational. Instead of focusing on confronting the accused with incriminating evidence, the technique encourages an open minded approach to asking the accused for information. Bull (2018) states:

> The PEACE approach is based on a number of basic principles that were chosen by the police. These include:
>
> 1) "The purpose of investigative interviewing is to obtain accurate and reliable information from suspects, witnesses or victims in order to discover the truth about matters under investigation."
>
> 2) "Interviews should be approached with an open mind. Information obtained from the person who is being interviewed should always be tested against what the investigator already knows or what can reasonably be established."

Advocates of the PEACE approach focus on the scope for obtaining information from the accused and suggest that its use can reduce false confessions. Gudjonsson and Pearse (2011) note, "Several authors have argued that the guilt-presumptive and confrontational processes inherent in the Reid technique should be replaced by the PEACE model or a similar noncoercive technique." Many of the criticisms of the Reid Technique focus on its confrontational approach, manipulation of the suspect, and goal of obtaining a confession. Leo (2008) highlights the focus on obtaining a confession instead of on truth. It's felt that the non-confrontational nature of the PEACE approach is much less likely to pressure a suspect into confessing.

[2]Although it receives little attention in the legal psychology literature, the Reid technique is permissive of fabrication of evidence within legal limits, saying:

> While the courts have consistently upheld the interrogator's use of deceptive evidence ploys, the interrogator should exercise great caution in utilizing them. In general courts recognize the practical necessity in allowing such tactics so long as they do not result in involuntary or false confessions.

This issue is not addressed in this paper, but this paper developed from Bull (2012a), which studied fabrication of evidence by police for interrogation as is legally permissible in the U.S.

In addition to the differences regarding psychological manipulation, there are differences in the information disclosed to the accused. The analysis in this paper is aimed at better understanding the effects of these differences. This is done in the context of a game-theoretic model that extends the model of plea bargaining of Baker and Mezzetti (2001) to an interrogation setting. The model studied here allows for the police officer to offer a reduced sentence for confessing that the prosecutor will approve. While the prosecutor is not typically involved in interrogation, police may imply a reduced sentence in their interaction with the suspect and are likely to have a sense for the reduction the prosecutor provides given the evidence strength and nature of the crime. In that sense, this is modeled here as bargaining between the accused and the prosecutor that occurs "in the shadow of trial."[3] This model is focused on the interaction between the accused and the prosecutor, and requires sequential rationality, including in the prosecutor's decision of whether to go to trial following the accused not confessing. One view of this is that the prosecutor simply provides the police with a schedule of reduced sentences given the strength of the evidence, and is not involved directly with each case at this point.

A key difference from Baker and Mezzetti's model is that the evidence realization implies a probability of the suspect being found guilty at trial denoted by x that is on a continuum; depending on the interrogation method, this can be conveyed to the accused. Further, the probability of conviction at trial that leads to police interrogation may differ from the probability that is sufficient for the prosecutor to go to trial. This fits with reality as charges are commonly dismissed (Gershowitz, 2018). The accused has the following two channels of information: 1) he knows that x is large enough that the police officer interrogates him, and 2) depending on the interrogation method, the value of x is conveyed to him. The modeling distinction for confrontational vs. non-confrontational is that under confrontational interrogation, the suspect is confronted with x by police. That is, the suspect is shown the evidence. Under non-confrontational interrogation, the suspect is not shown the evidence, closing the second channel of information.

I compare requiring non-confrontational interrogation to requiring confrontational interrogation. These methods are viewed as being required since the interrogation method used is typically specified by the police department, agency, etc., and some training is provided. For this comparison, I focus on whether the interrogation method is confrontational. There are other methods in addition to the PEACE approach that are non-confrontational, but it is currently the most widely used non-confrontational approach.

[3]The "in the shadow of trial" language is used by Bibas (2004) to describe many economic models of plea bargaining. Thus, the modeling exercise here is very related to plea bargaining.

Under the confrontational interrogation method, there is a pure-strategy perfect Bayesian equilibrium. The reduced sentence that is offered is just the expected loss from going to court for the accused given the realized value of x, and induces some innocent to confess. Under the non-confrontational interrogation method, there is a semi-separating equilibrium along the lines of that in Baker and Mezzetti (2001). The value of x is not conveyed to the accused and a single reduced sentence is offered for all levels of x. Setting the reduced sentence equal to the guilty suspect's expected punishment at trial induces the innocent to not confess, and makes the guilty indifferent between confessing and not. The guilty randomize as to whether to confess, which allows for the semi-separating nature of the equilibrium. The innocent do not confess.

The analysis is presented by first studying a benchmark model in which there is no exogenous threshold on x for the prosecutor to take non-confessing suspects to trial. Instead, under the. non-confrontational method, the prosecutor's decision of whether to go to trial is based on observing x and the equilibrium mixed strategy of the guilty accused. In this case, the prosecutor's expected payoff, which is sensitive to wrongful convictions, may also be taken to represent society's preferences. The non-confrontational method weakly dominates the confrontational one in that it always yields an expected equilibrium payoff that is at least as high as that of the confrontational method. The non-confrontational method does not induce the innocent to confess, but innocent defendants have a positive probability of being found guilty at trial.

The reason the non-confrontational method sometimes performs strictly better is the following. For some distributions, a higher threshold for taking non-confessing accused to trial (and corresponding reduced sentence for confessing) can be used. This results in the confession of some guilty who would not be taken to trial following non-confession and reduces the number of innocent (and guilty) who are taken to trial following non-confession.

Then we study, the case where the prosecutor has an exogenous threshold on x for taking a non-confessing accused to trial that is higher than the threshold for police to interrogate. This may be due, for example, to external political pressure or to the prosecutor's office having a high standard for taking a case to trial. Here, the comparison is similar with the non-confrontational method weakly dominating the confrontational method in terms of expected equilibrium payoff and the reason is similar. The confrontational method has a similar equilibrium to the benchmark case. Since the accused are informed of x and know the prosecutor's threshold x for trial, those with an x for which the prosecutor will not go to trial cannot be induced to confess. However, the non-confrontational method can induce

some guilty who would not be taken to trial to confess. As the accused are not informed of x, they choose whether to confess based on their expected loss should they not confess, which accounts for there being some probability that x is high enough to be taken to trial and also some probability that x is low enough they are not taken to trial. This results in the confession of some guilty who turn out to have an x below the prosecutor's threshold.

Related Literature

To my knowledge there has been little work, in the economics, and law and economics literatures, focusing on interrogation techniques. Bull (2012a) considers interrogation and the fabrication of evidence. Baliga and Ely (2016) study interrogation commitment problems inherent to torture. Ispano and Vida (2020) and (2023) study interrogation using cheap talk models. Their focus is on information disclosure through cheap statements. Although, in Ispano and Vida (2023) they focus on a "back and fourth" interrogation and do not attempt to specifically model the widely used interrogation methods, they have some similar ideas related to police not revealing too much information early on. They also study commitment by the police. In my model, commitment comes from officers following police department protocol and not from a sense of one officer delegating tasks to another as Ispano and Vida study.

There is a large literature on plea bargaining that is related. Some work on plea bargaining focuses on prosecutor resource expenditures and information revelation. Grossman and Katz (1983) study a game-theoretic model in which the plea bargain is used as a screening device. In their model, the prosecutor takes to trial only defendants she knows are innocent. Until Baker and Mezzetti (2001), the plea bargaining literature relied on separating equilibria in which the guilty accepted the plea bargain and the innocent rejected it. This relies on a non-credible threat of the prosecutor to take a defendant who is likely to be innocent (or an innocent defendant) to trial. Again, the model studied here is closely related to that of Baker and Mezzetti, and requires credibility of the prosecutor regarding going to trial.

Other work focuses on the impact of plea bargaining on type-1 and type-2 errors. These include Bar-Gill and Gazal Ayal (2006), Mungan and Klick (2016), and Lundberg (2019). Bar-Gill and Gazal Ayal show that by restricting the permissible sentence reduction in a plea bargain, the law can preclude plea bargains in cases with a low probability of conviction.

Mungan and Klick (2016) show that large compensations for exonorees can reduce expected costs associated with wrongful determinations of guilt in trial and thereby reduce the number of innocent pleas. Lundberg (2019) focuses on the scope for police discretion to increase deterrence. When citizens are active in multiple directions of crime, there is scope

for sharing information related to a second crime provided police sometimes reduce penalties for a first crime that has been discovered.

The current paper does not model a search for evidence or for whether a crime has been committed; there is less focus on deterrence. There is a large literature that is focused on deterrence. See, for example, Polinsky and Shavell (2007). The focus here is on the effects of information conveyed to the accused on the incentive to confess. It's possible that incentives to confess could influence ex ante deterrence effects. Since the information conveyed to the accused can induce both the guilty and innocent to confess, the implications of more information differ from those for the police or a prosecutor being better informed.

