Judicial Mechanism Design

Ron Siegel and Bruno Strulovici*

October 6, 2020

Abstract

This paper introduces a mechanism-design framework to study criminal justice systems. We identify properties of the generically unique optimal mechanisms for two notions of welfare distinguished by their treatment of deterrence. These properties shed new light on features of the criminal justice system in the United States, from the prevalence of binary verdicts with extreme sentences in conjunction with plea bargains to the use of strict jury instructions and an adversarial system, all of which emerge as the result of informational, commitment, and incentive arguments.

1 Introduction

Criminal justice systems have features that may seem puzzling from an economic perspective. It is unclear, for example, why criminal trials usually have only two verdicts, "guilty" and "not guilty," with extreme (large or null) sentences, as opposed to a more gradual system that reflects the strength of evidence. False convictions (and presumably false acquittals) are not uncommon, reflecting the fact that evidence regarding defendants' guilt is generally imperfect.[1] A binary verdict is not well suited to reflect this imperfection, and imperfect evidence suggests that some punishment following an acquittal is

*We are grateful for numerous questions, discussions, and comments from various 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.

[1]Gross et al. (2014) finds that, out of 7,482 death row convictions from 1973 to 2004 in the United States, *at least* 4.1% of death-row defendants have been wrongfully convicted. Given the high burden of proof required for convictions, acquittals of guilty defendants are likely even more frequent. Liebman, et al. (2000) find that, of all capital sentences given by lower court in the United States between 1973 and 1995, 68% of capital sentences were found by higher courts to contain serious, prejudicial errors, and 7% of the defendants whose initial death sentence was overturned were later found to be innocent of the capital crime.

sometimes optimal.[2] Moreover, sentencing guidelines in the United States do not mention the strength of evidence as a relevant factor in determining the punishment for a convicted defendant.[3]

Binary verdicts, no punishment following an acquittal, and other features of criminal justice systems, such as plea bargaining, are exogenously imposed in much of the law and economics literature. Early work, pioneered by Becker (1968) and Stigler (1970), used equilibrium analysis to study law enforcement and criminal justice. More recent work, including Grossman and Katz (1983), Baker and Mezzetti (2001), Kaplow (2017), and Daughety and Reinganum (2016a,b), takes a game theoretic approach. This literature, which has led to many important insights, does not provide a foundation for the aforementioned features of many criminal justice systems: these features are assumed as part of the models.

This paper takes a different approach. Instead of studying a specific judicial system with an exogenous structure, we focus on the goals and constraints common to all justice systems: (i) The social planner aims to deter potential criminals, suitably punish guilty defendants, and avoid punishing innocent one, and (ii) The defendant privately knows whether he is guilty. The planner designs a system to generate evidence regarding the defendant's guilt and determine an appropriate sentence.[4]

Judicial Mechanism Design: To implement this approach, a major challenge is to define a tractable set of mechanisms over which to optimize welfare. In reality, a judicial process involves numerous stages and agents (prosecutors, witnesses, forensic experts, jurors) whose incentives depend on the specific rules of the judicial process. To identify general insights, we proceed in two steps: *Reduction* and *Invariance*.

In the Reduction Step, we model judicial processes as extensive-form games of incomplete information and show how to extract, from any judicial process, a *judicial mechanism* focused on the defendant.[5] A judicial mechanism is a truthful, direct-revelation mechanism in which the defendant reports his type (guilty or innocent), a signal is generated that depends on the defendant's true and reported types, and the defendant receives a sentence according to a *sentencing scheme* that maps the

[2]For example, Scotland uses a "Not Proven" verdict when the incriminating evidence is significant but does not rise to the standard of proof for a guilty verdict. See, Bray (2005), Daughety and Reinganum (2016a,b).

[3]The list of mitigating and aggravating factors recognized by the U.S. code (18 U.S.C. §3553) does not include strength of evidence.

[4]Mechanism design has been applied to study tort law: Spier (1994), Klement and Neeman (2005), and Demougin and Fluet (2006) analyze settlement and fee-shifting rules between plaintiffs and defendants.

