# Judicial Mechanism Design

Ron Siegel and Bruno Strulovici[*]

December 23, 2021

**Abstract**

This paper proposes a mechanism design approach to study criminal justice systems. We derive properties of optimal mechanisms for two notions of welfare distinguished by their treatment of deterrence. These properties (i) provide insights into the effects of defendants' private information about their guilt, (ii) highlight forces that may underlie certain features of existing systems, such as plea bargaining and binary verdicts and the separation of fact finding and sentencing, and (iii) indicate directions for possible improvements of criminal trials, such as varying the standard for conviction across crimes.

## 1 Introduction

Every criminal justice system must contend with the following challenges: (i) determine whether defendants are guilty, (ii) administer appropriate sentences, and (iii) achieve a satisfactory level of deterrence. This paper proposes a mechanism design approach to study these challenges. We consider a stylized criminal justice setting that focuses on a defendant's private information regarding his guilt, and distinguish two notions of welfare. The first notion aims to appropriately punish guilty defendants and avoid punishing innocent ones. The second notion aims to also deter potential criminals. We abstract from idiosyncratic features of criminal justice systems such as jury use and selection and procedures for calling witnesses, and from private information on dimensions other than a defendant's guilt, such as a defendant's degree of risk aversion. We model the criminal justice process as a direct-revelation mechanism with a single agent, the defendant, who privately knows whether he is guilty, and a signal

[*] We are grateful for questions, discussions, and comments from multiple audiences. Strulovici acknowledges financial support from an NSF CAREER Award (Grant No. 1151410) and a fellowship from the Alfred P. Sloan Foundation. Part of this research was conducted while Strulovici was visiting the University of Tokyo, whose hospitality is gratefully acknowledged. Siegel: Department of Economics, The Pennsylvania State University, University Park, PA 16801, rus41@psu.edu. Strulovici: Department of Economics, Northwestern University, Evanston, IL 60208, b-strulovici@northwestern.edu.

regarding the defendant's guilt. The mechanism maps the signal regarding the defendant's guilt and the defendant's report to a (possibly random) sentence.

Our main results describe properties of welfare-maximizing mechanisms. If a defendant reports that he is innocent, then his sentence depends on whether the likelihood ratio of the defendant's guilt implied by the signal exceeds some threshold. Below the threshold, the defendant receives a null sentence; above the threshold, the defendant receives the highest allowable sentence. This holds regardless of whether deterrence is part of the welfare objective. This extreme sentencing scheme is optimal because it is the least attractive one to a guilty defendant who considers reporting that he is innocent.[1] If the defendant reports that he is guilty, he receives either a fixed sentence or a lottery over two sentences, which can be chosen to be independent of the signal. The fixed sentence is optimal when ignoring deterrence if the welfare associated with a guilty defendant is concave in the sentence. Non-concavity or significant deterrence considerations can make a two-sentence lottery optimal.

Welfare-maximizing mechanisms can be implemented as follows. The defendant is first offered a "plea bargain," which carries a possibly random sentence. If he rejects the plea bargain, a "trial" generates "evidence" (the signal) and ends either with an "acquittal," which carries no punishment, or a "conviction," which carries a sentence more severe than the one associated with a plea bargain. *Plea bargains, binary trial verdicts, and no punishment following an acquittal are not assumed in our analysis. Instead, they emerge endogenously as features of optimal mechanisms.*

To derive these results, the key modeling challenge is to specify the set of mechanisms available to the designer. Our main assumption (Assumption 1) requires that given an available mechanism, modifying the sentence function in a way that does not violate the defendant's incentive to report truthfully always yields a mechanism that is also available to the designer. This assumption is consistent with the usual assumption in mechanism design that the designer can commit to any mapping from messages to outcomes. Intuitively, the assumption allows the designer to choose the mechanism's features (sentences) without worrying about their impact on the quality of the signal that she receives about the defendant. The assumption allows us to abstract from the complex game of incomplete information played by the different agents in the judicial system, which may have multiple equilibria and involve additional incentive compatibility constraints for these agents. This assumption is part of our focus on the defendant's incentive constraints, since it can be interpreted as either abstracting from the incentive constraints of the other (unmodeled) agents that are part of the signal-forming process (such as witnesses and jurors), or as giving the designer the ability to choose (unmodeled) agents that

[1] This screening motive differs from the enforcement cost consideration leading to extreme sentences in the work of Becker (1968).

agree with the designer's objective.[2]

Our stylized model abstracts from several realistic features of criminal justice systems. In particular, our main assumption, while in line with commitment assumptions common in the mechanism design literature, is strong. We also assume that the defendant's private information consists only of whether he has committed the crime. These and other simplifications of our stylized model lead to some features of the welfare-maximizing mechanisms that differ from what we see in practice. Welfare-maximizing mechanisms achieve perfect screening: guilty defendants admit their guilt and get a "plea bargain" and innocent defendants go through a "trial." In addition, welfare-maximizing mechanisms use "evidence" (the signal) only to determine the "verdict" (a null sentence or a high sentence) but not to determine the magnitude of the sentence conditional on a "conviction."

Despite these differences, our analysis highlights several important forces. Commitment in our setting makes it possible to punish innocent defendants; various actual practices, such as the separate roles of juries and judges and keeping jurors uninformed of the consequences of a conviction, may indicate that commitment has value in criminal justice systems. Screening, which drives the use of plea bargains in optimal mechanisms, may also be a driver of plea bargains in reality. No punishment following an acquittal and a relatively high penalty following a conviction make sense when ignoring type I and type II errors regarding the defendant's guilt; our analysis shows that they optimally support screening when these errors are taken into account. Evidence (the signal) is used in the optimal mechanisms only to incentivize defendants to report whether they are guilty. This role may also be important in reality in addition to using evidence to determine guilt.

Our results also suggest potential improvements to existing systems. One direction is the possibility of a more flexible conviction threshold. The current threshold of "Beyond a reasonable doubt" in the United States is independent of the crime and circumstances. In contrast, the conviction threshold in the optimal mechanisms may vary across crimes and (unmodeled) circumstances.[3] Another direction is the use of random sentences. Optimal mechanisms may include binary sentence lotteries following an admission of guilt if, for example, society is significantly less risk averse than the defendant. Uncertainty regarding the sentence following a plea bargain may achieve this in practice. Exploring the extent of these potential improvements would, of course, require a more elaborate model than the one we consider here.

[2] To the extent that additional incentive constraints arise in reality, the set of mechanisms that we consider here is likely a superset of those available in practice. In complementary work, Pei and Strulovici (2021) study the effect of the sentencing mechanism on witnesses' incentives to credibly report their observations, but they do not consider a defendant's incentives within the criminal justice system.

[3] This suggestion echoes the analysis of Kaplow (2011) on the optimal burden of proof.

More broadly, we hope that our analysis demonstrates the potential value of using a mechanism design approach to study criminal justice systems. Future work could consider less stylized models and weaker commitment assumptions, thereby uncovering more nuanced insights that may match additional realistic features of criminal justice systems.

Plea bargaining, criminal trials, and other features of criminal justice systems have been the subject of a large economics literature. Early work, pioneered by Becker (1968) and Stigler (1970), used equilibrium analysis to study law enforcement and criminal justice. A well-known result of Becker's analysis is that extreme sentences are optimal when the main objective is to reduce costs of law enforcement and only guilty defendants risk being apprehended, i.e., there is no risk of committing Type I errors. Grossman and Katz (1983), Baker and Mezzetti (2001), Kaplow (2017), and Daughety and Reinganum (2016a,b), take a game theoretic approach that specifies various features, such as plea bargaining and binary sentences as part of the game.

Our paper takes a mechanism design approach that does not specify features like plea bargaining and binary sentences in advance. Mechanism design has been applied to tort law by Spier (1994), Klement and Neeman (2005), and Demougin and Fluet (2006), who analyzed settlement and fee-shifting rules between plaintiffs and defendants. Silva (2019) used a mechanism design approach to study a setting with multiple defendants in which an admission of guilt by one defendant had informational spillovers for other defendants.

Self-reporting of harmful activities is a central feature of several earlier works, particularly Kaplow and Shavell (1994) and Innes (1999, 2000). Our analysis shares several features with Kaplow and Shavell (1994). An important difference is that signals in Kaplow and Shavell are perfectly revealing, which precludes the risk of Type I and Type II errors conditional on acquiring the signal.[4] Innes pays particular attention to the possibility that self-reporting defendants may remedy the harm they caused.

Private information about dimensions other than guilt, such as risk aversion, has been considered by various papers, such as Jordan (2020) in the context of trial and sentence differences across racial groups.

In our paper, a key assumption (Assumption 1) is that the mechanism designer can elicit information from the defendant before the evidence is revealed and can modify the sentencing scheme without affecting the available evidence. Pei and Strulovici (2021) consider specific adjudication systems in which defendants are arrested on the basis of witness reports whose credibility depends endogenously on the sentencing scheme chosen by the designer. Unlike in the present paper, they assume that

[4] The signal is therefore binary, since guilt is also binary, and the question of how to set sentences as the function of the strength of the evidence does not arise in their setting.

witnesses perfectly observe the act of the defendant.

Our welfare analysis considers both interim and ex ante perspectives. Most of the literature, such as Grossman and Katz (1983), considers an interim perspective. Notable exceptions include Polinsky and Rubinfeld (1988) and Reinganum (1993), who consider an ex-ante perspective and deterrence.

The mechanism design problem studied in this paper differs from standard mechanism design analysis in several ways. First, the designer has only one instrument, the sentence, at his disposal, and the defendant's utility need not be linear in the sentence. Second, there is an exogenous signal regarding the defendant's guilt, which can be thought of as "hard information," but regarding which there is no disclosure decision. Third, the defendant's type does not determine the defendant's preferences over the sentence, but rather the distribution of the signal regarding his guilt that different actions will generate. Fourth, in a standard mechanism design setting all mappings from messages to outcomes are available, whereas in our setting the set of available mechanisms may be limited by technological and other (unmodeled) constraints. Our main assumption (Assumption 1) provides enough structure on the set of available mechanisms to enable a mechanism design exercise.

The rest of the paper is organized as follows. Section 2 describes mechanisms, the main assumption on the set of available mechanisms, and the notions of welfare. Section 3 derives properties of optimal mechanisms. Section 4 discusses the results and concludes.

## 2 Judicial Mechanisms

We consider two perspectives: after a crime has been committed (interim perspective) and before it was committed (ex-ante perspective).

### 2.1 Interim Perspective

The interim perspective starts when an individual has been arrested in relation to a crime. This individual—hereafter, the defendant—is either guilty or innocent, and his type $\theta \in \Theta = \{g, i\}$ is privately known. The probability of guilt at the time of the arrest is $\lambda \in (0,1)$. The arrest gives rise to a judicial process, which may involve many agents and stages.