Much of the literature on settlement has focused on contract settings and other issues. See, for example, Shavell (1989), Shavell (1993), and Rosenberg and Shavell (2006). Recently in the economics literature there has been considerable work addressing questions related to bail decisions, pretrial detention, and diversion. See, for example, Mueller-Smith and Schnepel (2020), Arnold, Dobbie, and Yang (2018), and Dobbie, Goldin, and Yang (2018).

There is much recent work on evidence that is related. This includes Bull (2008a and 2008b), Bull and Watson (2007), Bull and Watson (2004), Sanchirico (1999, 2000, and 2001), Sanchirico and Triantis (2008). Sanchirico (2010) emphasizes the importance of incorporating fabrication into models of evidence. Bull and Watson (2019) study jury interpretation and the exclusion of evidence at trial in a setting with two channels of information. Although their model focuses on disclosure incentives and a different setting, it is related in there are two channels of information. Fluet and Lanzi (2018), in a different setting, also have two channels of information. Lundberg and Mungan (2022) show that, when the jury beliefs are rational, any prohibition on the use of defendant characteristics as a form of evidence will never simultaneously reduce conviction rates for the innocent and increase conviction rates for the guilty. Bull (2025) provides conditions under which prohibiting jurors from conditioning on observable defendant characteristics improves accuracy.

The law literature that has focused on the effects of Miranda rights on police interrogation has touched some upon police lying to suspects. Also, the legal psychology literature has investigated issues related to interrogation. See, for example, Leo and Ofshe (1998), Magid (2001), Ofshe and Leo (1997), and White (2001).

Some studies, such as Weisselberg (2001), focus more on the details of how the Miranda rights are followed regarding questioning of suspects.[4] Kassin, et. al. (2007), based on police survey results, found that 81 percent of suspects waive their Miranda rights. They also

[4]There is a law and economics literature on the right to silence and the 5th Amendment. See Seidman and Stein (2000), Mialon (2005), and Leshem (2010).

reported, "The typical interrogation often, but not always, includes confronting the suspect with evidence of his or her guilt and appealing to his or her self-interests." Additionally, their respondents estimated "that 69.48% of guilty suspects provide a confession" and "23.30% of innocent suspects provide some form of confession."

The rest of the paper is organized as follows. Section 2 presents the benchmark model. The equilibria for the two methods are contained in Section 3 and a comparison of the two methods is discussed in Section 4. Section 5 presents the analysis for when the prosecutor has an exogenous threshold for going to trial. Section 6 concludes by discussing policy implications and limitations.

2 Benchmark Model

This section describes the model of interaction between a police officer, a prosecutor, denoted by P, and a person accused of a crime, who is denoted by A. I assume that A knows whether he is guilty (G) or innocent (I), but the police officer and P do not.

The police officer observes a private signal x that conveys information about A's guilt. The information conveyed by x becomes available to P should the case proceed to P. We can motivate x as being implied by hard evidence that is uncovered during the investigation. As the trial process is not the focus here, I do not explicitly model it, but instead assume that x denotes the probability with which A is found guilty at trial.

A does not observe x directly. The threshold for interrogation, x_b, exogenously determined by the police department as is the interrogation method. If the police officer interrogates, the interrogation method in place determines whether she conveys the value of x to A. For the non-confrontational approach, A is not informed of the value of x but knows that $x \geq x_b$. Under the confrontational approach, A is informed of the value of x.[5]

Upon being interrogated, A chooses whether to confess. Confessing corresponds to accepting a perceived reduced sentence.[6] Confession is assumed to result in a reduced sentence of R. If A does not confess, the case goes to P. P forms an updated belief that A is guilty given x and that A rejected the reduced sentence R. This is denoted by $\mu[G \mid R, x]$. P's

[5]There is a substantial literature that suggests that once police officers interrogate a suspect, they wish to obtain a confession. However, as modeled, the police officer simply follows department procedure.

[6]In practice, confessing during interrogation does not necessarily equate to not going to trial. In many (probably most) cases it does; a plea bargain is agreed to. See, for example, Redlich, et al. (2018). If a defendant who has confessed later decides instead go to trial, it is a very challenging to overcome the confession with the jury. The assumption in this model can be motivated by assuming that A correctly anticipates the outcome of this later plea bargaining stage. This is not the main focus of the exercise here so this simplification is used.

posterior belief that A is innocent is denoted $\mu[I \mid R, x]$, which is equivalent to $1-\mu[G \mid R, x]$.

P is assumed to set R since in practice the prosecutor presents any plea bargain to the court. Since prosecutors are not involved in interrogation and are typically not informed of arrests and interrogation, I assume that P provides guidance ahead of time that police can follow. In the case of non-confrontational interrogation, I assume there is a single value of R that is offered for all $x \geq x_b$ since offering different values of R for different values of x would potentially reveal information about x to A and that is at odds with the assumptions for this method of interrogation. For confrontational interrogation, the police officer reveals the realization of x to A so I assume P provides, ahead of time, the police with a schedule of R as a function of x.

If P takes the case to trial, the probability with which the court finds A guilty is given by x. P's payoff as a function of sentence length s is

$$u_P = \begin{cases} s & \text{if} \quad \theta = G \\ -s & \text{if} \quad \theta = I \end{cases} .$$

This coincides with society's payoff function. The suspect A's loss is $u_A(s) = -s$.

The sequence of interaction is represented in the timeline below.

1. Interrogation method selected (exogenous).

2. Prosecutor chooses reduced sentence for confession, R, which is to be implied or conveyed by the police officer.

3. A observes his guilt or innocence: $\theta \in \{G, I\}$. P does not observe θ.

4. P observes a signal $x \in [0, 1] = X$. The realization is according to distribution $f(\theta, x)$.

5. P interrogates if and only if $x \geq x_b$. If P does not interrogate, the interaction ends. If P interrogates, we continue to 6 below.

6. If P interrogates and is required to use confrontational interrogation, A is informed of x. Under non-confrontational interrogation, A only knows that $x \geq x_b$. A forms an expectation of his loss should he not confess. A weighs this against the reduced sentence $R \leq H$ if he confesses.

7. A chooses whether to confess.

8. If A does not confess, the prosecutor observes x and chooses whether to go to trial or dismiss. The probability A is found guilty at trial is given by x. If found

guilty, A receives sentence H.

It will sometimes be useful to write the probability of x conditional on θ, which is given by the standard conditional-probability formula:

$$f(x \mid \theta) \equiv \frac{f(\theta, x)}{\int_0^1 f(\theta, x)dx}.$$

For notational simplicity, I assume a monotone likelihood ratio property (MLRP) with respect to x so that

$$\frac{f(x \mid G)}{f(x \mid I)} > \frac{f(x' \mid G)}{f(x' \mid I)},$$

for $x > x'$. So, in expectation, a guilty A is more likely to be found guilty at trial than an innocent A is. However, an innocent A may be wrongfully convicted. Assume that for $\underline{x}$ it is equally likely that A is innocent or guilty; that is

$$f(G \mid \underline{x}) = f(I \mid \underline{x}).$$

For $x < \underline{x}$ it is more likely that A is innocent, and for $x > \underline{x}$ it is more likely that A is guilty. Note that this is in terms of $\underline{x}$ as a *face-value signal*, following Bull and Watson (2019), in that it is based on the statistical properties of x and not on the behavior of the accused, which can depend on his private information of his type. P also receives information from A's choice to not confess, which can be viewed as a claim of being not guilty. So there are two channels of information for P. It is assumed that $\underline{x} < x_b$. In Section 5, studies the prosecutor having an exogenous threshold $x_P > x_b$ for which she will only take the accused to trial when $x_P \leq x$.

P's strategy is given by her specification of R and the probability she chooses to take A who has not confessed to trial. The probability P chooses to take a case to trial is denoted by $\phi : \mathbb{R}_+ \times X \to [0, 1]$, where $\mathbb{R}_+$ represents the space of all possible reduced sentences. So $\phi(R, x)$ is the probability P takes A who has not confessed for reduced sentence R to trial.

P's expected payoff from going to trial is given by

$$u_p(R, x) = \mu[G \mid R, x]xH - [1 - \mu[G \mid R, x]]xH.$$

Assume this coincides with society's payoff/social welfare function.

Under non-confrontational interrogation, A's expected loss from not confessing is

$$U_{N,\theta} = \int_{x_b}^{1} \phi(R,x) x H \left[\frac{f(x \mid \theta)}{1 - F(x_b \mid \theta)} \right] dx.$$

A of type θ confesses, given R, with probability $\alpha(R,\theta) \in [0,1]$. Under confrontational interrogation, A's expected loss from not confessing is

$$U_{C,\theta}(R,x) = \phi(R,x) H x.$$

A of type θ confesses, given R and x, with probability $\alpha(R,x,\theta) \in [0,1]$.