[5]Our approach is similar in spirit to Gershkov and Szentes (2009), who perform a mechanism design analysis of extensive-form voting mechanisms with costly information acquisition. See also Gerardi and Yariv (2008), Shi (2012), and Kremer et al. (2014). Dynamic mechanism design models that relax the standard commitment assumption include Skreta (2006) and Doval and Skreta (2020a,b).

signal and the defendant's reported type into a lottery over sentences. To formalize this reduction, we show that, while the sentence given to a defendant in a judicial process can generally depend on many aspects of the process, the sentence can without loss of generality (for incentive and welfare purposes) be expressed as a function of the information generated by the judicial process (Proposition 1). The Reduction Step enables us to conduct the mechanism design exercise in the space of judicial mechanisms. It is not clear, however, which judicial mechanisms are available to the planner. This is the role of the Invariance Step.

In the Invariance Step, we introduce our main assumption: *for any judicial mechanism available to the planner, any change of the sentencing scheme that preserves the defendant's truth-telling incentives yields a mechanism that is also available to the planner.* According to this assumption, the availability of a judicial mechanism is invariant to the sentencing scheme used by the mechanism, as long as the sentencing scheme is incentive compatible for the defendant. Realistic incarnations of this assumption include adversarial justice systems and jury instructions, as discussed in Section 2.

The assumption may be viewed as giving the designer the ability to insulate the generation and processing of information about the defendant from the sentence that the defendant faces as a function of this information. It allows us to focus on the informational challenge every judicial system must contend with, namely the defendant's private information regarding his guilt, while abstracting to a large extent from the other agents' incentives. 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.*

Results and Comparison to Existing Legal Systems: We find that welfare-maximizing mechanisms display striking similarities to existing judicial systems and can be implemented by the following procedure. The defendant is first offered a "plea bargain." If he accepts it, the mechanism ends and the defendant gets a positive sentence. If he rejects the offer, he must go through a "trial" that produces evidence and ends with one of two verdicts: the defendant is found "guilty" if the likelihood of guilt associated with the evidence exceeds some threshold and "not guilty" otherwise. The guilty verdict carries a sentence that is more severe than the one associated with a plea bargain; a "not guilty" verdict carries no punishment. Finally, the evidence produced during the trial is used to determine the *verdict* but not the *sentence* associated with the verdict.

All these features, which arise endogenously as part of the optimal mechanism, are interrelated. For example, if plea bargains were excluded from the class of possible mechanisms, it would be optimal to use more than two verdicts, and the optimal sentence would be gradually increasing in the strength of incriminating evidence (see Lando (2005), Fisher (2011), and Siegel and Strulovici (2020)).

Welfare-maximizing mechanisms have another feature, which is ubiquitous in contract theory and

mechanism design but seems problematic and unrealistic in a legal setting: the punishment of agents who are known to have taken the socially desirable action in equilibrium but were unlucky enough to draw a bad signal. In contract theory, for instance, agents who are known to produce high effort in equilibrium, but face stochastic output, receive a low payoff when they produce low output. This punishment is necessary to induce agents to choose high effort in the first place.

The logic is similar with optimal judicial mechanisms: a defendant who rejects a plea bargain and goes to trial is surely innocent (as in Grossman and Katz's (1983) seminal analysis of plea bargains). However, this defendant is punished if the evidence revealed against him during trial is sufficiently incriminating. In our simple model, the role of evidence is used to induce guilty defendants to take the plea, not to sort defendants who rejected it.

In reality, defendants who go to trial are guilty with positive probability and the evidence also serves to sort defendants at the trial stage. A realistic equilibrium should have the following features: (i) some guilty defendants go to trial, (ii) the evidence presented at trial serves *two roles*: induce guilty defendants to take a plea and sort defendants at the trial stage.

This realistic equilibrium turns out to be compatible with the welfare-maximizing mechanisms in this paper: in these mechanisms, guilty defendants are by design indifferent between taking the plea bargain and going to trial. Moreover, if a small fraction of guilty defendants goes to trial, the equilibrium is nearly optimal and evidence regains its role in determining actual guilt, in addition to its incentivizing role. We study this equilibrium in Section 5.2.