We model the process in reduced form as a single-agent mechanism focused on the defendant. This process produces, at some social cost, a signal about the defendant's guilt and assigns a sentence to the defendant. The cost and signal can depend on the defendant's type. For example, a guilty defendant is more likely to have left incriminating evidence at the crime scene, which may also affect the cost of searching for evidence. The cost and signal can also depend on the defendant's actions during the judicial process. We model those actions in reduced form as the defendant sending a report $\hat{\theta} \in \{\hat{g}, \hat{\imath}\}$

about his type. A defendant of one type can claim to be of the other type, but truth-telling is optimal for the defendant. The sentence depends on the defendant's reported type and the signal generated by the judicial process, but not on the defendant's true type.

Formally, a (direct-revelation) *judicial mechanism* is a tuple $(F, C, S)$. The first component is $F = \left(F_i^{\hat{\imath}}, F_g^{\hat{\imath}}, F_i^{\hat{g}}, F_g^{\hat{g}}\right)$, where $F_\theta^{\hat{\theta}}$ for each pair $(\theta, \hat{\theta})$ of actual and reported types of the defendant is a distribution over the set of signals $T = [0, 1]$ and, without loss of generality, $F_g^{\hat{\theta}}$ dominates the distribution $F_i^{\hat{\theta}}$ according to the monotone likelihood ratio property (MLRP). To simplify the analysis, we assume that $F_\theta^{\hat{\theta}}$ has a continuous density $f_\theta^{\hat{\theta}}$ and full support, and that the ratio $f_g^{\hat{\theta}}(t)/f_i^{\hat{\theta}}(t)$ *strictly* increases in $t$ (strict MLRP).[5] The second component $C = \left(C_i^{\hat{\imath}}, C_g^{\hat{\imath}}, C_i^{\hat{g}}, C_g^{\hat{g}}\right)$ assigns a cost for each pair $(\theta, \hat{\theta})$. The third component $S : (t, \hat{\theta}) \mapsto S(t, \hat{\theta})$ is a measurable *sentencing scheme* that maps the signal $t$ and the defendant's report $\hat{\theta}$ into a lottery over sentences $s \in [0, \bar{s}]$, where the upper bound $\bar{s}$ is crime-specific and exogenously imposed. The upper bound $\bar{s}$ may be viewed as a technical or ethical constraint on punishment, and is standard in the literature (see, for example, Becker (1968), Grossman and Katz (1983), and Kaplow (2011)).[6] The lower bound 0 on the sentence means that the defendant cannot be rewarded or compensated by the mechanism, which is consistent with existing practice.[7]

### 2.1.1 Invariance Assumption

A central modeling challenge is to determine the set of judicial mechanisms available to the designer. Realistically, the distribution tuples $F$ that are available should depend on the forensic technology available for producing evidence and on the interaction between various unmodeled agents (judge, jurors, witnesses, etc.). In general, there may be many constraints on $F$, which are context specific.

Our approach is to focus our analysis on the sentencing scheme $S$ and to assume that the designer can change the sentencing scheme without affecting the process of generating the signal, i.e., without affecting $F$. More precisely, given an available mechanism $(F, C, S)$ in which the defendant optimally reports his type truthfully, any mechanism $(F, C, \tilde{S})$ that results from changing the sentencing scheme to $\tilde{S}$ is also available, provided that the change does not incentivize the defendant to misreport his type.

[5] If atoms were allowed, they could be decomposed into an interval of signals corresponding to a constant likelihood ratio. The constructions used in the proofs of Theorems 1 and 2 go through but the optimal sentence scheme will generically involve randomization over two extreme sentences when the signal observed has the corresponding likelihood ratio.

[6] For example, the actual number of years that a defendant can spend in prison is naturally bounded. The Eighth Amendment of the United States Constitution bans "cruel and unusual" punishments and "excessive fines" (United States v. Bajakajian (1998)), which provides another upper bound justification. Instead of imposing $\bar{s}$ directly, we could assume that the ex-post welfare functions introduced later in this section are infinitely negative beyond the level $\bar{s}$. We impose the bound directly for simplicity.

[7] One could allow bounded rewards (i.e., negative sentences) for the defendant without affecting the analysis.

This assumption allows us to focus on the defendant's private information regarding his guilt as the main constraint that the designer faces.[8] We provide several interpretations after the formal definition.

Let $\mathcal{F}$ denote the set of all distribution tuples $F = \left(F_i^{\hat{\imath}}, F_g^{\hat{\imath}}, F_i^{\hat{g}}, F_g^{\hat{g}}\right)$ and $\mathcal{S}$ denote the set of all sentencing schemes. Given a tuple $F \in \mathcal{F}$, say that a sentencing scheme $S \in \mathcal{S}$ is **$F$-truthful** if truth-telling is optimal for the defendant given $(F, S)$:

$$E[u(S(t, \hat{g}))|F_g^{\hat{g}}] \geq E[u(S(t, \hat{\imath}))|F_g^{\hat{\imath}}] \tag{1}$$

$$E[u(S(t, \hat{\imath}))|F_i^{\hat{\imath}}] \geq E[u(S(t, \hat{g}))|F_i^{\hat{g}}] \tag{2}$$

where (i) $u(s)$ is the defendant's utility from sentence $s$ and (ii) expectations are taken with respect to the signal realization and, whenever the sentencing scheme $S$ involves randomization over sentences, the realization of the corresponding lottery. We assume that $u(\cdot)$ is continuous and strictly decreases in $s$, with $u(0) = 0$. We refer to a judicial mechanism $(F, C, S)$ in which the sentencing scheme $S$ is $F$-truthful, as a *truthful judicial mechanism.*

Let $\mathcal{M}$ denote the set of truthful judicial mechanisms available to the designer.

**Assumption 1** *If $(F, C, S) \in \mathcal{M}$, and $\tilde{S} \in \mathcal{S}$ is $F$-truthful, then $\left(F, C, \tilde{S}\right) \in \mathcal{M}$.*

Assumption 1 is often imposed, implicitly or explicitly, in law and economics.[9] In the present context, Assumption 1 allows us to separate the process of generating evidence, captured by $F$, from the sentencing stage, captured by $S$. This separation is consistent with the commitment assumption in standard mechanism design, where the designer can commit to any mapping from messages to outcomes and is subject to the agent's incentive constraints.

While it is commonly used, Assumption 1 is demanding as it requires that the choice of sentences does not affect the evidence available to the court as long as it does not affect the defendant's reporting behavior during the judicial process. Criminal justice systems in practice likely impose additional incentive constraints beyond (1) and (2). Considering such additional constraints could prove an interesting avenue for future research.[10]

Importantly, existing judicial practices often support the kind of separation captured by Assumption 1 or express the will to impose such a separation. Such practices include: (i) incentivizing one

[8] It would be interesting to combine this approach with the incentives of other agents, such as prosecutors and jurors. The first step we take in this paper should prove useful in considering these more complex questions.

[9] For example, Grossman and Katz (1983) assume that the probabilities of "guilty" and "not-guilty" verdicts are independent of the plea bargaining and conviction sentences. Similarly, Kaplow (2011) assumes that the signal distributions generated by guilty and innocent defendants are independent of the conviction threshold.

[10] To the extent that additional constraints arise in reality, the set of mechanisms considered here is likely a superset of those available in practice.

party to produce incriminating evidence and the other party to produce exculpatory evidence regardless of the eventual sentence;[11] (ii) separating the jury's fact-finding role from the judge's sentencing role;[12] and (iii) specifically instructing jurors to focus solely on determining the defendant's guilt.

Assumption 1 can also be interpreted as a richness assumption—if changing the sentence affects the evidence generation process (through its affect of juror's incentives, for example), the designer can replace the unmodeled agents with other agents to restore $F$. The mechanism design exercise we conduct uses Assumption 1 to identify properties of any sentencing scheme that is part of an optimal mechanism.

### 2.1.2 Welfare

From an interim perspective, society aims to punish guilty defendants and avoid punishing innocent ones, taking into account the cost of producing evidence and administering punishment. To capture this objective, we introduce conditional welfare functions: $W(s, \theta)$ denotes the social welfare corresponding to imposing a sentence $s$ on a defendant of type $\theta$. Any monetary cost of imposing the sentence, such as the cost of incarceration, is included in $W$.

**Assumption 2** *The welfare function $W$ satisfies the following conditions:*

- $W(s, \theta) \leq 0$ *for all* $(s, \theta) \in [0, \bar{s}] \times \{g, i\}$.[13]
- $W(s, i) = \phi(u(s))$ *where* $\phi : \mathbb{R}_- \to \mathbb{R}_-$ *is weakly convex and strictly increasing.*
- $W(s, g)$ *is continuous in* $s$.

The second point in this assumption does not imply that $W(s, i)$ is convex but only that it is less concave than $u$. This assumption captures the idea that society is comparatively less sensitive to increases in the sentence given to an innocent defendant than the defendant himself. The assumption may also be viewed as pertaining to the uncertainty caused by an imperfect signal about the defendant's guilt. From this perspective, the assumption means that society is weakly less averse than the defendant regarding uncertain sentences.

[11] The benefits of this practice were already noted by the High Lord Chancellor of the United Kingdom in 1822, who wrote that "truth is best discovered by powerful statement on both sides of the quest." See also Shin (1998), Dewatripont and Tirole (1999), and Deffains and Demougin (2008).

[12] For example, in *United States v. Patrick* (D.C. Circuit, 1974), the court affirmed that the jury's role is limited to a determination of guilt or innocence. See Sauer (1995) for a detailed study of this separation of tasks.

[13] The non-positivity guarantees that interim welfare is never so high as to offset the harm caused by the crime, and could be relaxed as long as this latter property holds.

The second point in the assumption also implies that $W(s,i)$ is strictly decreasing in $s$, which means that it is socially harmful to punish innocent defendants. A particular case is $W(s,i) = u(s)$ (so $\phi$ is the identity function), which is the assumption made by Grossman and Katz (1983) in their analysis of plea bargains.

Given a probability $\lambda$ that the defendant is guilty, the expected welfare from giving a sentence $s$ to the defendant is

$$\lambda W\left(s,g\right) + (1-\lambda) W\left(s,i\right).$$

To allow for stochastic sentencing, let $W\left(\tilde{s},\theta\right)$ denote the expected welfare from imposing a (possibly) random sentence $\tilde{s}$ on a defendant of type $\theta$.[14] Given a truthful judicial mechanism $(F,C,S)$, the resulting *interim welfare* is

$$\lambda\left(\left(\int_0^1 W(S(t,\hat{g}),g) f_g^{\hat{g}}(t)dt\right) - C_g^{\hat{g}}\right) + (1-\lambda)\left(\left(\int_0^1 W(S(t,\hat{\imath}),i) f_i^{\hat{\imath}}(t)dt\right) - C_i^{\hat{\imath}}\right). \tag{3}$$

## 2.2 Ex-Ante Perspective

From an ex-ante perspective, society also wishes to deter crime. For simplicity, we focus on a specific crime, which entails harm $h > 0$ for society. If an individual commits this crime, he obtains an idiosyncratic benefit $b$ (in utility terms) but faces a probability $\pi_g > 0$ of being arrested and prosecuted. For simplicity, we treat $\pi_g$ as exogenous.[15] We assume that at most one individual is prosecuted for the crime.[16] The planner chooses a judicial mechanism before individuals decide whether to commit the crime.