3 Equilibria in Benchmark Model

I consider a perfect Bayesian equilibrium for each interrogation method.[7] Perfect Bayesian equilibrium requires a specification of strategies and consistent beliefs. For A, we need probabilities of confessing for each type and interrogation method; these are for the non-confrontational and confrontational methods $\alpha(R,\theta), \alpha(R,x,\theta)$, respectively. For P, we need R for each interrogation method, and the probability of taking a case to trial $\phi(R,x)$. P selects R in period 2 and ϕ is selected in period 8. These must maximize the expected payoffs for each, given the equilibrium behavior and beliefs. P's belief of guilt, given R and x, $\mu[G \mid R,x]$, must be derived using Bayes' rule where possible.

Given that P wishes to avoid wrongful convictions, there is not a pure strategy separating equilibrium for either interrogation method.[8]

Theorem 1: *For each interrogation method, there is no separating equilibrium in which A completely reveals his type and* $R > 0$.

Proof: Consider first the non-confrontational method. A is not informed of the value of x but knows that it is at least x_b. By MLRP, the expected probability of being found guilty at trial is higher for the guilty type of A than for the innocent type. So any potential separating equilibrium would have to induce the guilty type to confess and the innocent type to not confess. However, given her payoff function, P will not take a defendant she

[7] See Watson (2025) for thorough discussion of perfect Bayesian equilibrium. Regarding sequential rationality, there are some settings where perfect Bayesian equilibrium does not imply subgame perfection. Watson shows a very weak application of moderate plain consistency is adequate to get subgame perfect equilibrium; this consistency notion is defined for infinite games.

[8] This is similar to a result in Baker and Mezzetti (2001).

knows is innocent to trial, and this would induce the guilty type of A to not confess. So such an equilibrium cannot exist.

Next consider the confrontational method. A is informed of the value of x and the value of R can depend on x. While higher values of x are more likely for guilty types of A, there is positive probability of those for innocent types as well. So for any $R > 0$, it's not possible to induce all guilty and innocent types of A to behave in different ways. □

Now let's consider an equilibrium for each of the two interrogation methods. Consider the confrontational method first.

Confrontational

Under the confrontational interrogation method, there is a pure-strategy equilibrium. However, as Theorem 1 indicates, this is not a separating equilibrium. Because the probability of being found guilty at trial depends only on x, which is revealed to A by the police officer, some innocent accused confess. Further, since x is revealed to A by the police officer, it is possible to condition R on x.[9] Given a realization of x, it's possible to induce A to confess by setting $R \leq xH$.

Theorem 2: *Under confrontational interrogation in the base model, the following strategies and beliefs constitute a pure strategy perfect Bayesian equilibrium:*

$$R^*(x) = \begin{cases} xH & \text{if } \text{Prob}\left[G \mid x\right]xH - \left[1 - \text{Prob}\left[G \mid x\right]xH \geq 0 \\ H & \text{if } \text{Prob}\left[G \mid x\right]xH - \left[1 - \text{Prob}\left[G \mid x\right]xH < 0 \end{cases}$$

A of either type θ confesses if and only if $R \leq xH$.

Following rejection of R, P goes to trial whenever x is such that

$$\text{Prob}\left[G \mid x\right]xH - \left[1 - \text{Prob}\left[G \mid x\right]\right]xH \geq 0.$$

The prosecutor's expected payoff under the confrontational method is

$$\text{Prob}\left[G\right]\int_{x_b}^{1} xH\left[\frac{f(x \mid G)}{1 - F(x_b \mid G)}\right]dx - \text{Prob}\left[I\right]\int_{x_b}^{1} xH\left[\frac{f(x \mid I)}{1 - F(x_b \mid I)}\right]dx$$

[9] Another possible modeling approach is to assume that P does not allow for the police officer to vary R with x and sets a value of R that holds for all x as was required for the non-confrontational approach. However, such an approach is not ideal from P's perspective and will perform worse, for both P and society, than the approach studied here.

$$= \text{Prob}\,[G]\overline{x}_G H - \text{Prob}\,[I]\overline{x}_I H,$$

where

$$\overline{x}_\theta = \int_{x_b}^{1} x \left[\frac{f(x \mid \theta)}{1 - F(x_b \mid \theta)}\right] dx.$$

Proof: In the last stage of the game, following rejection of R, P goes to trial whenever x is such that

$$\text{Prob}\,[G \mid x]xH - [1 - \text{Prob}\,[G \mid x]]xH \geq 0.$$

That is when going to trial has a (weakly) positive expected value. Each type of A confesses when it is favorable to him to do so, given x, or when $R \leq xH$.

In the interrogation stage, P does not wish to induce a confession from those who are more likely than not innocent. Setting $R = H$ for those induces them to not confess. So P specifies

$$R^*(x) = \begin{cases} xH & \text{if} \quad \text{Prob}\,[G \mid x]xH - [1 - \text{Prob}\,[G \mid x]xH \geq 0 \\ H & \text{if} \quad \text{Prob}\,[G \mid x]xH - [1 - \text{Prob}\,[G \mid x]xH < 0 \end{cases}$$

If A with x satisfying the first case does not confess, P takes him to court in the last stage.

Consider P's expected payoff. Since x is observed and conveyed to A, we have $R = xH$. So we have P's ex ante expected payoff given by

$$\text{Prob}\,[G] \int_{x_b}^{1} xH \left[\frac{f(x \mid G)}{1 - F(x_b \mid G)}\right] dx - \text{Prob}\,[I] \int_{x_b}^{1} xH \left[\frac{f(x \mid I)}{1 - F(x_b \mid I)}\right] dx$$

$$= \text{Prob}\,[G]\overline{x}_G H - \text{Prob}\,[I]\overline{x}_I H,$$

where

$$\overline{x}_\theta = \int_{x_b}^{1} x \left[\frac{f(x \mid \theta)}{1 - F(x_b \mid \theta)}\right] dx.$$

□

Non-Confrontational

Now, consider the non-confrontational method. Note that for appropriate values of R there is a semi-separating equilibrium that has the I type of A not confess and the G type randomize so that P is willing to take non-confessing A to trial. For R that is less than the expected punishment from going to trial for the innocent, both types will confess. For R that is larger than the expected punishment from going to trial for the guilty, neither type

will confess.

The semi-separating equilibrium involves the guilty A randomizing so as to make P willing to go to trial following non-confession. P's choice of R amounts to choosing a threshold x_t such that P goes to trial following non-confession and observing that x is at least x_t. P chooses R to maximize her expected payoff. So it is useful to describe the equilibrium of the proper subgame following the choice of R.

Lemma 1: *Assume*

$$\int_{x_b}^{1} xH \left[\frac{f(x \mid G)}{1 - F(x_b \mid G)}\right] dx \leq R < H.$$

Each such R implies a unique equilibrium of the remaining proper subgame that is part of a semi-separating equilibrium. That is, for each R there is a unique equilibrium α and cutoff x_t.

Proof: In the last stage of the game, following rejection of R, P goes to trial whenever x is such that

$$u_p(R, x) = \mu[G \mid R, x]xH - [1 - \mu[G \mid R, x]]xH \geq 0,$$

for posterior belief μ.

In period 7 of the game, A of type θ confesses if

$$R \leq \int_{x_t}^{1} \phi(R, x)xH \left[\frac{f(x \mid \theta)}{1 - F(x_b \mid \theta)}\right] dx.$$

Given R, P must choose the threshold for trial x_t so that it satisfies the guilty's indifference condition

$$R = \int_{x_t}^{1} \phi(R^*, x)xH \left[\frac{f(x \mid G)}{1 - F(x_b \mid G)}\right] dx,$$

ensuring that the guilty are exactly indifferent between confessing and rejecting.

Since the innocent type always rejects the plea offer, by Bayes' rule the updated belief upon observing a rejection is:

$$\mu[G \mid R, x] = \frac{[1 - \alpha] \operatorname{Prob}[G \mid x]}{[1 - \alpha] \operatorname{Prob}[G \mid x] + \operatorname{Prob}[I \mid x]}$$

P's expected payoff from going to trial is

$$u_p(R, x) = \mu[G \mid R, x]xH - [1 - \mu[G \mid R, x]]xH.$$

Since R is constant across x, P can be indifferent about going to trial for at most one value of x. So we consider that she goes to trial if

$$u_p(R,x) = \mu[G \mid R,x]xH - [1-\mu[G \mid R,x]]xH \geq 0.$$

So

$$\phi(R,x) = \begin{cases} 1 & \text{if} \quad u_t \geq 0 \\ 0 & \text{if} \quad u_t < 0 \end{cases}.$$

The guilty A must randomize so that P is indifferent as to whether to go to trial following non-confession and $x = x_t$. This requires

$$u_p(R,x_t) = \mu[G \mid R,x_t]x_tH - [1-\mu[G \mid R,x_t]]x_bH = 0,$$

or that

$$\frac{[1-\alpha]\,\text{Prob}\,[G \mid x_t]}{[1-\alpha]\,\text{Prob}\,[G \mid x_t] + \text{Prob}\,[I \mid x_t]}x_tH - \frac{[\text{Prob}\,[I \mid x_t]}{[1-\alpha]\,\text{Prob}\,[G \mid x_t] + \text{Prob}\,[I \mid x_t]}x_tH = 0.$$

This implies

$$\frac{[1-\alpha]\,\text{Prob}\,[G \mid x_t]}{[1-\alpha]\,\text{Prob}\,[G \mid x_t] + \text{Prob}\,[I \mid x_t]} = \frac{\text{Prob}\,[I \mid x_t]}{[1-\alpha]\,\text{Prob}\,[G \mid x_t] + \text{Prob}\,[I \mid x_t]},$$

or

$$[1-\alpha]\,\text{Prob}\,[G \mid x_t] = \text{Prob}\,[I \mid x_t],$$

or

$$1 - \frac{\text{Prob}\,[I \mid x_t]}{\text{Prob}\,[G \mid x_t]} = \alpha.$$

Since we study a pure-strategy equilibrium for the confrontational method, I suppress the arguments of α.