Finally, we show that the optimal plea bargain may sometimes involve a random sentence. When defendants are risk averse, uncertainty offers a way of punishing guilty defendants and increase deterrence. This feature is consistent with plea sentences in practice, because plea bargain agreements between a defendant and a prosecutor can be altered by the judge overseeing the case.[6] We show, however, that random plea sentences are optimal only if achieving deterrence is more important—in a sense that we make precise—than avoiding Type I/II errors at the trial stage.

2 Judicial Mechanisms

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

[6]Recent examples include the case of Jared Fogle, a former Subway spokesman who accepted a plea bargain with 5 years in prison but received a much larger sentence. See "Jared Fogle, Former Subway Pitchman, Gets 15-Year Prison Term," *New York Times*, November 19, 2015.

2.1 Interim Perspective

Consider a defendant who has just been arrested for a crime. 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 that produces, at some cost c, a signal t about the defendant's guilt and assigns a sentence $s \in [0, \bar{s}]$ to the defendant.

The signal's probability distribution depends on the defendant's type and actions during the judicial process (e.g., admitting guilt can interrupt evidence collection). Likewise, the cost of the judicial process generally depends on the defendant's type and actions.

2.1.1 Reduction to a Truthful Direct-Revelation Mechanism

A judicial process involves several players and stages. Formally, we view it as an extensive-form game of incomplete information with multiple players and a finite horizon. From this process we extract a direct-revelation mechanism focused on the defendant, in which the strategy of the defendant is to send a report $\hat{\theta} \in \{\hat{g}, \hat{\imath}\}$ about his type θ, and in which truth-telling is optimal for the defendant.

Formally, a *(direct-revelation) judicial mechanism* is a tuple (F, C, S), where $F = \left(F_i^{\hat{\imath}}, F_g^{\hat{\imath}}, F_i^{\hat{g}}, F_g^{\hat{g}}\right)$ contains a signal distribution for each pair $(\theta, \hat{\theta})$ of actual and reported types of the defendant, $C = \left(C_i^{\hat{\imath}}, C_g^{\hat{\imath}}, C_i^{\hat{g}}, C_g^{\hat{g}}\right)$ assigns a cost for each of these pairs, and $S : (t, \hat{\theta}) \mapsto S(t, \hat{\theta})$ is a *sentencing scheme* that maps the signal t and 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.[7] The lower bound 0 on the sentence means that the defendant cannot be rewarded or compensated by the mechanism and is consistent with existing practice.[8]

Given a judicial process in extensive form and an equilibrium strategy profile, we construct the direct-revelation mechanism in four steps:

- **Step 1:** The defendant reports his type $\hat{\theta}$.
- **Step 2:** A signal t is generated about the defendant's type θ. This signal has the same distribution

[7]The upper bound $\bar{s}$ may be viewed as a technical or ethical constraint on punishment. 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. This restriction and its relation to Crémer and McLean's (1988) is discussed in Appendix F.

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

as the signal generated in the judicial process if the defendant follows the equilibrium strategy of type $\hat{\theta}$ in the judicial process.

- **Step 3:** A sentence s is drawn from a distribution that depends on the report $\hat{\theta}$ and on the signal t. The sentence has the same distribution as the sentence generated by the judicial process when the defendant follows the strategy of a type-$\hat{\theta}$ defendant *and* the likelihood ratio of guilt at the end of the judicial process is equal to the likelihood ratio of guilt associated with signal t.

- **Step 4:** The mechanism generates a cost $C_{\theta}^{\hat{\theta}}$ equal to the expected cost of the judicial process when the defendant has type θ and follows the equilibrium strategy of type $\hat{\theta}$.

Section 4 formalizes this reduction from any judicial process to a single-agent judicial mechanism. In particular, it shows that taking the mechanism's signal to be one dimensional, a common assumption in the law and economics literature, is without loss of generality, *even though evidence generated by the judicial process may be multi-dimensional.* Section 4 also shows why it is always possible to represent the sentence given in a judicial process through the signal decomposition used in Step 3.