Given a truthful judicial mechanism $(F,C,S)$, an individual with benefit $b$ commits the crime if

$$b + \pi_g\left(\int_0^1 u(S(t,\hat{g})) f_g^{\hat{g}}(t)dt\right) > 0. \tag{4}$$

There is a measure 1 of individuals in the population, who vary in the benefit $b$ from committing the crime. The distribution of $b$ in the population is described by a probability measure $G_b$. Letting $H(F,S)$ denote the fraction of individuals who commit the crime, i.e., the number of instances of the crime, we have

$$H\left(F,S\right) = 1 - G_b\left(-\pi_g\left(\int_0^1 u(S(t,\hat{g})) f_g^{\hat{g}}(t)dt\right)\right). \tag{5}$$

[14] Formally, if $\tilde{s}$ represents a probability distribution over sentences in $[0,\bar{s}]$, then $W(\tilde{s},\theta) = \int W(s,\theta)d\tilde{s}(s)$.

[15] This probability can be endogenized by including the amount of costly law enforcement as a decision variable. This would not change any of the results, which would apply to the optimal level of law enforcement.

[16] This allows us to abstract from interdependencies between multiple defendants, an issue that is tangential to the focus of this paper. See Silva (2019) for an analysis of this issue.

In a large society, the probability that a *specific* individual is arrested for a particular crime that he did not commit is small. Furthermore, such erroneous arrests affect both individuals who committed a (different) crime and those who are innocent. For simplicity, these events are omitted from the incentive equation (4).[17] However, the probability that *some* innocent individual is arrested for the crime is not negligible, as this probability is the aggregation of infinitesimally small probabilities across a large number of individuals.

Let $\pi_i \in (0, 1-\pi_g]$ denote the probability, for any given instance of the crime, that some innocent individual is arrested for this crime. The ex-ante social welfare corresponding to a truthful judicial mechanism $(F, C, S)$ is

$$H(F,S)\left[\pi_g\left(\left(\int_0^1 W(S(t,\hat{g}),g)f_g^{\hat{g}}(t)dt\right)-C_g^{\hat{g}}\right)+\pi_i\left(\left(\int_0^1 W(S(t,\hat{\imath}),i)f_i^{\hat{\imath}}(t)dt\right)-C_i^{\hat{\imath}}\right)-h\right]. \quad (6)$$

The relation between (6) and (3) is as follows. First, by the time an individual is arrested, the crime has already been committed, so from an interim perspective the social harm $h$ from the crime is "sunk" and omitted. Second, an arrested individual's probability of guilt is $\lambda = \pi_g/(\pi_g+\pi_i)$. Using these observations, we recover (3) from (6).

# 3 Optimal Judicial Mechanisms

This section derives properties of optimal judicial mechanisms for interim and ex-ante welfare objectives.

## 3.1 Interim Welfare

A judicial mechanism $(F, C, S)$ in $\mathcal{M}$ is *interim optimal* if, given the prior probability $\lambda$ that the defendant is guilty, the mechanism maximizes interim welfare (3) among all the mechanisms in $\mathcal{M}$. Studying interim-optimal mechanisms allows us to disentangle deterrence from other welfare considerations and makes the arguments in the proof easier to follow.

Our first theorem describes properties of interim optimal mechanisms. One of the properties requires the following assumption:

**Assumption 3** *Functions $W(\cdot, g)$ and $u(\cdot)$ are concave and at least one is strictly concave. Function $W(\cdot, g)$ has a unique maximizer, denoted $\hat{s}$.*

[17] Our results hold even if the probability of being wrongfully arrested and convicted has a non-negligible impact on the expected utility from not committing the crime, because the welfare-improving mechanisms constructed in Section 3 keep the expected utility of a guilty defendant unchanged and increase the expected utility of an innocent defendant. If the probability that a given innocent individual is arrested and convicted is treated as strictly positive, the constructed mechanisms would have the additional benefit of increasing deterrence by increasing the utility differential between an innocent defendant and a guilty one.

**Theorem 1** *The following holds:*[18]

*(i) In any interim-optimal mechanism, the sentence received by an innocent defendant is a step function of the signal $t$, which jumps from $0$ to $\bar{s}$ at some cutoff $\bar{t}$.*

*(ii) There is an interim-optimal mechanism such that the sentence received by a guilty defendant is* independent of the signal $t$*, and is either deterministic (i.e., with a one-point support) or a random variable with a two-point support. Moreover, for a* generic *set of welfare functions, the support is the same across all interim-optimal mechanisms.*[19]

*(iii) If Assumption 3 holds, the guilty defendant's sentence is deterministic and does not exceed $\hat{s}$.*

*(iv) If the guilty defendant's sentence has a two-point support and $W(\cdot, g)$ is single-peaked in $s$ with maximizer $\hat{s}$, then the two-point support lies in $[0, \hat{s}]$.*

*(v) The guilty defendant is indifferent between reporting truthfully and misreporting, i.e.,* (1) *holds as an equality.*

*When Assumption 3 holds, any mechanism that violates (i) or (iii) is strictly welfare-dominated for* all *non-degenerate priors $\lambda$ by a* single mechanism *with these properties.*

**Comments on Theorem 1**

Theorem 1 shows that interim-optimal mechanisms resemble a system in which plea bargains are available and trials end in one of two verdicts. If the defendant accepts the plea by pleading "guilty" he forgoes a trial and receives a sentence that is without loss of generality independent of the evidence against him.[20] Otherwise, he faces a trial, in which he may be acquitted and receive a null sentence or convicted and receive a high sentence. He is convicted if the evidence against him is sufficiently strong (above some threshold). The availability of a "plea bargain," a binary verdict following a "trial," and a null sentence following an acquittal are not assumed features of the mechanism, and instead emerge as part of any optimal mechanism.[21]

The last statement of Theorem 1 reveals its non-Bayesian nature.[22] In fact, the proof of Theorem 1 proceeds by showing that, starting from any mechanism that violates the properties of the theorem, there is another mechanism with these properties that improves upon the initial mechanism *conditional*

[18] All statements in this section are up to a set signals that has probability 0.

[19] "Generically" is in the sense of *prevalent sets* over the vector space of welfare functions. See Appendix D.

[20] Although this entails no social benefit, the signal $t$ could in principle be used to perform the randomization when the sentence is stochastic. However, this is strictly suboptimal if generating signal $t$ is at all costly.

[21] In fact, the interim-optimal sentences following an acquittal are strictly positive when pleas are not allowed. See Lando (2005) and Siegel and Strulovici (2020).

[22] For simplicity, the last statement is formulated under Assumption 3. A similar statement holds for a generic set of welfare functions even when Assumption 3 is not imposed.

*on each defendant type $\theta$.* In the language of statistical decision theory, this means that the class of mechanisms described by Theorem 1 a *complete class* (Karlin and Rubin (1956)).[23]

If the defendant pleads guilty, the signal is not used by the mechanism to determine his sentence, even if the signal $t$ conditional on the report $\hat{\theta} = \hat{g}$ is informative about the defendant's guilt. The signal is used only to induce the defendant to reveal his type and prevent deviations. The screening value of plea bargains has already been noted and emphasized by Grossman and Katz (1983), but that paper does not show the *optimality* of plea bargains among other mechanisms: it takes as given the structure of a two-verdict system with a plea bargain. Under Assumption 3, Theorem 1 shows that such a system is in fact *globally optimal* from an interim perspective. When Assumption 3 is not imposed, the deterministic pleas assumed by Grossman and Katz (1983) may be suboptimal. Theorem 1 shows that in this case a stochastic sentence with a two-point support is optimal. In fact, we will see in Theorem 2 that fixed plea sentences are not generally optimal even when Assumption 3 holds but deterrence is taken into account.[24]

Theorem 1 shows that the use of extreme sentences, 0 and $\bar{s}$, is optimal. Extreme sentences are known to be optimal in some models of law enforcement, starting with Becker (1968). In Becker's framework, extreme sentences are used to save on enforcement costs. In our framework, extreme sentences play a different role: they serve to maximize the screening power of plea bargaining.[25] In our framework, the optimal mechanism is sensitive to $\bar{s}$ in two ways: $\bar{s}$ is the sentence given following a conviction and it affects the plea sentence through the incentive compatibility constraint of guilty defendants.

### Intuition for Theorem 1

The proof of Theorem 1 is in the Appendix. It constructs a welfare improvement conditional on each defendant type. The signal is used to devise a sentencing scheme that induces the defendant to report truthfully, and the relevant incentive constraint is dissuading a guilty defendant from reporting that he is innocent. Starting from a truthful judicial mechanism, we construct a mapping from signals to sentences that minimizes the disutility of an innocent defendant subject to maintaining the same expected utility for a guilty defendant as in the initial mechanism. The MLRP of the signal distribution

[23] This result is reminiscent of the Neyman-Pearson lemma and the Karlin-Rubin theorem concerning uniformly most powerful tests, which show that likelihood-based estimators maximize the power of a test subject to a given size. In contrast to these papers, the question here is not whether to accept or reject a hypothesis but how to choose a continuous sentence, and the objective involves not only Type I and Type II errors but also the magnitude of the errors as measured by the sentence given relative to the ideal one.

[24] Grossman and Katz (1983) focus on interim welfare and do not consider deterrence.

[25] Becker's (1968) model assumes that only guilty defendants are punished, so screening is not an issue.

(which, we recall, is without loss of generality) shows that the optimal mapping is the two-step sentence function in part (i) of the theorem. This step does not rely on any concavity assumption for the utility or welfare function and holds without imposing Assumption 3.

When we impose Assumption 3, concavity guarantees that the optimal sentence for a guilty defendant, which is considered in the second step of the proof, must be constant. To see this, suppose by contradiction that a guilty defendant was receiving a stochastic sentence in the initial judicial mechanism. We show that moving from this stochastic sentence to its certainty-equivalent constant sentence relaxes the defendant's incentive constraint and increases social welfare as long as the constant sentence does not exceed $\hat{s}$, the socially optimal sentence conditional on facing a guilty defendant. If it exceeds $\hat{s}$, then we can decrease the sentence to $\hat{s}$, which gives the highest possible social welfare conditional on facing the guilty defendant.[26]

When Assumption 3 is relaxed, it may be optimal to give a guilty defendant a lottery over two sentences, which are different from the ones faced by the innocent defendant. In this case, the key change is as follows: we consider the guilty defendant's expected utility from his sentence, rather than the sentence itself, and use a concavification argument to find the utility distribution that maximizes social welfare while maintaining the guilty defendant's expected utility.