Given α and R, A with $x = x_t$ is equally likely to be guilty or innocent. Given MLRP, this value of x_t is unique. Note that setting the threshold higher than x_t reduces P's expected payoff since it involves not taking some A who are more likely to be guilty than innocent to trial. Similarly, setting the threshold lower than x_t reduces P's expected payoff since it involves taking some A who are more likely to be innocent to trial. □

In light of this, in period 2 P chooses R to maximize her expected payoff given equilbrium

behavior in the subsequent subgame. The lowest possible value of R is that which would induce a trial threshold of x_b in the subgame following the choice of R. This results in A who do not confess being taken to trial for all values of $x \geq x_b$. That is, for such a choice of R, all who are interrogated and do not confess are taken to trial.

Alternatively, the choice of a larger value of R with corresponding threshold $x_t > x_b$ results in only those non-confessing A with values of x of at least $x_t > x_b$ being taken to trial. So not all non-confessing A are taken to trial since P observes x prior to deciding whether to go to trial and will not go to trial when $x < x_t$. In this case, some guilty A who turn out to have a value of x low enough to not be taken to trial following non-confession are induced to confess. This is because they do not observe their x prior to choosing whether to confess and base their decision to confess on their expected value of x. Since they are guilty, their expected value of x is higher than that of the innocent due to MLRP.

Theorem 3: *Under non-confrontational interrogation in the base model, the following strategies and beliefs constitute a semi-separating perfect Bayesian equilibrium:*

P selects R^* and corresponding x_t to maximize her expected payoff.

$$R^* = \int_{x_t}^{1} xH\left[\frac{f(x \mid G)}{1 - F(x_b \mid G)}\right] dx,$$

A of type I confesses if

$$R \leq \int_{x_t}^{1} \phi(R, x)xH\left[\frac{f(x \mid I)}{1 - F(x_b \mid I)}\right] dx.$$

A of type G confesses if

$$R \leq \int_{x_t}^{1} \phi(R, x)xH\left[\frac{f(x \mid G)}{1 - F(x_b \mid G)}\right] dx,$$

and does not confess if

$$R > \int_{x_t}^{1} \phi(R, x)xH\left[\frac{f(x \mid G)}{1 - F(x_b \mid G)}\right] dx.$$

A of type G randomizes and confesses with probability α if

$$R = \int_{x_t}^{1} \phi(R, x)xH\left[\frac{f(x \mid G)}{1 - F(x_b \mid G)}\right] dx,$$

where

$$\alpha = 1 - \frac{\text{Prob}\,[I \mid x_t]}{\text{Prob}\,[G \mid x_t]}.$$

Following rejection of R, P goes to trial whenever x is such that

$$u_p(R, x) = \mu[G \mid R, x]xH - [1 - \mu[G \mid R, x]]xH \geq 0,$$

where

$$\mu[G \mid R, x] = \frac{[1-\alpha]\,\text{Prob}\,[G \mid x]}{[1-\alpha]\,\text{Prob}\,[G \mid x] + \text{Prob}\,[I \mid x]}$$

$$\phi(R, x) = \begin{cases} 1 & \textit{if} \quad u_P \geq 0 \\ 0 & \textit{if} \quad u_P < 0 \end{cases}.$$

Proof: This follows from Lemma 1 and sequential rationality. □

The intuition for this equilibrium is that P chooses R^* and corresponding x_t so as to make the guilty A indifferent between confessing and not. The innocent A do not confess. The guilty A randomize so as to make P willing (indifferent for x_t) to go to trial following A not confessing. When R^* is such that $x_t > x_b$, the guilty A's randomization is such that higher value of x is needed for P to anticipate that A is equally likely to be guilty or innocent for x_t. That is, α is higher so that the probability G confesses is higher, which raises the threshold value of x, denoted by x_t, so that a non-confessing A is equally likely to be guilty or innocent.[10]

These possible choices of R are discussed below. Naturally, P chooses the value of R, and the implied threshold for trial following non-confession, that maximizes her payoff. The next Theorem describes the expected payoffs of P and society for these possible values of R. A comparison of these is contained in Section 4.

Theorem 4: *Under non-confrontational interrogation in the base model, P's expected payoffs (and society's) are as follows.*

In equilibrium R^* corresponds to the optimal $x_t \in [x_b, 1]$, P's ex ante expected payoff is

$$\text{Prob}\,[G] \int_{x_t}^{1} xH \left[\frac{f(x \mid G)}{1 - F(x_b \mid G)}\right] dx - \text{Prob}\,[I] \int_{x_t}^{1} xH \left[\frac{f(x \mid I)}{1 - F(x_b \mid I)}\right] dx.$$

[10]Naturally, the guilty A are indifferent between confessing and not, but it's typical to set the decision when there's indifference so there's equilibrium behavior. See for example, Baliga and Sjostrom (2009) and Bull (2012b). Since this involves randomization by the guilty A, P does not actually observe the choice of mixed strategy.

When R^* corresponds to the optimal $x_t = x_b$, P's ex ante expected payoff can be written as

$$\text{Prob}\,[G]\overline{x}_G H - \text{Prob}\,[I]\overline{x}_I H,$$

where

$$\overline{x}_\theta = \int_{x_b}^{1} x \left[\frac{f(x \mid \theta)}{1 - F(x_b \mid \theta)}\right] dx.$$

Proof: Given the optimal R^* and corresponding x_t, denoted by $R(x_t)$, consider the components of P's ex ante expected payoff. Averaged over all realizations of x and decisions by the guilty defendant, the ex ante expected payoff when the prosecutor encounters a guilty defendant is:

$$\alpha R(x_t) + [1 - \alpha] \int_{x_t}^{1} xH \left[\frac{f(x \mid G)}{1 - F(x_b \mid G)}\right] dx.$$

The first term represents those guilty who confess, with probability α, and receive R. The second term represent those who do not confess, with probability $[1 - \alpha]$ and go to trial if $x \geq x_t$. In equilibrium,

$$R^*(x_t) = \int_{x_t}^{1} xH \left[\frac{f(x \mid G)}{1 - F(x_b \mid G)}\right] dx.$$

So this reduces to

$$\int_{x_t}^{1} xH \left[\frac{f(x \mid G)}{1 - F(x_b \mid G)}\right] dx.$$

The innocent do not confess, and P's expected payoff when encountering an innocent is

$$\int_{x_t}^{1} xH \left[\frac{f(x \mid I)}{1 - F(x_b \mid I)}\right] dx.$$

Thus, P's ex ante expected payoff is given by

$$\text{Prob}\,[G] \int_{x_t}^{1} xH \left[\frac{f(x \mid G)}{1 - F(x_b \mid G)}\right] dx - \text{Prob}\,[I] \int_{x_t}^{1} xH \left[\frac{f(x \mid I)}{1 - F(x_b \mid I)}\right] dx.$$

When the optimal R^* and corresponding x_t are such that $x_t = x_b$,

$$= \text{Prob}\,[G]\overline{x}_G H - \text{Prob}\,[I]\overline{x}_I H,$$

where

$$\overline{x}_\theta = \int_{x_b}^{1} x \left[\frac{f(x \mid \theta)}{1 - F(x_b \mid \theta)}\right] dx.$$

□

We now consider details of the choice of R and the corresponding x_t

Theorem 5: *Under non-confrontational interrogation in the base model, P selects R^* so that the corresponding $x_t > x_b$ if and only if there exists $x_t > x_b$ such that*

$$\text{Prob}\,[I]\int_{x_b}^{x_t} x\left[\frac{f(x \mid I)}{1-F(x_b \mid I)}\right]dx > \text{Prob}\,[G]\int_{x_b}^{x_t} x\left[\frac{f(x \mid G)}{1-F(x_b \mid G)}\right]dx,$$

.

Proof: Let $u_P(x_b)$ denote P's expected payoff when R and the corresponding threshold imply that P will take all non-confessing A with $x \geq x_b$ to trial. Similarly, let $u_P(x_t)$ denote when P selects an R with a threshold $x_t > x_b$.