By a logic similar to that of the Revelation Principle (Myerson 1979), it is optimal for the defendant to report his type *truthfully* in the direct-revelation mechanism.[9]

2.1.2 Invariance Assumption

Our mechanism design objective is to select, within the set of judicial mechanisms, the optimal sentencing scheme (i.e., map from signal and report to sentence lotteries). In the mechanism design literature, the ability to choose any scheme is often framed as a commitment assumption: the designer is committed to a map from messages to outcome, no matter how undesirable the outcomes are ex post. Alternatively, this ability can be interpreted as "richness" condition over the set of game forms available to the designers: for any scheme that the designer might consider, there is some game available to the designer that implements this sentencing scheme.

In this spirit, we formulate the designer's ability to choose a sentencing scheme as an invariance assumption: given (i) any judicial mechanism available to the designer and (ii) any sentencing scheme, the designer can use this sentencing scheme in the initial judicial mechanism without affecting the signals generated by the judicial mechanism.[10] Intuitively, we wish to capture the idea that the designer has

[9] If it were profitable for the defendant to misreport his type, the strategy in the original judicial process corresponding to this other type would be a profitable deviation in the original judicial process.

[10] More precisely, we require this condition as long as the sentencing scheme is truthful given the signals generated by the judicial mechanism.

access to a large class of judicial processes, sufficiently rich that the judicial mechanisms satisfy this assumption.[11]

Formally, the assumption states that given an available truthful mechanism (F, C, S), any mechanism $(F, C, \tilde{S})$ that results from changing the sentencing scheme to $\tilde{S}$ is also available, provided that it is truthful. This assumption allows us to ignore actors of the judicial system and focus on our main concern, which is the defendant's private information regarding his guilt.[12]

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*, i.e., measurable functions from $T = [0, 1]$ (signals) and $\{\hat{g}, \hat{\imath}\}$ (reported type) to $\Delta([0, \bar{s}])$ (lotteries over sentences). 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 satisfies the following conditions:

- $u(s)$ is strictly decreasing and continuous in s.
- $u(0) = 0$.

Let $\mathcal{M}$ denote the set of truthful judicial mechanisms available to the designer. We can now state our key invariance assumption:

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

Interpretation: Assumption 1 may be interpreted in several ways: (a) The planner can insulate agents' incentives to seek, reveal, or process information from the sentencing scheme used to punish the defendant. This insulation finds some incarnation in the form of *adversarial justice systems* and

[11] Under this interpretation, the judicial processes underlying each judicial mechanism as not controlled by the designer: each judicial process features an equilibrium involving various agents who are not directly controlled by the designer. But there are sufficiently many judicial processes to choose from that it looks as if the designer could directly choose the sentencing scheme.

[12] It would be interesting to combine this approach with the incentives of other agents, such as prosecutors and jurors. This paper's first step should prove useful to consider these more complex questions.

jury instructions. First, an adversarial justice system splits agents into two groups: one that produces incriminating evidence and one that produces exculpatory evidence. This split and agents' roles and incentives do not depend on the specific sentence faced by the defendant. [13] Second, jury instructions in the United States aim to insulate jurors from the sentence that the defendant may face if convicted so as to separate as much as possible jurors' fact-finding task from the judge's sentencing task. [14] (b) The planner can choose agents whose interests are more aligned with his (as in the case of jury selection), or to provide detailed and forceful instructions to agents that reduce the scope for deviations. (c) Finally, to the extent that incentive constraints of agents other than the defendant arise in reality, *the set of mechanisms that we consider here is likely a superset of those available in practice*. When welfare-maximizing mechanisms in the class of mechanisms that satisfy Assumption 1 share some features with existing criminal justice systems, these features may suggest that any additional constraints pertaining to these features are not binding in practice, and vice versa.

Versions of Assumption 1 appear in numerous works in law and economics.[15] Section 5.1 interprets Assumption 1 in light of existing features of the US criminal justice system.

Regularity Conditions for Signal Distributions: Section 4 shows that, given a report $\hat{\theta}$, the signal t can be assumed without loss generality to equal the likelihood ratio of the defendant's type: given report $\hat{\theta}$, if $t' > t$, then a guilty defendant is more likely to generate signal t' than t, compared to an innocent defendant. By using an appropriate increasing transformation of the likelihood ratio, we can further assume without loss that, for each report $\hat{\theta}$, the signal t takes values in $[0,1]$ and that the distribution $F_g^{\hat{\theta}}$ dominates the distribution $F_i^{\hat{\theta}}$ according to the monotone likelihood ratio property (MLRP).