## 3.2 Ex-Ante Welfare and Deterrence

While interim welfare is only concerned with appropriately punishing defendants, ex-ante welfare also takes into account the number of crimes committed. This number depends on the mechanism, because different mechanisms deter crime to different extents. Any modification of a mechanism must take into account the modification's impact on deterrence. The proof of Theorem 1 suggests that under Assumption 3 this consideration need not necessarily lead to a radically different analysis of the optimal sentencing scheme. In that proof, if a guilty defendant's certainty equivalent $s^{ce}$ does not exceed $\hat{s}$ (the socially optimal sentence conditional on facing a guilty defendant), each step of the proof alters the initial mechanism in a way that increases interim welfare but leaves the expected utility of a guilty defendant unchanged. Since this expected utility is unchanged, so is the set of individuals who commit the crime.[27] In this case, therefore, ex-ante welfare also increases. In particular, Theorem 1 identifies

[26] This last point is not generally true for ex-ante optimal mechanisms because reducing the sentence may reduce deterrence.

[27] Recall our simplifying assumption that the ex-ante probability that an individual is arrested for a crime that he did not commit is negligible and/or independent of whether the individual committed another a crime. Therefore, only changes in the expected utility of a guilty defendant affect the incentives to commit crime. Moreover, the improvements constructed to prove Theorem 1 and Theorem 2 increase the expected utility of an innocent defendant, so if this utility

properties of the mechanisms that maximize ex-ante welfare among all available mechanisms in which the certainty equivalent of a guilty defendant does not exceed $\hat{s}$.

In general, however, even under Assumption 3 optimal deterrence may lead to sentences that exceed $\hat{s}$. In this case, the improvements constructed in Theorem 1 require decreasing these sentences. While this increases interim welfare, it also increases the utility of guilty defendants. This increases the set of individuals who commit the crime, and may therefore decrease ex-ante welfare.

Our next result identifies properties of ex-ante optimal mechanisms. A judicial mechanism $(F, C, S)$ in $\mathcal{M}$ is *ex-ante optimal* if the mechanism maximizes ex-ante welfare (6) among all the mechanisms in $\mathcal{M}$. The result shows that ex-ante optimal mechanisms (with or without Assumption 3) are similar to interim optimal mechanisms without Assumption 3. This similarity comes from the fact that our construction in the interim case modifies the sentence function is a way that does not change the guilty defendant's utility and thus does not change the set of individuals who commit the crime. Thus, only some minor adaptations of the proof of Theorem 1 are required to derive Theorem 2. Part (iv) of Theorem 1 must be modified for ex-ante optimal mechanisms because decreasing the guilty's sentence below $\hat{s}$ may increase crime and decrease ex-ante social welfare even when social welfare conditional on facing the guilty is single peaked at $\hat{s}$.

**Theorem 2** *(i) In any ex-ante optimal mechanism, the innocent defendant's sentence is a step function of the signal $t$, which jumps from $0$ to $\bar{s}$ at some cutoff $\bar{t}$.*[28]

*(ii) There is an ex-ante optimal mechanism in which the guilty defendant's sentence is either deterministic and independent of the signal or is a random variable with a two-point support. Moreover, this property must generically hold for any ex-ante optimal mechanism.*[29] *The guilty defendant's sentence can be chosen to be statistically independent of the signal.*

*(iii) If the guilty defendant's sentence in an ex-ante optimal mechanism is random with a two-point support and $W(\cdot, g)$ is single-peaked at $\hat{s}$, then the two-point support lies in $[0, \hat{s}]$ or in $[\hat{s}, \bar{s}]$, but* cannot *straddle $\hat{s}$. If, in addition, $W(\cdot, g)$ and $u(\cdot)$ are concave and at least one of them is strictly concave, then the two-point support lies in $[\hat{s}, \bar{s}]$.*

*(iv) The guilty defendant is indifferent between reporting truthfully and misreporting, i.e., (1), holds as an equality.*

Theorem 2 shows that it may be optimal to give the guilty defendant a fixed deterministic sentence even when this sentence exceeds $\hat{s}$. For some intuition regarding when a random sentence is optimal,

had any impact on the incentives to commit a crime, these improvements would reduce crime incentives even further.

[28] Necessity follows from Appendix C.

[29] The notion of genericity is the same as in Theorem 1.

suppose that Assumption 3 holds. Then two things must happen for a random sentence to be optimal. First, the optimal level $U^g$ of utility for the guilty defendant must be lower than $u(\hat{s})$, which never happens in an interim optimal mechanism, and happens in an ex-ante optimal mechanism when the tradeoff between deterring individuals from committing the crime and the loss of welfare from punishing the ones who do too severely leans toward deterrence. Second, society must be sufficiently less risk averse than the individuals contemplating committing the crime so that, referring to the notation from the proof of Theorem 3, $\hat{W}$ is not concave below $u(\hat{s})$, and in addition $\hat{W}(U^g) < \bar{W}(U^g)$.[30]

# 4 Discussion

We considered a stylized criminal justice setting and conducted a mechanism design analysis under Assumption 1, which is in the spirit of the commitment assumption in standard mechanism design. Our setting focused on the defendant's private information regarding his guilt, which is a central issue for any criminal justice system, and abstracted from various other realistic features and constraints, which may vary across criminal justice systems. Despite the stylized nature of our setting, the results we obtained deliver several potentially useful insights.

First, the results show that commitment has value in our setting. Optimal mechanisms require punishing innocent defendants with some probability, even though these mechanisms achieve complete separation between guilty and innocent defendants. This indicates that commitment may also have value in actual criminal justice settings, and this in turn may underlie some existing judicial practices, such as separating guilt determination (fact finding) from sentencing and keeping jurors unaware of the magnitude of the sentence following a conviction.[31] Second, the results show that the screening value of plea bargaining, which was first identified by Grossman and Katz (1983), is so large that any interim- or ex ante- optimal mechanism can be thought of as offering the defendant a plea bargain, which may involve stochastic sentences not considered by Grossman and Katz (1983), even without cost savings or other expediency considerations. Third, binary, extreme sentences are optimal because they provide the strongest incentives for defendants to separate while minimizing the disutility to innocent defendants. This incentivizing role is distinct from any ex-post justice considerations that

[30] For example, if $\bar{s} = 4$, $u^{-1}(U) = \sqrt{-U}$, and $W(s) = -2 + s$ for $s \leq 2$ and $2 - s$ for $s > 2$, then for $U^g < -4$ the optimal sentencing scheme randomizes between $s = 2$ and $\bar{s} = 4$.

[31] Recent judicial practice has been to keep the jury uninformed about the punishment faced by the defendant, with the explicit goal of minimizing any undue influence on the jury's decision (Sauer (1995)). As noted by Lee (2014), jurors are generally instructed to reach a verdict based only on the presented evidence (see, for example, the California Code of Civil Procedure - Section 232 (b)). The Capital Jury Project found that most jurors "grossly underestimated" the amount of jail time associated with a guilty verdict.

may in practice make it undesirable or impossible to punish defendants following a "not guilty" verdict. Fourth, the signal is optimally used only to incentivize defendants to admit their guilt. This highlights a potentially important role for evidence in actual criminal justice systems, in addition to its usual role in determining guilt during a trial. Fifth, the standard for conviction in the optimal mechanisms may vary across crimes and circumstances (as captured by the welfare function),[32] but in practice the conviction standard in criminal trials is "beyond a reasonable doubt" (BARD), which is independent of the crime and circumstances.[33] This indicates that a variable conviction threshold may have the potential to improve existing systems. Finally, our results show that the ability to use random sentences following an admission of guilt may be valuable, i.e., lotteries may be an efficient way of a punishing guilty defendants.[34] Random sentences following an admission of guilt could be implemented, for example, with plea bargains that do not specify a particular sentence or ones in which the judge can decide on a sentence other than the one specified without allowing the defendant to withdraw his plea.[35]

These insights also suggest the future potential of following a mechanism design approach to studying criminal justice systems. Additional work could examine features of existing criminal justice systems, such as the fact that some guilty defendants go to trial, which are not captured by the optimal mechanisms in our stylized framework. Assumption 1 could be relaxed by imposing additional constraints, such as jurors taking into account the defendant's decision to reject a plea bargain, which will naturally lead to some guilty defendants going to trial in equilibrium. While adding such constraints goes beyond the scope of this paper, Appendix F shows that Bayesian updating can be accommodated by "nearly optimal" mechanisms in which a small fraction of guilty defendants claim to be innocent. Another direction is allowing for richer private information on the part of the defendant, beyond knowing whether he is guilty or not. For example, the defendant may know how likely he is to have left incriminating evidence at the crime scene. These directions provide interesting avenues for future research.

[32] The evidence (signal) threshold is high for the particular crime and circumstances considered if it is more important to acquit innocent defendants than to punish guilty ones, a preference that will be captured by the welfare functions $W(\cdot, g)$ and $W(\cdot, i)$.

[33] In reality, jurors may interpret BARD differently depending on the severity of the crime, leading to effectively different conviction criteria. Such differences, to the extent they exist, deviate from the usual interpretations of BARD.

[34] Intuitively, the stochastic element that may optimally follow a guilty plea captures the fact that the welfare function conditional on facing a guilty defendant may be locally convex in the defendant's utility, i.e., social preferences may be risk loving in a guilty defendant's utility. This feature can arise at sentence levels at which the ex-post welfare function $W(\cdot, g)$ is decreasing, which creates the possibility that the function $U \rightarrow W(u^{-1}(U), g)$ is convex, even when both $u$ and $W(\cdot, g)$ are concave. (The composition $g \circ f$ of two concave functions is guaranteed to be concave only if $g$ is increasing.)

[35] See Federal Rules of Criminal Procedure 11(c)(1)(C) and 11(c)(1)(B).

# A Proof of Theorem 1

## A.1 Proof of Theorem 1 under Assumption 3

To introduce features of optimal mechanisms gradually, we first prove Theorem 1 under Assumption 3. We show that any available mechanism can be improved upon (weakly) by another available mechanism that satisfies (i) and (iii) in the statement of Theorem 1. Appendix C shows that the improvement is strict if the original mechanism does not satisfy (i) and (iii).

Consider an available mechanism $(F, C, S)$. We modify the sentencing scheme $S$ in a way that maintains truthfulness and increases interim welfare. We do not change the signal distributions $F$ and the cost function $C$. Assumption 1 ensures that the resulting mechanism is also available.