Consider

$$u_P(x_t) - u_P(x_b) = \text{Prob}\,[G]\int_{x_t}^{1} xH\left[\frac{f(x \mid G)}{1-F(x_b \mid G)}\right]dx - \text{Prob}\,[G]\int_{x_b}^{1} xH\left[\frac{f(x \mid G)}{1-F(x_b \mid G)}\right]dx$$

$$-\text{Prob}\,[I]\int_{x_t}^{1} xH\left[\frac{f(x \mid I)}{1-F(x_b \mid I)}\right]dx + \text{Prob}\,[I]\int_{x_b}^{1} xH\left[\frac{f(x \mid I)}{1-F(x_b \mid I)}\right]dx.$$

If this expression is positive, P has a higher expected payoff from choosing R so that the threshold is $x_t > x_b$. Since $H > 0$ the comparison reduces to asking when

$$\text{Prob}\,[I]\int_{x_b}^{x_t} x\left[\frac{f(x \mid I)}{1-F(x_b \mid I)}\right]dx > \text{Prob}\,[G]\int_{x_b}^{x_t} x\left[\frac{f(x \mid G)}{1-F(x_b \mid G)}\right]dx.$$

Consider the denominator in each expression,

$$1 - F(x_b \mid G)$$

and

$$1 - F(x_b \mid I).$$

It's possible to have

$$1 - F(x_b \mid G) < 1 - F(x_b \mid I).$$

So, although MLRP tells us that higher values of x imply it's more likely that the accused

is guilty, there are distributions that can result in

$$\text{Prob}\,[I]\int_{x_b}^{x_t} x\left[\frac{f(x\mid I)}{1-F(x_b\mid I)}\right]dx > \text{Prob}\,[G]\int_{x_b}^{x_t} x\left[\frac{f(x\mid G)}{1-F(x_b\mid G)}\right]dx,$$

which is equivalent to $u_P(x_t) > u_P(x_b)$. □

The Appendix contains a numerical example showing that such an outcome is possible. The $x_t > x_b$ equilibrium performs better for some distributions because the non-confrontational method allows for inducing guilty who would not be taken to trial following non-confession (those with $x < x_t$). Those distributions are relatively innocent heavy enough in the $[x_b, x_t]$ interval that the benefit from avoiding taking the innocent in that interval to trial outweighs the loss from not taking non-confessing guilty in that interval to trial. Importantly, that some of the guilty in that interval confess lessens the loss associated with not taking those guilty to trial following non-confession.

P goes to trial when A is at least as likely to be guilty than innocent. Note that xH, the expected punishment at trial, does not affect this decision. However, the value of x does influence the value of μ.

Observation 1: *P chooses to go to trial, given her posterior belief, when A is at least as likely to be guilty as innocent. This is independent of xH.*

4 Comparison in Benchmark Model

We now turn to a comparison of the two interrogation methods. As proponents of non-confrontational interrogation methods suggest, in this model, non-confrontational interrogation results in fewer false confessions than confrontational interrogation does. In fact, it results in no false confessions. However, under the non-confrontational interrogation method, innocent are taken to trial and some are convicted.

When under the non-confrontational method the equilibrium involves R and a corresponding threshold of x_b, the two methods perform equally well based on the expected payoff to society, which is equivalent to the prosecutor's expected payoff. Importantly, P chooses the value of R at time 2 so as to maximize her expected payoff, which is the same as society's expected payoff. A value R that corresponds to $x_t > x_b$ may or may not result in a higher expected payoff. However, if P optimally selects R so that the equilibrium involves a threshold x_t with $x_t > x_b$, it must be that the non-confrontational method yields a weakly and possibly strictly higher payoff for society. Since, under the non-confrontational

method, P can always select a value of R that corresponds to $x_t = x_b$ and an equilibrium expected payoff equal to that under the confrontational method, the non-confrontational method yields a weakly higher expected payoff for society.

Theorem 6: *In terms of society's payoff, the non-confrontational interrogation method weakly dominates the confrontational interrogation method. When the non-confrontational method results in a strictly higher payoff for society (and the prosecutor), it induces confession from some guilty who would not be taken to trial following non-confession.*

Proof: This follows from payoffs for P, which are the same as those of society, as specified in Theorems 2 and 4. Consider the non-confrontational method. When P selects R corresponding $x_t = x_b$, society's expected payoffs under both methods are equal. For some, but not all, distributions and priors, it's possible for R that corresponds to $x_t > x_b$ to yield a payoff for society that is higher than that for an R that corresponds to $x_t = x_b$. In that case, as described in Theorem 5, some guilty who would not be taken to trial following non-confession under either the x_b equilibrium or the equilibrium in the confrontational case confess. Since in period 2, P selects R to maximize her expected payoff, which is equivalent to maximizing society's expected payoff, the non-confrontational method results in a weakly higher expected payoff than the confrontational method. □

5 Threshold x_P for Prosecutor to go to Trial

We now consider the case where the prosecutor has an exogenously given threshold for going to trial x_P. Motivation for x_P could be that the prosecutor does not want to take too low of probability cases to trial because she is careful or wishes to be viewed as being careful. It could be that a political cost is incurred for pursuing weaker cases. There is anecdotal evidence that prosecutors may have a higher threshold for going to trial than police do for interrogating.

In such a setting, the non-confrontational method is able to induce some guilty who have $x < x_P$ to confess because A does not observe the realized value of x prior to choosing whether to confess and bases these decision on the expected value of x given his type. This is not possible under the confrontational method because A is confronted with the realized value of x and A knows that the prosecutor will not go to trial if $x < x_P$.

Consider first the confrontational method. The confrontational method yields very similar results to those of the benchmark model.

Theorem 7: *Consider the confrontational interrogation method. Suppose $x_b < x_P$. The following strategies and beliefs constitute a pure strategy perfect Bayesian equilibrium:*

$$R^*(x) = \begin{cases} xH & \text{if } \text{Prob}\left[G \mid x\right]xH - \left[1 - \text{Prob}\left[G \mid x\right]xH \geq 0 \\ H & \text{if } \text{Prob}\left[G \mid x\right]xH - \left[1 - \text{Prob}\left[G \mid x\right]xH < 0 \end{cases}$$

A of either type θ confesses if and only if $R \leq xH$.

Following rejection of R, P goes to trial whenever x is such that

$$\text{Prob}\left[G \mid x\right]xH - \left[1 - \text{Prob}\left[G \mid x\right]xH \geq 0.$$

P's expected payoff is given by

$$\text{Prob}\left[G\right]\int_{x_P}^{1} xH\left[\frac{f(x \mid G)}{1 - F(x_b \mid G)}\right]dx - \text{Prob}\left[I\right]\int_{x_P}^{1} xH\left[\frac{f(x \mid I)}{1 - F(x_b \mid I)}\right]dx.$$

Proof: This is similar to the proof for Theorem 2 so the main differences are provided. Here, A knows that P will not go to trial for $x < x_P$. The calculation of the expected payoff from not confessing must include an expression for the probability that $x \geq x_P$. That is, That is,

$$\text{Prob}\left[x \geq x_P \mid x \geq x_b, \theta\right] = \frac{1 - F(x_P \mid \theta)}{1 - F(x_b \mid \theta)}.$$

Since A is informed of x, R is as before.

$$R^*(x) = \begin{cases} xH & \text{if } \text{Prob}\left[G \mid x\right]xH - \left[1 - \text{Prob}\left[G \mid x\right]xH \geq 0 \\ H & \text{if } \text{Prob}\left[G \mid x\right]xH - \left[1 - \text{Prob}\left[G \mid x\right]xH < 0 \end{cases}$$

Incorporating that it's possible that $x < x_P$ and those cases will not be taken to trial following A not confessing yields an expected payoff for P of

$$\text{Prob}\left[G\right]\int_{x_P}^{1} xH\left[\frac{f(x \mid G)}{1 - F(x_b \mid G)}\right]dx - \text{Prob}\left[I\right]\int_{x_P}^{1} xH\left[\frac{f(x \mid I)}{1 - F(x_b \mid I)}\right]dx.$$

□

Next, consider the non-confrontational approach. The analysis is very similar to that for the benchmark model, but the probability that $x < x_P$ must be accounted for.

Theorem 8: *Assume the non-confrontational interrogation method is in place. The following strategies and beliefs constitute a pure strategy perfect Bayesian equilibrium: P selects*

R^* and corresponding x_t to maximize her expected payoff.

$$R^* = \int_{x_t}^{1} xH \left[\frac{f(x \mid G)}{1 - F(x_b \mid G)} \right] dx,$$

with $x_t \geq x_P$.