To simplify the analysis, we impose the following conditions:

(i) For each $(\theta, \hat{\theta})$, the signal distribution $F_\theta^{\hat{\theta}}$ has a continuous density $f_\theta^{\hat{\theta}}$ and has full support over $T = [0,1]$.[16]

[13] This benefit was 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).

[14] 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.

[15] 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.

[16] The domain of t is just a normalization: in Section 4, we show that the signal summarizing the evidence against the defendant can be taken without loss of generality to be the likelihood ratio $\ell \in (0, \infty)$. From this we can always redefine

(ii) For each $\hat{\theta} \in \{\hat{g}, \hat{\imath}\}$, the ratio $f_g^{\hat{\theta}}(t)/f_i^{\hat{\theta}}(t)$ *strictly* increases in t (Strict MLRP).

The assumption that signals have continuous, strictly positive densities implies that the set of likelihood ratios $\{f_g^{\hat{\imath}}(t)/f_i^{\hat{\imath}}(t)\}_{t\in T}$ is a subinterval of $(0, +\infty)$. This assumption is used in the construction of welfare-improving schemes that keep the defendant's expected utility unchanged.[17]

2.1.3 Welfare

From an interim perspective, society wishes to punish guilty defendants and avoid punishing innocent ones, and takes into account the cost of producing evidence. Maximizing social welfare after the crime has been committed is similar to minimizing Type I and Type II errors (convicting an innocent defendant and acquitting a guilty ones) when facing a binary-verdict decision, but in our setting the designer can choose a sentence on a continuum rather than a binary verdict.

We denote by $W(s, \theta)$ the social welfare of imposing a sentence s on a defendant of type θ. 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:*[18]

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

The second assumption implies that (i) $W(s, i)$ is strictly decreasing in s, which means that it is socially harmful to punish innocent defendants, and (ii) the social planner is weakly less averse than the defendant with regard to uncertainty over sentences. A particular case is $W(s, i) = u(s)$, which is the assumption made by Grossman and Katz (1983) in their analysis of plea bargains.

Given a probability λ that the defendant is guilty, the expected welfare of giving a sentence s to the defendant is

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

$t = \ell/1 + \ell$, or if ℓ's support given the defendant's report $\hat{\theta}$ is a subinterval $(\underline{\ell}(\hat{\theta}), \bar{\ell}(\hat{\theta}))$ of $\mathbb{R}$, we can normalize the signal by setting $t = \ell - \underline{\ell}(\hat{\theta})/(\bar{\ell}(\hat{\theta}) - \underline{\ell}(\hat{\theta}))$ to guarantee that t has full support in $(0, 1)$ for each report $\hat{\theta}$.

[17] 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 Theorem 1 and 3 go through, but the optimal scheme will generically involve randomization over two extreme sentences when the signal observed has the corresponding likelihood ratio.

[18] The non-positivity of welfare functions guarantees that interim welfare is never so high as to offset the harm caused by the crime in the first place, and could be relaxed accordingly.

To allow for stochastic sentencing, let $W(\tilde{s}, \theta)$ denote the expected welfare of imposing a (possibly) random sentence $\tilde{s}$ on a defendant of type θ.[19] Also let $C_\theta^{\hat{\theta}} \geq 0$ denote the expected cost of the judicial process when the defendant has type θ and follows the strategy of type $\hat{\theta}$. This cost captures information acquisition costs as well as procedural costs. Given a judicial mechanism (F, C, S) in which the defendant reports his type truthfully, 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 here 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 expositional convenience, we treat π_g as exogenous.[20] We assume that at most one individual is prosecuted for the crime.[21] The planner chooses a judicial mechanism before individuals decide whether to commit the crime.