For expositional simplicity we assume in this proof that $W(s,i) = u(s)$. The general case $W(s,i) = \phi(u(s))$ is addressed in Appendix E.

First, we replace the sentence function $S(\cdot,\hat{\imath})$ by a step function $\tilde{S}(\cdot,\hat{\imath})$ with cutoff $\bar{t}$ such that $\tilde{S}(t,\hat{\imath}) = 0$ for $t < \bar{t}$ and $\tilde{S}(t,\hat{\imath}) = \bar{s}$ for $t > \bar{t}$. The cutoff $\bar{t}$ is chosen so that a guilty defendant is indifferent between $S(\cdot,\hat{\imath})$ and $\tilde{S}(\cdot,\hat{\imath})$ when misreporting:

$$\int_0^1 u(\tilde{S}(t,\hat{\imath}))f_g^{\hat{\imath}}(t)dt = u(0)F_g^{\hat{\imath}}([0,\bar{t}]) + u(\bar{s})F_g^{\hat{\imath}}([\bar{t},1]) = \int_0^1 u(S(t,\hat{\imath}))f_g^{\hat{\imath}}(t)dt. \tag{7}$$

The cutoff $\bar{t}$ exists because distribution $F_g^{\hat{\imath}}$ has no atoms.[36] Rearranging (7) yields

$$\int_0^1 [u(S(t,\hat{\imath})) - u(\tilde{S}(t,\hat{\imath}))]f_g^{\hat{\imath}}(t)dt = 0. \tag{8}$$

The function $t \mapsto u(S(t,\hat{\imath})) - u(\tilde{S}(t,\hat{\imath}))$ crosses 0 once from below, since $u(S(t,\hat{\imath}))$ lies in the interval $[u(\bar{s}), u(0)]$ for all $t$ and any sentence function $S(\cdot,\hat{\imath})$, while $u(\tilde{S}(t,\hat{\imath}))$ equals $u(0)$ for $t \le \bar{t}$ and jumps down to $u(\bar{s})$ at $t = \bar{t}$. The density ratio $f_i^{\hat{\imath}}(t)/f_g^{\hat{\imath}}(t)$ is decreasing in $t$, by MLRP. A standard result in comparative statics analysis[37] then implies that

$$\int_0^1 [u(S(t,\hat{\imath})) - u(\tilde{S}(t,\hat{\imath}))]f_i^{\hat{\imath}}(t)dt \le 0. \tag{9}$$

This increases social welfare, provided that truthfulness is maintained. Truthfulness is maintained because (9) and the fact that (2) holds for mechanism $(F, S)$ show that (2) continues to hold when $S(\cdot,\hat{\imath})$ is replaced with $\tilde{S}(\cdot,\hat{\imath})$.

Next, let $s^{ce}$ denote the fixed sentence ("certainty equivalent") that makes a guilty defendant indifferent between $s^{ce}$ and $S(\cdot,\hat{g})$. This means that

$$u(s^{ce}) = \int_0^1 u(S(t,\hat{g}))f_g^{\hat{g}}(t)dt.$$

Denote by $s^a = \int_0^1 E[S(t,\hat{g})]f_g^{\hat{g}}(t)dt$ the average sentence. Then $s^{ce} \ge s^a$ because $u$ is concave and decreasing. Since $W(\cdot,g)$ is also concave, $W(s^a,g) \ge \int_0^1 W(S(t,\hat{g}),g)f_g^{\hat{g}}(t)dt$. Since $W(\cdot,g)$ is single-peaked at $\hat{s}$, it decreases in $s$ for $s \ge \hat{s}$, so if $s^{ce}$ is sufficiently greater than $s^a$, it might be that $W(s^{ce},g) < \int_0^1 W(S(t,\hat{g}),g)f_g^{\hat{g}}(t)dt$.

Thus, to set the welfare-improving constant sentence $s^g$ for a guilty defendant, there are two cases to consider. If $s^{ce}$ is less than $\hat{s}$, we set $s^g = s^{ce}$. Since $s^{ce} \ge s^a$ and $W(\cdot,g)$ is increasing up to $\hat{s}$, we have $W(s^{ce},g) \ge$

[36] If there is an atom at the relevant signal, randomizing between 0 and $\bar{s}$ with the correct probability generates the requisite indifference.

[37] The result is proved by a simple integration by parts, and follows from a result initially proved by Karlin (1968).

$W(s^a, g) \geq \int_0^1 W(S(t,\hat{g}),g) f_g^{\hat{g}}(t)d$, so $s^g$ increases welfare conditional on facing a guilty defendant. If instead $s^{ce} > \hat{s}$, we set $s^g = \hat{s}$. This sentence yields the highest possible social welfare conditional on facing a guilty defendant.

By construction the guilty defendant is indifferent between $s^{ce}$ and reporting truthfully with the sentence function $S(\cdot,\hat{g})$. Since $s^g \leq s^{ce}$, he thus prefers $s^g$ to reporting truthfully with $S(\cdot,\hat{g})$. By construction of $\tilde{S}(\cdot,\hat{\imath})$ and the fact that (1) holds for mechanism $(F,S)$, he prefers reporting truthfully with sentence function $S(\cdot,\hat{g})$ to misreporting with sentence function $\tilde{S}(\cdot,\hat{\imath})$. Thus, he prefers sentence $s^g$ to misreporting with sentence function $\tilde{S}(\cdot,\hat{\imath})$, so truthfulness is maintained for the guilty defendant, i.e., (1) continues to hold when $S(\cdot,\hat{g})$ is replaced with $s^g$.

If (1) holds strictly when $S(\cdot,\hat{g})$ is replaced with $s^g$, we increase the cutoff $\bar{t}$ until the guilty defendant becomes indifferent between $s^g$ and misreporting with sentence function $\tilde{S}(\cdot,\hat{\imath})$. This modification increases welfare since it increases the utility of an innocent defendant. It also maintains truthfulness of an innocent defendant. To see this, note that the indifference condition for a guilty defendant implies that

$$u(s^g) = \int_0^1 u(\tilde{S}(t,\hat{\imath})) f_g^{\hat{\imath}}(t)dt \Rightarrow \int_0^1 [u(s^g) - u(\tilde{S}(t,\hat{\imath}))] f_g^{\hat{\imath}}(t)dt = 0,$$

so, as in the first part of the proof, MLRP implies that

$$\int_0^1 [u(s^g) - u(\tilde{S}(t,\hat{\imath}))] f_i^{\hat{\imath}}(t)dt \leq 0 \Rightarrow \int_0^1 u(\tilde{S}(t,\hat{\imath})) f_i^{\hat{\imath}}(t)dt \geq \int_0^1 u(s^g) f_i^{\hat{g}}(t)dt,$$

where the second inequality follows from the first because $s^g$ is constant.

## A.2 Proof of Theorem 1 without imposing Assumption 3

Consider an available mechanism $(F,S)$. Similarly to the proof when Assumption 3 is imposed, we modify the mechanism by changing the sentence function in a way that maintains truthfulness and increases ex-ante welfare.

We replace the sentence function $S(\cdot,\hat{\imath})$ with a step function $\tilde{S}(\cdot,\hat{\imath})$ that is equal to zero below $\bar{t}$ and equal to $\bar{s}$ above it, with $\bar{t}$ chosen to make a guilty defendant indifferent between $\tilde{S}(\cdot,\hat{\imath})$ and $S(\cdot,\hat{\imath})$ when misreporting his type, so an innocent defendant prefers $\tilde{S}(\cdot,\hat{\imath})$ to $S(\cdot,\hat{\imath})$ when reporting truthfully. The cutoff $\bar{t}$ is now increased until the guilty defendant is indifferent between $S(\cdot,\hat{g})$ and $\tilde{S}(\cdot,\hat{\imath})$. This change increases the utility of an innocent defendant, and therefore social welfare.

We now modify the sentence function $S(\cdot,\hat{g})$ in a way that keeps the guilty defendant's expected utility, $U^g$, unchanged. We wish to find a sentence function $\tilde{S}(\cdot,\hat{g})$ that maximizes social welfare when facing the guilty defendant subject to giving the guilty defendant utility $U^g$. Thus, we are looking for a sentence function $\tilde{S}(\cdot,\hat{g})$ that solves

$$\max_{s(\cdot)\in(\Delta([0,\bar{s}]))^T} \int_0^1 W(s(t),g) f_g^{\hat{g}}(t)dt$$

subject to

$$\int_0^1 u(s(t)) f_g^{\hat{g}}(t)dt = U^g.$$

To solve this problem, it is convenient to reformulate it in terms of the defendant's utility, i.e., to find a mapping from types to lotteries over utilities that solves

$$\max_{\hat{u}(\cdot)\in(\Delta([u(\bar{s}),u(0)]))^T} \int_0^1 E[\hat{W}(\hat{u}(t))] f_g^{\hat{g}}(t)dt \tag{10}$$

subject to

$$\int_0^1 E[\hat{u}(t)] f_g^{\hat{g}}(t)dt = U^g, \tag{11}$$

where $\hat{W}(U) = W\left(u^{-1}(U), g\right)$ for any $U \in [u(\bar{s}), 0]$. The two formulations are equivalent because the defendant's utility $u(\cdot)$ is strictly decreasing in the sentence.

To characterize the solution of (10) subject to (11), it is useful to consider a simpler optimization problem:

$$\max_{\dot{u} \in \Delta([u(\bar{s}), u(0)])} E[\hat{W}(\dot{u})] \tag{12}$$

subject to

$$E[\dot{u}] = U^g. \tag{13}$$

Consider a stochastic process $\hat{u} : T \to \Delta[u(\bar{s}), u(0)]$ whose sample paths are Lebesgue measurable and that satisfies (11). This process induces a measure $F^u$ over $[u(\bar{s}), u(0)]$ such that for any Borel subset $\mathcal{B}$ of $[u(\bar{s}), u(0)]$, $F^u(\mathcal{B}) = \int_0^1 Pr(\{\hat{u}(t) \in \mathcal{B}\}) f_g^{\hat{g}}(t)dt$. Intuitively, $F^u(\mathcal{B})$ is the probability that the defendant receives a utility level in $\mathcal{B}$ given the utility process $\hat{u}$ and given that the signal $t$ is distributed according to $F_g^{\hat{g}}$. Let $\dot{u}$ denote a random variable distributed according to $F^u$. By construction, $\dot{u}$ satisfies (13) and

$$\int_0^1 E[\hat{W}(\hat{u}(t))] f_g^{\hat{g}}(t)dt = E[\hat{W}(\dot{u})]. \tag{14}$$

Therefore, $\hat{u}$ is a solution of (10) subject to (11) if and only if $\dot{u}$ solves (12) subject to (13).