A of type I confesses if

$$R \leq \int_{x_t}^{1} \phi(R, x) xH \left[\frac{f(x \mid I)}{1 - F(x_b \mid I)} \right] dx.$$

A of type G confesses if

$$R \leq \int_{x_t}^{1} \phi(R, x) xH \left[\frac{f(x \mid G)}{1 - F(x_b \mid G)} \right] dx,$$

and does not confess if

$$R > \int_{x_t}^{1} \phi(R, x) xH \left[\frac{f(x \mid G)}{1 - F(x_b \mid G)} \right] dx.$$

A of type G randomizes and confesses with probability α if

$$R = \int_{x_t}^{1} \phi(R, x) xH \left[\frac{f(x \mid G)}{1 - F(x_b \mid G)} \right] dx,$$

where

$$\alpha = 1 - \frac{\text{Prob}\,[I \mid x_t]}{\text{Prob}\,[G \mid x_t]}.$$

Following rejection of R, P goes to trial whenever x is such that

$$u_p(R, x) = \mu[G \mid R, x] xH - [1 - \mu[G \mid R, x]] xH \geq 0 \text{ and } x \geq x_P,$$

where

$$\mu[G \mid R, x] = \frac{[1 - \alpha]\,\text{Prob}\,[G \mid x]}{[1 - \alpha]\,\text{Prob}\,[G \mid x] + \text{Prob}\,[I \mid x]}$$

$$\phi(R, x) = \begin{cases} 1 & \text{if} \quad u_P \geq 0 \\ 0 & \text{if} \quad u_P < 0 \end{cases}.$$

P's ex ante expected payoff is

$$\text{Prob}\,[G]\int_{x_t}^{1} xH\left[\frac{f(x\mid G)}{1-F(x_b\mid G)}\right]dx-\text{Prob}\,[I]\int_{x_t}^{1} xH\left[\frac{f(x\mid I)}{1-F(x_b\mid I)}\right]dx,$$

where x_t corresponds to the optimal R.

This ex ante expected payoff is at least

$$\text{Prob}\,[G]\int_{x_P}^{1} xH\left[\frac{f(x\mid G)}{1-F(x_b\mid G)}\right]dx-\text{Prob}\,[I]\int_{x_P}^{1} xH\left[\frac{f(x\mid I)}{1-F(x_b\mid I)}\right]dx.$$

Proof: This is similar to the proofs for Theorems 3 and 4. The lower bound on P's ex ante expected payoff is due it being possible for P to choose R with corresponding threshold for going to trial following non-confession of $x_t = x_P$.

P's ex ante expected payoff is

$$\text{Prob}\,[G]\int_{x_t}^{1} xH\left[\frac{f(x\mid G)}{1-F(x_b\mid G)}\right]dx-\text{Prob}\,[I]\int_{x_t}^{1} xH\left[\frac{f(x\mid I)}{1-F(x_b\mid I)}\right]dx.$$

The terms in this expression represent similar As as those in the proof of Theorem 4.

Since, as noted previously, in period 2, P selects R to maximize her expected payoff (which is equivalent to maximizing society's expected payoff), society's expected payoff is at least

$$\text{Prob}\,[G]\int_{x_P}^{1} xH\left[\frac{f(x\mid G)}{1-F(x_b\mid G)}\right]dx-\text{Prob}\,[I]\int_{x_P}^{1} xH\left[\frac{f(x\mid I)}{1-F(x_b\mid I)}\right]dx$$

That is, by sequential rationality, society's expected payoff must be at least that which could be attained by selecting R that corresponds to $x_t = x_b$ so that all non-confessing A are taken to trial. □

Here, under the non-confrontational method, guilty with $x < x_P$ can be induced to confess, but they randomize so that P is indifferent going to trial and not when $x = x_P$.

We now turn to comparing the two methods based on society's expected payoff.

Theorem 9: *When $x_b < x_P$, in terms of society's payoff, the non-confrontational interrogation method weakly dominates the confrontational interrogation method.*

Further, the non-confrontational method results induces some guilty to confess who would not confess under the confrontational method.

The non-confrontational method does not induce false confessions, but some of those innocent who do not confess are found guilty at trial.

Proof: From Theorems 8 and 7,

Confrontational: P's expected payoff is

$$\text{Prob}\,[G]\int_{x_P}^{1} xH\left[\frac{f(x\mid G)}{1-F(x_b\mid G)}\right]dx-\text{Prob}\,[I]\int_{x_P}^{1} xH\left[\frac{f(x\mid I)}{1-F(x_b\mid I)}\right]dx.$$

Non-confrontational: P's expected payoff is

$$\text{Prob}\,[G]\int_{x_t}^{1} xH\left[\frac{f(x\mid G)}{1-F(x_b\mid G)}\right]dx-\text{Prob}\,[I]\int_{x_t}^{1} xH\left[\frac{f(x\mid I)}{1-F(x_b\mid I)}\right]dx,$$

which must be at least

$$\text{Prob}\,[G]\int_{x_P}^{1} xH\left[\frac{f(x\mid G)}{1-F(x_b\mid G)}\right]dx-\text{Prob}\,[I]\int_{x_P}^{1} xH\left[\frac{f(x\mid I)}{1-F(x_b\mid I)}\right]dx.$$

That is, when P's expected payoff under the non-confrontational method is at least as high (and possibly higher) than that under the confrontational method. The non-confrontational method results in the confession of some guilty who would not be taken to trial if they did not confess (those with $x < x_P$ who confess with probability $\alpha > 0$) under either method. No such guilty confess under the confrontational method because they observe their x and know that they will not be taken to trial if they do not confess. □

Under the non-confrontational method, some guilty accused (fraction α), who would not be taken to trial due to lack of evidence strength, can be induced to confess. This is because the guilty accused weigh the reduced sentence against the expected sentence should he not confess, which accounts for the probability that the case may be dismissed due to $x < x_P$.

It's also important to note that in practice it is expected that these confessions of low x guilty accused would result in convictions. Three reasons for this are: 1) it seems unlikely that the prosecutor (or judge who handles it) would closely review the evidence to ascertain x, 2) a confession adds new evidence that is very convincing, and 3) it is very likely that a defendant who has confessed to police will not proceed to trial. Redlich, et al (2018) note, "Famously a legal scholar once proclaimed that confessions make the introduction of other types of evidence superfluous." There is a large literature supporting this view.

The practical arguments in favor of non-confrontational interrogation methods focus on the scope for not inducing innocent to confess. In equilibrium in the model of non-confrontational interrogation studied here, innocent are not induced to confess. However, the innocent (who have $x \geq x_P$) are taken to court (along with some guilty who do not

confess), and the expected number who are convicted is the same as those who confess in the model of confrontational interrogation. The reason the non-confrontational method, as modeled here, performs better is that it induces some guilty who would not be taken to court to confess. This is because, given that x is not conveyed to A, they decide whether to confess based on the expected outcome of trial.

Interestingly, when $x_b < x_P$, lowering x_b can allow more guilty to be interrogated and induced to confess.

Corollary 1: *Suppose $x_b < x_P$. Under the non-confrontational negotiation method, lowering x_b can result in more guilty confessing.*

6 Concluding Discussion and Limitations

Theorem 6 suggests that the non-confrontational method performs at least as well as a confrontational approach. Further, Theorem 9 suggests also that from an information standpoint a non-confrontational approach performs weakly better when police interrogate some suspects that the prosecutor would choose not to take to trial. The threshold for interrogation is assumed to be high enough that only those who are more likely to be guilty than innocent are interrogated. This analysis relies on rational behavior. In practice, it seems that in many instances charges are dismissed by the prosecutor. See, for example, Gershowitz (2018). This occurs for a variety of reasons. So, I argue this is a reasonable case to study.

In many models of plea bargaining, the timing of when evidence is available to the prosecutor is not a factor because the models assume rationality and that the distribution of evidence is common knowledge. So the defendant is able to form an expectation for his loss at trial. Baker and Mezzetti (2001) discuss this in detail. In the benchmark model studied here, a similar result is seen in the comparison of the two methods when the optimal R corresponds to the threshold being equal to x_b and the two methods result in the same expected payoff for the prosecutor and society. Of course, the confrontational method results in false confessions but the non-confrontational method does not induce false confessions. If society's preferences differ from those of the prosecutor and put more weight on the avoidance of false confessions, the comparison would differ.

However, there is scope, depending on the distributions, for the non-confrontational method to perform better in terms of ex ante expected payoff of society and the prosecutor. This occurs in the benchmark model when there is an optimal R that corresponds to $x_t > x_b$ and in the richer model when there is an optimal R that corresponds to $x_t > x_P$.

For some distributions, the non-confrontational method performs strictly better because it's possible for the prosecutor to choose R that corresponds to $x_t > x_b$ and induce some guilty who will not be taken to trial following non-confession to confess. This is possible since under the non-confrontational method the accused is not informed of the specific realization of the evidence x. Because the confrontational method involves confronting the accused with x, this is not possible under the confrontational method.

Limitations

The main results rely on the accused knowing the reduced sentence for confessing and being able to anticipate the prosecutor's decision about trial. I would argue that these are reasonable assumptions. It's worth noting there is, in some jurisdictions, an increased effort by prosecutors to work with police, which may increase coordination. See, for example, Hamann and Delaney (2018). On the other hand, it's possible that this increased involvement at an earlier stage by prosecutors could reduce the scope for $x_b < x_P$. However, given resource and time constraints, it seems unlikely to eliminate this even if there was a desire to do so. It seems likely that police may arrest and interrogate someone they believe is guilty but there is not sufficient evidence for the prosecutor to be willing to go to trial. Alternatively, there may be situations where a prosecutor is not reluctant to take an accused to trial as long as he is more likely guilty than not.