Given a judicial mechanism (F, C, S), an individual commits the crime if

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

The mechanism faced by the defendant could a priori depend on the evidence gathered between the time of the crime and the time of the defendant's arrest. If this evidence is stochastic from the viewpoint of someone committing a crime, the second term of (4) should aggregate a distribution of mechanisms, one for each belief assigned to the defendant's guilt at the time of arrest. Our results hold in this more general environment because the welfare-improving mechanisms constructed in Section 3 keep the expected utility of a guilty defendant in the mechanism unchanged and can be performed for each mechanism. These changes would improve welfare without affecting deterrence.[22] For expositional simplicity, we focus here on a single belief at the time of arrest, which means that deterrence is evaluated

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

[20] 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 we would simply apply for the optimal level of law enforcement.

[21] 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.

[22] As explained in Section 4, our results also hold if the welfare functions $W(\cdot, i)$ and $W(\cdot, g)$ and the utility function $u(\cdot)$ depend on the evidence gathered at the time of arrest and if the mechanisms and the welfare and utility functions are allowed to depend on some observable characteristics of the defendant, such as age or socioeconomic status. We simply apply the sentence modifications used in our proofs separately for each observable characteristic of the defendant, without affecting deterrence, in a way that improves interim welfare.

when anticipating a particular mechanism with certainty following an arrest. This would also be the case if the planner were constrained to use the same mechanism regardless of the information acquired about the defendant at the time of arrest.

The individual benefit b from committing the crime varies in the population, and is distributed according to some probability measure G_b. Letting $H(F,S)$ denote the fraction of individuals who commit the crime, we have

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

In a large society, the probability that a *specific* individual be arrested for a crime the he did not commit is small. Furthermore, such erroneous arrests affect both individuals who committed a (different) crime or those who are innocent. For simplicity, these events are omitted from the incentive equation (4).[23]

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

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

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 $\lambda=\pi_g/\left(\pi_g+\pi_i\right)$. Using these observations, we recover (3) from (6).

3 Optimal Judicial Mechanisms

This section derives key properties of interim-optimal judicial mechanisms and of ex-ante optimal mechanisms. These properties are stated in three closely related but distinct theorems that are based on the same core arguments.,

3.1 Interim Welfare

A judicial mechanism (F,C,S) is *interim optimal* if, given the prior probability λ that the defendant is guilty, the mechanism maximizes interim welfare (3) among all the mechanisms in $\mathcal{M}$. Studying

[23] Our results hold even if the probability of being wrongfully convicted had 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 any given innocent individual is prosecuted 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.

interim-optimal mechanisms allows us to disentangle deterrence from other welfare considerations and makes the arguments in the proof easier to follow.

For our first result, we assume that the welfare function conditional on facing a guilty defendant is single peaked. Single-peakedness of the welfare function for a guilty defendant is consistent with US sentencing guidelines, which state that "The court shall impose a sentence sufficient, but not greater than necessary, to...reflect the seriousness of the offense... and to provide just punishment for the offense."[24] We also assume that the defendant's utility and the welfare objective function exhibit risk aversion.

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

Theorem 1 *Under Assumption 3, any interim-optimal mechanism has the following properties:*[25]

(i) The innocent defendant's sentence is a step function of the signal t, which jumps from 0 to $\bar{s}$ at some cutoff $\bar{t}$.

(ii) The guilty defendant's sentence is constant in t.

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

Any mechanism without these properties is welfare-dominated for all *non-degenerate priors λ by a* single mechanism *with these properties.*

All proofs for this section are in the Appendix.

Theorem 1 shows that an optimal mechanism resembles a system in which plea bargains are available and trials end in one of two verdicts. If the defendant pleads guilty, he receives a fixed sentence and forgoes a trial. 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). We emphasize that a binary verdict following a trial and a null sentence following an acquittal were *not* assumed features of the mechanism, but rather emerge as part of the optimal mechanism.[26]

[24]See 18 U.S.C §3553. These guidelines also state that another goal of sentencing is "to protect the public from further crimes of the defendant." This incapacitation motive presumably increases at a rate that decreases in the sentence, whereas the disutility a prisoner experiences increases with his sentence, which together may also give rise to single-peaked social welfare.

[25]All statements in this section may be violated on a set signals that has zero probability. We ignore this for simplicity.

[26]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).