We now solve for (12) subject to (13). For any $U$ in the interval $[u(\bar{s}), u(0)]$, let

$$\bar{W}(U) = \sup\{x : (U, x) \in co(\hat{W})\}$$

where $co(\hat{W})$ denotes the convex hull of the graph of $\hat{W}$. $\bar{W}$ is the *concavification* of $\hat{W}$; it is the smallest concave function that is everywhere above $\hat{W}$.

It is well-known that $\bar{W}(U^g)$ is the value function of the optimization problem (12) subject to (13):[38] If $\hat{W}(U^g) = \bar{W}(U^g)$, the maximal value is achieved by the constant sentence $u^{-1}(U^g)$. In this case, by (14), an optimal $\hat{u}$ is achieved by the sentence function $\tilde{S}(\cdot, \hat{g}) \equiv u^{-1}(U^g)$, which is constant in the signal $t$. If $\hat{W}(U^g) < \bar{W}(U^g)$, the maximal value is achieved by randomizing between $u^{-1}\left(\underline{U}\right)$ and $u^{-1}\left(\overline{U}\right)$, where $\underline{U} = \max\left\{v < U^g : \hat{W}(v) = \bar{W}(v)\right\}$ and $\overline{U} = \min\left\{v > U^g : \hat{W}(v) = \bar{W}(v)\right\}$, with probabilities $\alpha$ and $1 - \alpha$ such that $\alpha\underline{U} + (1-\alpha)\overline{U} = U^g$. In this case, again by (14), the constant stochastic sentence function $\tilde{S}(\cdot, \hat{g})$ (which is independent of the signal) that for every signal $t$ assigns sentence $u^{-1}\left(\underline{U}\right)$ with probability $\alpha$ and sentence $u^{-1}\left(\overline{U}\right)$ with probability $1 - \alpha$ is optimal.

If $W$ is single peaked at $\hat{s}$, then the fact that $u$ is decreasing implies that $\hat{W}$ is single peaked at $u(\hat{s})$, which proves that if $\hat{W}(U^g) < \bar{W}(U^g)$, then the two-point support lies in $[0, \hat{s}]$.[39] If, in addition, $u$ and $W(\cdot, g)$ are concave on $[0, \hat{s}]$, then $\hat{W}$ is concave on the utility interval $[u(\hat{s}), u(0)]$. In this case, $\hat{W}$ coincides with $\bar{W}$ for $U \geq u(\hat{s})$, so $U^g \geq u(\hat{s})$ is optimally achieved by a single sentence.

The resulting mechanism is truthful. Indeed, by construction guilty defendants are indifferent between the sentence functions $\tilde{S}(\cdot, \hat{g})$ and $\tilde{S}(\cdot, \hat{\imath})$, i.e.,

$$\int_0^1 u(\tilde{S}(t, \hat{g})) f_g^{\hat{g}}(t)dt - \int_0^1 u(\tilde{S}(t, \hat{\imath})) f_g^{\hat{\imath}}(t)dt = 0,$$

so (1) holds when $S$ is replaced with $\tilde{S}$. Moreover, since function $\tilde{S}(\cdot, \hat{g})$ is independent of the signal, the last equality can be written as

$$\int_0^1 [u(\tilde{S}(t, \hat{g})) - u(\tilde{S}(t, \hat{\imath}))] f_g^{\hat{\imath}}(t)dt = 0.$$

[38] Concavification with respect to beliefs has been used repeatedly since the works of Aumann and Maschler. See Aumann et al. (1995). Concavification is also used in contract theory to show that a principal's payoff function is concave in the agent's promised utility. See, e.g., Spear and Srivastava (1987).

[39] Sentences higher than $\hat{s}$ can be replaced by $\hat{s}$, which increases interim welfare and relaxes the incentive constraint.

This is equivalent to

$$\int_0^1 [u(s^{ce}) - u(\tilde{S}(t, \hat{\imath}))] f_g^{\hat{\imath}}(t) dt = 0,$$

where $s^{ce}$ is the certainty equivalent of the stochastic sentence $\tilde{S}(t, \hat{g})$ (which is independent of the signal $t$), i.e., $u(s^{ce}) = u(S(t, \hat{g}))$. As in the first and last parts of the proof of Theorem 1, MLRP then implies that

$$\int_0^1 [u(s^{ce}) - u(\tilde{S}(t, \hat{\imath}))] f_i^{\hat{\imath}}(t) dt \leq 0,$$

which is equivalent to

$$\int_0^1 [u(\tilde{S}(t, \hat{g})) - u(\tilde{S}(t, \hat{\imath}))] f_i^{\hat{\imath}}(t) dt \leq 0.$$

This can be written as

$$\int_0^1 u(\tilde{S}(t, \hat{g})) f_i^{\hat{g}}(t) dt - \int_0^1 u(\tilde{S}(t, \hat{\imath})) f_i^{\hat{\imath}}(t) dt \leq 0$$

because $\tilde{S}(\cdot, \hat{g})$ is independent of the signal. This shows that (2) holds when $S$ is replaced with $\tilde{S}$.

Appendix D proves the genericity claim in part (ii).

# B Proof of Theorem 2

Consider an available mechanism $(F, S)$ and construct the same improving available mechanism $(F, \tilde{S})$ as in Appendix A.2. This mechanism also improves ex-ante welfare (6). To see this, note that the two mechanisms lead to the same number of crimes (because they give the same utility $U^g$ to guilty defendants) and have the same cost (because they have the same signal distributions $F$). But function $\tilde{S}(\cdot, \hat{\imath})$ increases the utility of innocent defendants relative to mechanism $S(\cdot, i)$, and therefore increases welfare when facing an innocent defendant, and function $\tilde{S}(\cdot, \hat{g})$ maximizes welfare when facing a guilty defendant among all sentence functions that give the guilty defendant utility $U^g$. Thus, $(F, \tilde{S})$ increases (6). This proves parts (i), (ii), and (iv).

For part (iii), continuing with the notation from Appendix A.2, if $W$ is single peaked at $\hat{s}$, then the fact that $u$ is decreasing implies that $\hat{W}$ is single peaked at $u(\hat{s})$, which proves that the two-point support lies in $[0, \hat{s}]$ or in $[\hat{s}, \bar{s}]$. If, in addition, $u$ and $W(\cdot, g)$ are concave on $[0, \hat{s}]$, then $\hat{W}$ is concave on the utility interval $[u(\hat{s}), u(0)]$. In this case, $\hat{W}$ coincides with $\bar{W}$ for $U \geq u(\hat{s})$, so $U^g \geq u(\hat{s})$ is optimally achieved by a single sentence. This also implies that when $U^g < u(\hat{s})$ is optimally achieved by randomizing between two sentences, these sentences both exceed $\hat{s}$.

# C Proof of uniqueness in Theorem 1, Part (iii)

Consider a truthful mechanism $(F, C, S)$ and suppose, first, that $S$ violates Condition (i) of Theorem 1 on a positive measure of signals. In this case, the step function $\tilde{S}(t, \hat{\imath})$ constructed in the first part of the proof is such that the difference $S(t, \hat{\imath}) - \tilde{S}(t, \hat{\imath})$ is strictly positive over a subset $T_1$ of $[0, \hat{t})$ that has positive Lebesgue measure and strictly negative over a subset $T_2$ of $(\hat{t}, 1)$ that has positive Lebesgue measure.[40] Since $u$ is strictly

[40] Indeed, the difference must be non-zero over a set of positive measure. Since $t \mapsto S(t, \hat{\imath}) - \tilde{S}(t, \hat{\imath})$ is single crossing from positive to negative, this implies that the existence of one of the two sets mentioned. Finally, since $S(t, \hat{\imath})$ and $\tilde{S}(t, \hat{\imath})$ give the same expected utility to an innocent defendant, and $u$ is decreasing it must be that the second set also exists: for example, if $S(t, \hat{\imath})$ strictly exceeds $\tilde{S}(t, \hat{\imath})$ over a set of positive measure, it must also be exceeded by it over a set of positive measure.

decreasing, this implies that the single-crossing function $\delta : t \mapsto \delta(t) = u(S(t, \hat{\imath})) - u(\tilde{S}(t, \hat{\imath}))$ is strictly negative over $T_1$ and strictly positive over $T_2$. Let $H(t) = \int_t^1 \delta(\tau) f_g^{\hat{\imath}}(\tau) d\tau$. By construction, we have $H(0) = H(1) = 0$, $H(t) \geq 0$ for all $t$, and $H(t) > 0$ for all $t$ in the interior of the convex hull of $T_1 \cup T_2$. [41] Let $\gamma(t) = f_i^{\hat{\imath}}(t)/f_g^{\hat{\imath}}(t)$. By strict MLRP, $\gamma$ is a strictly increasing function and thus almost everywhere differentiable. Therefore,

$$\int_{[0,1]} \delta(t) f_i^{\hat{\imath}}(t) dt = \int_{[0,1]} \delta(t) f_g^{\hat{\imath}}(t) \gamma(t) dt = \int_{[0,1]} -H'(t)\gamma(t) dt = \int_{[0,1]} H(t)\gamma'(t) dt < 0$$

where the strict inequality comes from the fact that $\gamma'$ is strictly negative except on a set of measure zero, while $H$ is strictly positive over a set of positive measure.

This shows that the innocent defendant strictly benefits from replacing $S(\cdot, \hat{\imath})$ with $\tilde{S}(\cdot, \hat{\imath})$, so welfare strictly increases.[42] Truthfulness is maintained because the original mechanism was truthful by assumption and, by construction, the guilty defendant is indifferent between $S(\cdot, \hat{\imath})$ and $\tilde{S}(\cdot, \hat{\imath})$ when misreporting.

Suppose now that $S$ violates Condition (iii) in Theorem 1, i.e., that $S(t, g)$ is non-constant. There are two cases to consider. If $u$ is strictly concave, then the certainty equivalent $s^{ce}$ is strictly higher than $s^a$: it is possible to increase a guilty defendant's expected punishment without violating incentive compatibility. If $s^{ce} \leq \hat{s}$, then since $W(s, g)$ is strictly increasing up to $\hat{s}$, setting $s^g = s^{ce}$ strictly increases the expected welfare conditional on facing a guilty defendant. If $s^{ce} > \hat{s}$, then setting $s^g = \hat{s}$ uniquely achieves the highest possible welfare conditional on facing a guilty defendant while preserving incentive compatibility, which constitutes a strict improvement. Suppose now that $W(s, g)$ is strictly concave. In this case, if $s^{ce} \leq \hat{s}$, setting $s^g = s^{ce}$ strictly improves welfare conditional on facing a guilty defendant, even if $u$ is only weakly concave, because $s^{ce}$ leads to a weakly higher expected punishment but eliminates the uncertainty about the punishment, which is strictly preferable according to the welfare function $W(s, g)$. If instead $s^{ce} > \hat{s}$, then setting $s^g = \hat{s}$ uniquely achieves the highest possible welfare conditional on facing a guilty defendant, and is a strict improvement because $S(t, \hat{g}) \neq \hat{s}$ (it is non-constant), while preserving truthfulness.