There are some arguments that the outcomes of plea bargains do not always reflect the defendant's chances at trial. Bibas (2004) discusses this. Some reasons have to do with psychological reasons and biases. Others have to do with "lumpiness and rigidity" in sentencing laws. It's worth noting that the arguments Bibas puts forth are focused on plea bargaining, not interrogation. However, future research with a more behavioral economics approach may yield more insights about interrogation. Further, different preferences for the suspect, the prosecutor, and society may yield differences in the behavior and in the comparisons.

A Numerical Example – Higher Payoff with $x_t > x_b$

Recall that $X = [0, 1]$, and suppose the joint densities are given by

$$f(G, x) = \frac{20x}{19} \qquad \text{and} \qquad f(I, x) = \frac{9}{19}.$$

Then

$$\Pr(G) = \int_0^1 \frac{20x}{19}\, dx = \frac{10}{19}, \qquad \Pr(I) = \int_0^1 \frac{9}{19}\, dx = \frac{9}{19}.$$

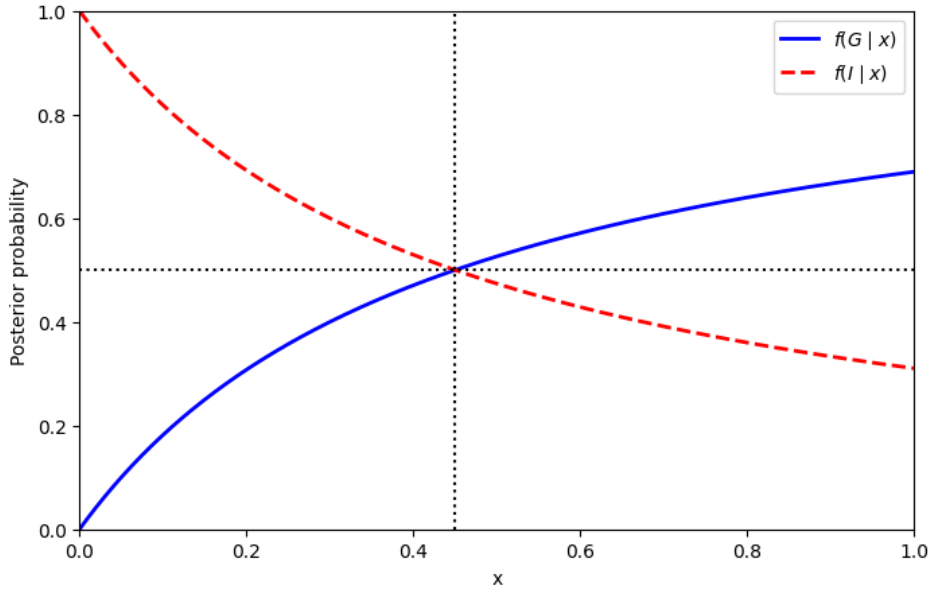

Figure 1: Posterior Probabilities.

The conditional densities are therefore

$$f(x \mid G) = \frac{f(G,x)}{\Pr(G)} = 2x, \qquad f(x \mid I) = \frac{f(I,x)}{\Pr(I)} = 1.$$

Note that $(x \mid G) \sim$ Beta$(2,1)$ and $(x \mid I) \sim$ U$[0,1]$ or Beta$(1,1)$.

We have

$$\frac{f(x \mid G)}{f(x \mid I)} = 2x,$$

which is increasing in x, so the MLRP assumption is satisfied.

The posterior probabilities are

$$f(G \mid x) = \frac{f(G,x)}{f(G,x) + f(I,x)} = \frac{20x}{20x+9},$$

and

$$f(I \mid x) = \frac{f(I,x)}{f(G,x) + f(I,x)} = \frac{9}{20x+9}.$$

The posterior probabilities are represented in Figure 1.

Let

$$\underline{x} = .45.$$

Then

$$f(G \mid \underline{x}) = \frac{20(.45)}{20(.45)+9} = \frac{9}{18} = \frac{1}{2},$$

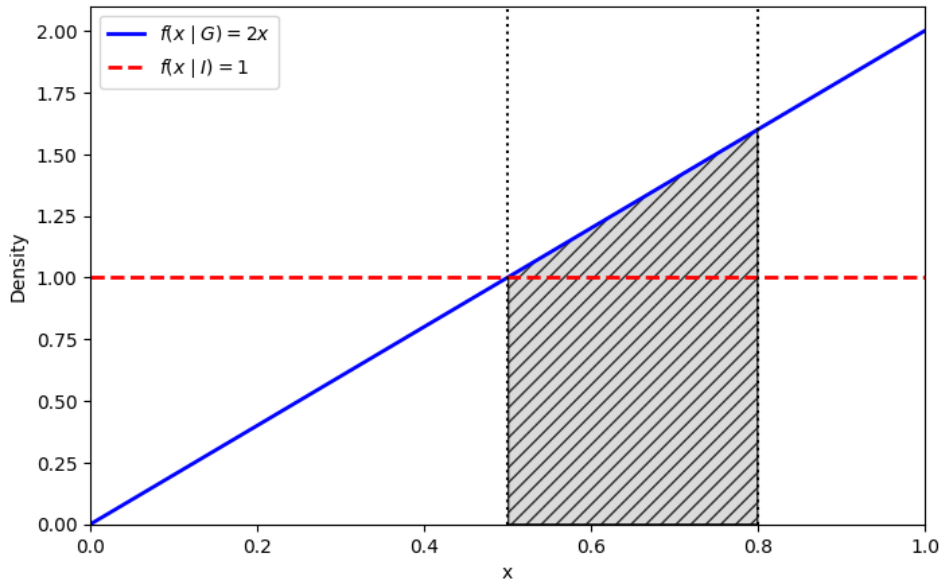

Figure 2: Conditional Densities with Interval $[x_b, x_t] = [0.5, 0.8]$.

and similarly

$$f(I \mid \underline{x}) = \frac{9}{18} = \frac{1}{2}.$$

Thus the assumption that the accused is equally likely to be guilty or innocent at $\underline{x}$ is satisfied.

Now let

$$x_b = .5, \qquad x_t = .8, \qquad H = 1.$$

Figure 2 provides a graphical representation of the densities and the $[x_b, x_t]$ interval. Note that since these are continuous densities, the density (not the probability) for the guilty goes above 1.

Since

$$F(x_b \mid G) = x_b^2 = .25, \qquad F(x_b \mid I) = x_b = .5,$$

we have

$$1 - F(x_b \mid G) = .75, \qquad 1 - F(x_b \mid I) = .5.$$

The payoff in the x_b equilibrium is

$$u_P(x_b) = \Pr(G) \int_{x_b}^{1} x \frac{f(x \mid G)}{1 - F(x_b \mid G)}\, dx - \Pr(I) \int_{x_b}^{1} x \frac{f(x \mid I)}{1 - F(x_b \mid I)}\, dx.$$

Substituting the values above,

$$u_P(x_b) = \frac{10}{19} \int_{.5}^{1} x \frac{2x}{.75}\, dx - \frac{9}{19} \int_{.5}^{1} x \frac{1}{.5}\, dx.$$

Thus

$$u_P(x_b) = \frac{10}{19} \cdot \frac{2}{.75} \int_{.5}^{1} x^2 \, dx - \frac{9}{19} \cdot 2 \int_{.5}^{1} x \, dx.$$

Since

$$\int_{.5}^{1} x^2 \, dx = \frac{1 - .5^3}{3} = \frac{7}{24},$$

and

$$\int_{.5}^{1} x \, dx = \frac{1 - .5^2}{2} = \frac{3}{8},$$

we have

$$u_P(x_b) = \frac{10}{19} \cdot \frac{8}{3} \cdot \frac{7}{24} - \frac{9}{19} \cdot 2 \cdot \frac{3}{8}.$$

So

$$u_P(x_b) = \frac{70}{171} - \frac{27}{76} = \frac{37}{684} \approx .0541.$$

Next consider the x_t equilibrium with threshold $x_t = .8$. Its payoff is

$$u_P(x_t) = \Pr(G) \int_{x_t}^{1} x \frac{f(x \mid G)}{1 - F(x_b \mid G)} \, dx - \Pr(I) \int_{x_t}^{1} x \frac{f(x \mid I)}{1 - F(x_b \mid I)} \, dx.$$

Substituting,

$$u_P(x_t) = \frac{10}{19} \int_{.8}^{1} x \frac{2x}{.75} \, dx - \frac{9}{19} \int_{.8}^{1} x \frac{1}{.5} \, dx.$$

Equivalently,

$$u_P(x_t) = \frac{10}{19} \cdot \frac{8}{3} \int_{.8}^{1} x^2 \, dx - \frac{9}{19} \cdot 2 \int_{.8}^{1} x \, dx.$$

Now

$$\int_{.8}^{1} x^2 \, dx = \frac{1 - .8^3}{3} = \frac{1 - .512}{3} = \frac{61}{375},$$

and

$$\int_{.8}^{1} x \, dx = \frac{1 - .8^2}{2} = \frac{.36}{2} = \frac{9}{50}.$$

Hence

$$u_P(x_t) = \frac{10}{19} \cdot \frac{8}{3} \cdot \frac{61}{375} - \frac{9}{19} \cdot 2 \cdot \frac{9}{50}.$$

So

$$u_P(x_t) = \frac{976}{4275} - \frac{81}{475} = \frac{247}{4275} \approx .0578.$$

Thus

$$u_P(x_t) - u_P(x_b) = \frac{247}{4275} - \frac{37}{684} = \frac{47}{12825} > 0.$$

So, in this example,

$$u_P(x_t) > u_P(x_b).$$

This shows that the equilibrium in which the prosecutor selects the higher threshold

$x_t = .8$ can yield a strictly higher expected payoff than the x_b equilibrium.