Non-Bayesian Statement: The last statement in Theorem 1 shows that Theorem 1 is non-Bayesian in nature: 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 on each defendant type θ. In the language of statistical decision theory, this shows that the class of mechanisms described by Theorem 1 forms a *complete class* (Karlin and Rubin (1956)). 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.

Singular Role of Evidence: If the defendant pleads guilty, the signal is not used by the mechanism to determine his sentence, even if the signal t conditional on $\hat{\theta} = \hat{g}$ is informative about the defendant's guilt. This highlights the role of evidence in the optimal mechanism, which is only to induce the defendant to reveal his type through screening. On the equilibrium path, evidence adds no further information because the defendant has already revealed his type, but evidence plays a key role in dissuading deviations from the equilibrium path.

Plea Bargaining as a Screening Device: The screening value of plea bargains has already been noted and emphasized by Grossman and Katz (1983). That paper does not show the optimality of plea bargains among other mechanisms: it takes as exogenously given the structure of a two-verdict system with a plea bargain, whereas we show that such a system is in fact globally optimal from an interim perspective and under Assumption 3. Theorems 2 and Theorem 3 below show that fixed plea sentences are in fact not generally optimal when Assumption 3 is relaxed or when deterrence is taken into account.[27]

Extreme Sentences: In his seminal analysis of law enforcement, Becker (1968) noted the optimality of extreme sentences. In Becker's framework, extreme sentences increase the expected punishment from committing crime and, hence, improve deterrence. Becker's analysis ignores the possibility of Type I errors: in that paper, extreme sentences work because they are only ever given to criminals. By contrast, we find that extreme sentences are optimal when Type I errors are possible, *even in the absence of a deterrence motive*. In our framework, extreme sentences work because they maximize the screening power of plea bargaining.

Intuition for Theorem 1: The proof of Theorem 1 constructs a welfare improvement conditional on each defendant type. The signal is used to devise a sentencing scheme that induces the defendant to

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

report truthfully, and the relevant incentive constraint is dissuading a guilty defendant from pretending to be an innocent one. Starting from a (truthful) judicial mechanism, we construct a mapping from signals to sentences that minimizes the punishment to 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 (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. In fact, as Theorem 2 will make clear, the same result holds when Assumption 3 is relaxed. The role of concavity is to guarantee that the optimal sentence for a guilty defendant, which is considered in the second step of the proof, must be constant. Suppose 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 relevant 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 does, then we can decrease the sentence to $\hat{s}$, which gives the highest possible social welfare conditional on facing the guilty defendant.[28]

Interim-Optimal Mechanism without Risk Aversion: The optimality of a constant sentence following an admission of guilt relies on Assumption 3. Without this assumption, it is not immediately clear how the optimal sentencing function $S(\cdot, \hat{g})$ should depend on the signal, whether it should involve random sentences (recall that for a given signal t, $S(t, \hat{g})$ can be random), and if so, what properties these sentences should satisfy. The following result answers these questions.

Theorem 2 *Regardless of whether Assumption 3 holds:*

(i) In any interim 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 $\bar{t}$.[29]

(ii) There is an interim optimal mechanism such that 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 holds generically for interim optimal mechanisms.[30] *The guilty defendant's sentence can be chosen to be statistically independent of the signal.*

(iii) If the guilty defendant's sentence in an interim-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 $[0, \hat{s}]$. If $W(\cdot, g)$ and $u(\cdot)$ are concave and at least one of them is strictly concave, then a random plea cannot be optimal.

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

[29] The necessity of this property follows from the uniqueness proof for Theorem 1. See Appendix D.

[30] Here "generically" is understood in the sense of *prevalent sets* over the vector space of welfare functions. See Appendix E.

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

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. The only new step to prove Theorem 2 is the following one. We consider the guilty defendant's utility from his sentence rather than the sentence itself, and find the utility distribution that maximizes social welfare while maintaining the defendant' expected utility for the guilty, using a concavification argument that is often used in contract theory and strategic communication.