# D Proof of generic uniqueness in Theorem 1 Part (ii) and in Theorem 2

We will show that for "almost all" $u$ and $W(\cdot, g)$, in a sense to be made precise, the function $\hat{W}$ defined in the main text and its concavification $\bar{W}$ are such that whenever $\bar{W}$ is linear over some maximal interval $I$ (i.e., there is no interval strictly containing $I$ over $\bar{W}$ is linear), it coincides with $\hat{W}$ only at the endpoints of $I$. This property—which we call the "two-contact property"—implies that over the interior any such interval, the only way to achieve the optimal value $\bar{W}$ is to randomize over the endpoints of $I$, i.e., to use a two-point lottery. Over the remaining domain of $\hat{W}$, $\bar{W}$ and $\hat{W}$ coincide, and because $\bar{W}$ is locally strictly concave (since it is always concave and it is nonlinear over any subinterval of the remaining domain), the only way to achieve the optimum is a deterministic sentence.

The notion of "almost all" that we choose is the mathematical notion of " prevalence," which is used to formalize genericity for infinite-dimensional sets like the set of functions that we consider here.[43]

[41] The fact that $H(0) = 0$ is simply a restatement of(8). Nonnegativity of $H$ comes from the fact that the integrand of $H$, $\delta(t) f_g^i(t)$, is first negative and then positive and integrates up to 0, and the strict inequalities come from the fact that the integrand is strictly negative over $T_1$ and strictly positive over $T_2$.

[42] This is immediate if $W(\cdot, i) = u(s)$. The general case is explained in Appendix E. See Equation (17).

[43] The concept of prevalent sets was developed by Hunt et al. (1992) and coincides with the usual measure-theoretic

Given a topological vector space $\mathcal{W}$, a subset $\mathcal{G}$ is said to be *prevalent* if there exists a *finite* dimensional subspace $\mathcal{V}$ of $\mathcal{W}$ such that for all $w \in \mathcal{W}$, we have $w + v \in \mathcal{G}$ for all $v \in \mathcal{V}$ except for a set of $v$ that has Lebesgue measure zero in $\mathcal{V}$. Intuitively, it means that almost all translations of $w$ by elements in $\mathcal{V}$ belong to $\mathcal{G}$, where "almost all" is now understood in the usual sense of the Lebesgue measure over finite dimensional vector spaces.

In our case, the functions of interest are of the form $U \mapsto \hat{W}(U) = W(u^{-1}(U), g)$. Since $u^{-1}$ is continuous[44] and strictly monotonic, the transformation $u^{-1}$ amounts to a mere re-scaling (and direction change) of the function $s \mapsto W(s; g)$. Moreover, the domain of $[0, \bar{s}]$ can be without loss of generality taken to be $[0, 1]$.

This leads us to the following formulation of the genericity problem:

**Problem Statement**: Let $\mathcal{W}$ denote the vector space of all real-valued, continuous functions over $[0, 1]$ and $\mathcal{G}$ be the subset of $\mathcal{W}$ consisting of all functions $w$ whose concavification $\bar{w}$ over any maximal interval $I$ where it is linear coincides with $w$ only at the endpoints of $I$. Show that $\mathcal{G}$ is prevalent in $\mathcal{W}$.

To prove this result, the finite-dimensional subset $\mathcal{V}$ that we choose[45] is the set $\{af : a \in \mathbb{R}\}$, where $f(x) = x^2$. $\mathcal{V}$ is thus one dimensional.

Given a function $w \in \mathcal{W}$, let $w_a = w + af$, and let $A(w) = \{a \in \mathbb{R} : w_a \text{ violates the two-contact property}\}$. We wish to show that $A(w)$ has zero Lebesgue measure. For any fixed $a$, let $\{I_k^a\}_k$ denote the collection of maximal intervals of $[0, 1]$ over which the concavification $\bar{w}_a$ of $w_a$ is linear and coincides with $w_a$ at three or more points points of these intervals. Since these intervals are maximal, they are closed. Moreover, if $a$ is increased slightly, it is straightforward to see,[46] by strict convexity of $f$, that there are at most two points of contact over $I_k^a$ for all $a' > a$: all interior points $x$ of $I_k^a$ are such that $w_{a'}(x) < \bar{w}_{a'}(x)$.

If $w_a$ violates the two-contact property for some $a$, this implies that for any $a' > a$ the set of maximal intervals over which $w_{a'}$ violates the two-contact property consists of intervals $I_{k'}^{a'}$ that are either in the closure of the complement of $\cup_k \{I_k^a\}$, or consist of intervals that strictly contain some $I_k^a$. In particular, one may associate to each new interval a rational number $r_{a',k'}$ that belongs to $I_{k'}^{a'}$ but not to any other interval $I_k^a$.

Starting from any $a \in \mathbb{R}$, there must therefore exist for each $a' > a$ for which $w_{a'}$ violates the two-contact property an associated rational number $r_{a'}$ that belongs only to a maximal interval associated with $a'$. This implies that the set of $a' \geq a$ for which $w_{a'}$ violates the two-contact property is countable, because each such $a'$ is associated with a unique rational number. Since the statement is true for all $a \in \mathbb{R}$, we conclude that the set $A(w)$ is countable and, hence, has zero Lebesgue measure.

# E Welfare versus Utility Difference in Risk Attitude

While social preferences may be broadly aligned with those of the defendant when he is innocent, they need not be identical. We relax the assumption that $W(\cdot, i) = u(\cdot)$ and assume instead that there exists a strictly

notion of generic sets for finite-dimensional spaces. It has been in used in the mechanism design literature by Jehiel et al. (2006) and advocated by Anderson and Zame (2001) as a relevant measure of genericity for infinite-dimensional spaces in economics.

[44] It is well-known, and straightforward to check, that the inverse of a continuous, real-valued bijection over a compact domain is always continuous.

[45] Any strictly convex (or strictly concave) function would work equally well.

[46] Indeed, letting $\underline{x} < \bar{x}$ denote the endpoints of any such interval, we have for any $x = \lambda\underline{x} + (1-\lambda)\bar{x}$ in the interior of $[\underline{x}, \bar{x}]$, $f(x) < \lambda f(\underline{x}) + (1-\lambda) f(\bar{x})$. Since by assumption $\bar{w}_a$ is linear over the interval, we have $w_a(x) \leq \lambda w_a(\underline{x}) + (1-\lambda) w_a(\bar{x})$, which implies that $w_{a'}(x) = w_a(x) + (a'-a)f(x) < \lambda w_a(\underline{x}) + (1\lambda) w_a(\bar{x}) + (a'-a)(\lambda f(\underline{x}) + (1-\lambda) f(\bar{x})) = \lambda w_{a'}(\underline{x}) + (1-\lambda) w_{a'}(\bar{x})$. This shows that $w_{a'}(x) < \bar{w}_{a'}(x)$ for $x \in (\underline{x}, \bar{x})$.

increasing transformation $\phi : \mathbb{R}_- \to \mathbb{R}_-$ such that $W(s, i) = \phi(u(s))$. The weak convexity of $\phi$ means that the social preference over sentence lotteries when facing an innocent defendant exhibits less risk aversion than the defendant's own preference, i.e., that society need fully not internalize an innocent defendant's risk exposure to the judicial process.

Since this extension works in the same way for Theorems 1 and 2, we focus without loss of generality on Theorem 2.

**Proposition 1** *Suppose that $\phi$ is increasing and convex and that the assumptions of Theorem 2 are otherwise unchanged. Then, there exists a welfare-maximizing optimal mechanism that satisfies all the conclusions of Theorem 2.*

**Proof.** The construction is identical to the proof of Theorem 2. The welfare function $W(s, i)$ enters only the first step of the proof of Theorem 2, and it suffices to verify that expected welfare conditional on facing an innocent defendant is still increasing in this step. The first step replaces the sentence function $S(\cdot, \hat\imath)$ with a step function $\tilde S(\cdot, \hat\imath)$ that is equal to zero below $\bar t$ and equal to $\bar s$ above it, with $\bar t$ chosen to make a guilty defendant indifferent between $S(\cdot, \hat\imath)$ and $\tilde S(\cdot, \hat\imath)$.

For expositional simplicity, let us normalize the utility functions as follows: $u(0) = 0$, $u(\bar s) = -1$, $\phi(0) = 0$ and $\phi(-1) = -M$. This normalization is without loss of generality, as is easily checked. We must show the following inequality

$$\int_0^1 W(S(t, \hat\imath)) f_i^{\hat\imath}(t) dt \le \int_0^1 W(\tilde S(t, \hat\imath)) f_i^{\hat\imath}(t) dt = -M F_i^{\hat\imath}([\bar t, 1]),$$

where the equality follows from the normalization and the definition of the two-step sentence $\tilde S$. Since $W(s, i) = \phi(u(s))$, the previous inequality becomes

$$\int_0^1 \phi(u(S(t, \hat\imath))) f_i^{\hat\imath}(t) dt \le -M F_i^{\hat\imath}([\bar t, 1]), \tag{15}$$

It follows from the indifference equation (7) and the argument following it that

$$\int_0^1 u(\tilde S(t, \hat\imath)) f_i^{\hat\imath}(t) dt \ge \int_0^1 u(S(t, \hat\imath)) f_i^{\hat\imath}(t) dt \tag{16}$$

with a strict inequality if $S$ did not have the form of a step function. Using the above normalization for $u$ and definition of the cutoff $\bar t$ for $\tilde S$ then yields

$$-F_i^{\hat\imath}([\bar t, 1]) \ge \int_0^1 u(S(t, \hat\imath)) f_i^{\hat\imath}(t) dt \tag{17}$$

with a strict inequality if $S$ was not a step function.

Since $u(\cdot)$ takes values in $[-1, 0]$ we can view $-u(\tilde s)$ as a weight in a convex combination. Since also $u(0) = \phi(0) = 0$, $u(\bar s) = -1$, and $\phi(-1) = -M$, we have[47]

$$\begin{aligned}\phi(u(S(t, \hat\imath))) &= \phi\left[(-u(S(t, \hat\imath)))(-1) + (1 - (-u(S(t, \hat\imath))))(0)\right] \\ &\le (-u(S(t, \hat\imath)))\phi(-1) + (1 - (-u(S(t, \hat\imath))))\phi(0) \\ &= Mu(S(t, \hat\imath)).\end{aligned}$$

[47] The inequality is a direct application of the definition of $\phi$'s convexity if $t \mapsto S(t, \hat\imath)$ is deterministic. If $S(t, \hat\imath)$ is a lottery, the proof is also straightforward. For example, fixing some $t$, suppose that $S(t, \hat\imath)$ is a lottery with distribution $g$. Then $\phi(u(S(t, \hat\imath))) = \int_{[0,\bar s]} \phi(u(\tilde s)) g(\tilde s) d\tilde s$. For each $\tilde s$, the convexity of $\phi$ and together with $u(\tilde s) \in [-1, 0]$, $u(\bar s) = -1$, $u(0) = 0$, $\phi(0) = 0$, and $\phi(-1) = -M$, imply $\phi(u(\tilde s)) = \phi((-u(\tilde s)(-1) + (1 - (-u(\tilde s)))(0))) \le (-u(\tilde s))\phi(-1) + (1 - (-u(\tilde s)))\phi(0) = Mu(\tilde s)$. Integrating over $\tilde s$ then yields $\phi(u(S(t, \hat\imath))) \le M \int_{[0,\bar s]} u(\tilde s) g(\tilde s) d\tilde s = Mu(S(t, \hat\imath))$.