References

Arnold, David, Will Dobbie, and Crystal S. Yang. 2018."Racial Bias in Bail Decisions." *The Quarterly Journal of Economics*, 1885–1932.

Baker, Scott and Claudio Mezzetti. 2001. "Prosecutorial Resources, Plea Bargaining, and the Decision to Go to Trial." *Journal of Law, Economics, & Organization*, 17,1, 149-167.

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

Baliga, Sandeep and Tomas Sjöström (2009) "Contracting with Third Parties" *American Economic Journal: Microeconomics* **1**, 75–100.

Bar-Gill, Oren and Oren Gazal Ayal. 2006. "Plea Bargains Only for the Guilty." *The Journal of Law and Economics*, 49: 1, 353–364.

Bibas, Stephanos. 2004. "Plea Bargaining outside the Shadow of Trial." *Harvard Law Review*, 117: 8, 2463–2547

Bull, Jesse. 2008a. "Mechanism Design with Moderate Evidence Cost." *B.E. Journal of Theoretical Economics*, 8, article 15.

Bull, Jesse. 2008b. "Costly Evidence Production and the Limits of Verifiability,"" *B.E. Journal of Theoretical Economics* 8 article 18.

Bull, Jesse. 2012a. "Interrogation and Evidence Fabrication," working paper.

Bull, Jesse. 2012b. "Third-Party Budget Breakers and Side Contracting in Team Production." *Economics Bulletin* 32: 3.

Bull, Jesse. 2025. ""Jury Priors and Observable Defendant Characteristics," *International Review of Law & Economics*, 81.

Bull, Jesse and Joel Watson. 2019. "Statistical Evidence and the Problem of Robust Litigation." *The RAND Journal of Economics*, 50, 4: 974–1003.

Bull, Jesse and Joel Watson. 2007. "Hard Evidence and Mechanism Design." *Games and Economic Behavior*, 58, 75–93.

Bull, Jesse and Joel Watson. 2004. "Evidence Disclosure and Verifiability." *Journal of Economic Theory*, 118: 1–31.

Bull, Ray. 2018. "PEACE-ful Interviewing/Interrogation: What Research Can Tell Us." In Shigemasu, Kazuo, Sonoko Kuwano, Takao Sato, and Tetsuro Matsuzawa(Eds.), *Diversity in Harmony – Insights from Psychology: Proceedings of the 31st International Congress of Psychology.*

Dobbie, Will, Jacob Goldin, and Crystal S. Yang. 2018. "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, 108(2): 201–240.

Fluet, Claude and Thomas Lanzi. 2019. "Adversarial Persuasion with Cross-Examination," working paper.

Gershowitz, Adam M. 2018." Prosecutorial Dismissals as Teachable Moments (and Databases) for the Police." *George Washington Law Review*, 86(6), 1525-1551.

Grossman, Gene M. and Michael L. Katz. 1983. "Plea Bargaining and Social Welfare." *The American Economic Review*, 73, No. 4: 749–757.

Gudjonsson, Gisli H. and John Pearse. 2011. "Suspect Interviews and False Confessions." *Current Directions in Psychological Science*, 20(1): 33–37.

Hamann, K. and J. Delaney. 2018. "Investigating Violent Crime:The Prosecutor's Role." Bureau of Justice Assistance, Office of Justice Programs, U.S. Department of Justice.

DOI: 10.1093/acprof:osobl/9780199844807.003.0005

Inbau, Fred E., John E. Reid, Joseph P. Buckley, and Brian C. Jayne. 2013. *Criminal Interrogation and Confessions*, 5th Edition. Jones & Bartlett Learning.

Ispano, Alessandro and Péter Vida. 2020. "Custodial Interrogations." working paper.

Ispano, Alessandro and Péter Vida. 2023. "Designing Interrogations." working paper.

Kassin, Saul M., Richard A. Leo, Christian A. Meissner, Kimberley D. Richman, Lori H. Colwell, Amy-May Leach, and Dana La Fon. 2007 "Police Interviewing and Interrogation: A Self-Report Survey of Police Practices and Beliefs." *Law and Human Behavior*, 31: 381–400.

Leo, Rchard A. 2008. Police Interrogation and American Justice. Cambridge, MA: Harvard University Press.

Leo, Richard A. and Richard Ofshe. 1998. "The Consequences of False Confessions: Deprivations of Liberty and Miscarriages of Justice in the Age of Psychological Interrogation." *Journal of Criminal Law and Criminology*, 88: 429.

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

Lundberg, Alexander, 2019. "Leniency Can Increase Deterrence." *International Review of Law and Economics*, 60.

Lundberg, Alexander and Murat Mungan (2022), "The Effect of Evidentiary Rules on Conviction Rates," *Journal of Economic Behavior and Organization*. 203, 563–576.

Mialon, Hugo M. 2005. "An Economic Theory of the Fifth Amendment." *The RAND Journal of Economics*, 36, 4: 833–848.

Mueller-Smith, M. and K. Schnepel. 2020. "Diversion in the Criminal Justice System," working paper.

Mungan, Murat and Jonathan Klick. 2016. "Reducing False Guilty Pleas and Wrongful Convictions through Exoneree Compensation." *The Journal of Law and Economics*, 59, 173–189.

Ofshe, Richard J. and Richard A. Leo. 1997. "The Decision To Confess Falsely: Rational Choice and Irrational Action." *Denver University Law Review*, 74: 979–1122.

Polinsky, A. Mitchell and Steven Shavell. 2007. "The Theory of Public Enforcement of Law," in *Handbook of Law and Economics*, A. Mitchell Polinsky and Steven Shavell, eds., Amsterdam: Elsevier Science Publishing.

Redlich, Allison D., Shi Yan, Robert J. Norris, Shawn D. Bushway. 2018. "The Influence of Confessions on Guilty Pleas and Plea Discounts." *Psychology Public Policy and Law.* 24, 2: 147–157.

Rosenberg, David and Steven M. Shavell. 2006. "A Solution to the Problem of Nuisance Suits: The Option to Have the Court Bar Settlement." *International Review of Law and Economics*, 26: 42.

Sanchirico, Chris. 1999. "Relying on the Information of Interested - and Potentially Dishonest - Parties." *American Law and Economics Review.*

Sanchirico, Chris. 2000. "Games, Information, and Evidence Production: With Application to English Legal History." *American Law and Economics Review*, 2: 342–380.

Sanchirico, Chris. 2001. "Character Evidence and the Object of Trial." *Columbia Law Review*, 101: 1227.

Sanchirico, Chris. 2010. "Models of Evidence: Survey and Assessment." University of Pennsylvania Institute for Law & Economics Research Paper No. 10-28.

Sanchirico, Chris and George Triantis. 2005. "Evidentiary Arbitrage: the Fabrication of Evidence and the Verifiability of Contract Performance." *Journal of Law, Economics, and Organization*, 24:1.

Seidman, Daniel J., and Alex Stein. 2000. "The Right to Silence Helps the Innocent: A Game-Theoretic Analysis of the Fifth Amendment Privilege." *Harvard Law Review*, 114, 2: 430–510.

Shavell, Steven M. 1993. "Suit versus Settlement When Parties Seek Nonmonetary Judgments." *Journal of Legal Studies*, 22: 1.

Shavell, Steven M. 1989. "Sharing of Information Prior to Settlement or Litigation." *Rand Journal of Economics*, 20: 183–195.

Watson, Joel. 2025. "Perfect Bayesian Equilibrium: Consistency Conditions for Practical Definitions." Manuscript, University of California, San Diego.

Weisselberg, Charles D. 2001. "In the Stationhouse after Dickerson." *Michigan Law Review*, 99: 1121–1167.

White, Welsh S. 2001. "Miranda's Failure to Restrain Pernicious Interrogation Practices." *Michigan Law Review*, 99: 1211–1247.