Integrating this equation for $t = 0$ to 1 with respect to the density $f_i^{\hat{\imath}}$ yields

$$\int_0^1 \phi(u(S(t,\hat{\imath})))f_i^{\hat{\imath}}(t)dt \leq M \int_0^1 u(S(t,\hat{\imath}))f_i^{\hat{\imath}}(t)dt.$$

Combining this with (17) then yields (15). ■

## F Trials with Guilty Defendants

By the last part of each of the theorems, guilty defendants in the optimal mechanisms are indifferent between taking a plea and going to trial. Thus, if a small fraction of guilty defendants goes to trial, the resulting welfare is close to optimal. As we now demonstrate, this allows for both a large fraction of convicted defendants to be guilty, and for jurors to use Bayesian updating to determine a defendant's guilt in a way that approximates the optimal mechanisms.

Suppose that under the optimal sentencing scheme a fraction $\alpha$ of guilty defendants reject the plea and go to trial. The jury's belief, upon seeing a defendant going to trial and observing signal $t$ regarding the defendant's guilt, is a combination of both pieces of information (rejecting the plea and generating signal $t$). With Bayesian updating, the posterior probability of guilt corresponding to some signal $t$ can be computed in two steps. First, given a prior $\lambda$ and the fact that the defendant rejected the plea and went to trial, the probability at the outset of the trial that the defendant is guilty is

$$\hat{\lambda} = \frac{\lambda\alpha}{\lambda\alpha + (1-\lambda)}. \tag{18}$$

Next, at the end of the trial, given signal $t$ the probability that the defendant is guilty is

$$\hat{p}(t) = \frac{\hat{\lambda}f_g(t)}{\hat{\lambda}f_g(t) + (1-\hat{\lambda})f_i(t)} = \frac{\hat{\lambda}r(t)}{\hat{\lambda}r(t) + (1-\hat{\lambda})},$$

where $r(t) = f_g(t)/f_i(t)$ is the likelihood ratio associated with signal $t$. Replacing $\hat{\lambda}$ by (18), we have

$$\hat{p}(t) = \frac{\lambda\alpha r(t)}{\lambda\alpha r(t) + (1-\lambda)}.$$

Thus, for any fraction $\alpha > 0$ and conviction threshold $\hat{t}$ there corresponds a posterior belief $\hat{p}(\hat{t})$ of guilt. To get a rough sense of this threshold, suppose that the likelihood ratio at the optimal threshold $\bar{t}$ is equal to ten, i.e., the evidence necessary to convict a defendant must be ten times more likely to have come from a guilty defendant than from an innocent one. This is consistent with the doctrine of "beyond a reasonable doubt" (BARD) used in criminal cases.[48] Also suppose that, consistent with criminal data in the United States, 90% of defendants are in fact guilty.[49] These assumptions correspond to $\lambda = 0.9$ and $r(\bar{t}) = 10$. The associated posterior probability that the defendant is guilty is

$$\hat{p} = \frac{9\alpha}{9\alpha + 0.1} = 1 - \frac{.1}{9\alpha + 0.1}.$$

For $\alpha = 0.1$, for instance, this implies that the posterior probability of guilt of a defendant who is barely convicted under the optimal scheme is 0.9, or 90%. Thus, even if the BARD doctrine is applied to posterior beliefs that take into account the decision of the defendant to reject the plea, instead of being based purely on the evidence

[48] William Blackstone, Commentaries on the Laws of England, Volume 2, edited by William Carey-Jones, Bancroft–Whitney, San Francisco, 1916 (Books 3 & 4) Book 4, *358, page 2596.

[49] More than 90% of criminal cases in the United States lead to a conviction. More than 90% plead guilty, and of those going to trial, more than 90% are found guilty.

presented at trial, the mechanism proposed here leads to a certainty threshold of 90% regarding the guilt of convicted defendants when 10% of guilty defendants reject the plea.

Thus, under realistic assumptions with regard to the evidence conviction threshold $\bar{t}$ and the prior $\lambda$ of guilt, our modified mechanism remains consistent with BARD and the observation that most defendants are guilty. With a fraction $\alpha$ of guilty defendants going to trial, we incur a welfare loss relative to the optimal mechanism since these guilty defendants are sometimes acquitted and sometimes punished too severely. But this loss concerns only a small fraction of guilty defendants. In addition, once some guilty defendants go to trial, evidence is used to determine the defendant's guilt, in addition to its role in incentivizing most guilty defendants to accept the plea bargain.

## References

Anderson, R., Zame, W. (2001) "Genericity with Infinitely Many Parameters," *Advances in Theoretical Economics*, Vol. 1, pp. 1–62.

Aumann, R., Maschler, M., and Stearns, R. (1995) *Repeated Games with Incomplete Information*, MIT Press.

Baker, S., Mezzetti, C. (2001) "Prosecutorial Resources, Plea Bargaining, and the Decision to Go to Trial," *Journal of Law, Economics, and Organization,* Vol. 17, pp. 149–167.

Becker, G. (1968) "Crime and Punishment: An Economic Approach," *Journal of Political Economy,* Vol. 76, pp. 169–217.

Daughety, A., Reinganum, J. (2016a) "Informal Sanctions on Prosecutors and Defendants and the Disposition of Criminal Cases," *Journal of Law, Economics, and Organization*, Vol. 32, pp. 359–394.

Daughety, A., Reinganum, J. (2016b) "Selecting Among Acquitted Defendants: Procedural Choice vs. Selective Compensation," *Journal of Institutional Theoretical Economics*, Vol. 172, pp. 113–133.

Deffains, B. and Demougin, D. (2008) "The Inquisitorial and the Adversarial Procedure in a Criminal Court Setting," *Journal of Institutional and Theoretical Economics*, Vol. 164, pp. 31–43.

Demougin, D. and Fluet, C. (2016) "Preponderance of Evidence," *European Economic Review*, Vol 50, pp. 963–976.

Dewatripont, M. and Tirole, J. (1999) "Advocates," *Journal of Political Economy*, Vol 107, pp. 1–39.

Grossman, G., and Katz, M. (1983) "Plea Bargaining and Social Welfare," *American Economic Review*, Vol. 73, pp. 749–757.

Hunt, B., Sauer, T., and J. Yorke (1992) "Prevalence: A Translation-Invariant "Almost Every" on Infinite-Dimensional Spaces," *Bulletin of the American Mathematical Society*, Vol. 27, pp. 217–238.

Innes, R. (1999) "Remediation and Self-Reporting in Optimal Law Enforcement," *Journal of Public Economics*, Vol. 72, pp. 379–393.

Innes, R. (2000) "Self-Reporting in Optimal Law Enforcement When Violators Have Heterogeneous Probabilities of Apprehension," *Journal of Legal Studies*, Vol. 29, pp. 287-300.

Jehiel, P., Meyer-ter-Vehn, M., Moldovanu, B., and W. Zame (2006) "The Limits of Ex Post Implementation," *Econometrica*, Vol. 74, pp. 585–610.

Jordan, A. (2020) "What Can Plea Bargaining Teach Us About Racial Bias in Criminal Justice?" *Working Paper.*

Kaplow, L. (2011) "On the Optimal Burden of Proof," *Journal of Political Economy*, Vol. 119, pp. 1104–1140.

Kaplow, L. (2017) "Optimal Multistage Adjudication," *Journal of Law, Economics, and Organizations,* Vol. 33, pp. 613–652.

Kaplow, L and Shavell, S. (1994) "Optimal Law Enforcement with Self-Reporting of Behavior," *Journal of Political Economy,* Vol. 102, pp. 583–606.

Karlin, S. (1968) *Total Positivity, Volume 1*, Stanford University Press.

Karlin, S., and Rubin, H. (1956) "The Theory of Decision Procedures for Distributions with Monotone Likelihood Ratio." *The Annals of Mathematical Statistics*, Vol. 27, pp. 272–299.

Klement, A. and Neeman, Z. (2005) "Against Compromise: A Mechanism Design Approach," *Journal of Law, Economics, and Organization*, Vol. 21, pp. 285–314.

Kremer, I., Mansour, Y., and Perry, M. (2014) "Implementing the "Wisdom of the Crowd," *Journal of Political Economy*, Vol. 122, pp. 988–1012.

Lando, H. (2005) "The Size of the Sanction should Depend on the Weight of the Evidence," *Review of Law and Economics*, Vol. 1, pp. 277–292.

Lee, S. (2014) "Plea Bargaining: On the Selection of Jury Trials," *Economic Theory*, Vol. 57, pp. 59–88.

Pei, H and Strulovici, B. (2021) "Crime Aggregation, Deterrence, and Witness Credibility," *Working Paper.*

Polinsky M and Rubinfeld, D. (1988) "The Deterrent Effects of Settlements and Trials," *International Review of Law and Economics*, Vol. 8, pp. 109–116.

Reinganum, J. (1993) "The Law Enforcement Process and Criminal Choice," *International Review of Law and Economics*, Vol. 13, pp. 115–134.

Sauer, K. (1995) "Informed Conviction: Instructing the Jury About Mandatory Sentencing Consequences," *Columbia Law Review*, Vol. 95, pp. 1232–1272.

Shin, H. S. (1998) "Adversarial and Inquisitorial Procedures in Arbitration," *The RAND Journal of Economics*, Vol. 29, pp. 378–405.

Siegel, R., and Strulovici, B. (2020) "The Economic Case for Probability-Based Sentencing," *Working Paper.*

Silva, F. (2019) "If We Confess Our Sins," *International Economic Review*, Vol. 60, pp. 1389–1412.

Spear, S., Srivastava, S. (1987) "On Repeated Moral Hazard with Discounting," *Review of Economic Studies*, Vol. 54, pp. 599–617.

Spier, K.E. (1994) "Pretrial Bargaining and the Design of Fee-Shifting Rules," *The RAND Journal of Economics*, pp. 197–214.

Stigler, G. (1970) "The Optimum Enforcement of Laws," *Journal of Political Economy,* Vol. 78, pp. 526–536.