3.2 Ex-Ante Welfare and Deterrence

While interim welfare is 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.[31] In this case, therefore, ex-ante welfare also increases. In particular, Theorem 1 identifies 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 the ex-ante optimal mechanisms, which maximize ex-ante welfare (6) among all available mechanisms. The result shows that ex-ante optimal mechanisms (with or without Assumption 3) are similar to interim optimal mechanisms without Assumption 3, as described

[31] 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 Theorems 1 and 3 increase the expected utility of an innocent defendant, so if this utility had any impact on the incentives to commit a crime, these improvements would reduce crime incentives even further.

in Theorem 2. 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 a minor adaptation of the proof of Theorem 2 is required to show that parts (i), (ii), and (iv) of Theorem 2 also hold for ex-ante optimal mechanisms. Part (iii) of Theorem 2 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 3 *(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 $\bar{t}$.*[32]

(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.[33] *The guilty defendant's sentence is 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, that is (1), holds as an equality.

Theorem 3 shows that it may be optimal to give the guilty defendant a fixed deterministic sentence even when this sentence exceeds $\hat{s}$. To understand when a random sentence is optimal, 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 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 $\hat{W}(U^g) < \bar{W}(U^g)$.[34]

[32] The necessity of this property follows from the uniqueness proof for Theorem 1. See Appendix D.

[33] The notion of genericity is the same as in Theorem 2.

[34] 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 $\bar{s} = 4$.

4 From Judicial Processes to Judicial Mechanisms

This section provides a microfoundation for the mechanism design analysis performed in earlier sections. It formalizes the definition of judicial processes and their reduction to judicial mechanisms, considering *interim* and *ex-ante* perspectives.

Interim perspective:

- We formally define a *judicial case* and a *judicial process* from an interim perspective, i.e., after the arrest of a suspect. A judicial case is crime- and defendant-specific: all the primitives can depend on the crime that the defendant is accused of, the circumstances of the crime and, more generally, all the information available at the time of the arrest.
- We show how to construct, from any judicial process, a truthful direct-revelation mechanism focused on the defendant.
- We prove the following statistical lemma: the sentence given to a defendant can be chosen without loss of generality so as to depend only on (i) the information generated by the judicial process about the defendant's guilt and (ii) the defendant's equilibrium strategy in the judicial process.
- We introduce the main *invariance property* that the set of all judicial processes available to the social planner must satisfy, and state a general mechanism design problem from an interim welfare perspective.

Ex-ante perspective:

- We consider a setting in which each individual decides whether to commit a crime and each individual can be prosecuted.
- We define a *judicial system* as a map from each judicial case to a judicial process.
- The objective is to find an optimal judicial system according to an ex-ante notion of welfare that includes deterrence.
- We characterize the set of individuals who commit crime as a function of the judicial system. If an individual commits a crime, he does not necessarily know what judicial case he will face if and when he is arrested, since this may depend on the information accumulated after the crime is committed and before the arrest.
- We generalize our the invariance assumption to this setting and state a general mechanism design problem from an ex-ante perspective.

4.1 Definitions (Interim Perspective)

4.1.1 Judicial Case

Definition 1 *A **judicial case** J specifies:*

- *A set $\mathcal{I}$ of players that includes Nature and a defendant.*
- *For each player $j \in \mathcal{I}$, a private type τ_j in some type space T_j.*
- *A distribution μ of players' types defined over the type space $\mathcal{T} = \times_{j \in \mathcal{I}} T_j$, endowed with some σ-algebra.*

The set of players, the players' type space, and their type distribution all depend on the judicial case J. For notational simplicity, we omit the dependence on J in the interim analysis. This will not lead to any confusion since J is fixed at the interim stage.

Intuitively, a judicial case describes the situation of a defendant who was just arrested in relation to a crime. A judicial case consists of a set of players, which include the defendant and every other person related to the case, as well as Nature to capture uncertainty of the judicial process, such as whether the defendant is lucky or unlucky with the evidence discovered in his case. The type distribution μ is conditional on all the public information accumulated until the time of the arrest.

Let $\mathcal{J}$ denote the set of all possible judicial cases. We make the following assumptions:

Reduction Assumption 1 *For any judicial case $J \in \mathcal{J}$:*

- *There is exactly **one** defendant.*
- *The defendant has a **binary** private type $T_d = \{g, i\}$: either he committed the crime ($\tau_d = g$), or he did not($\tau_d = i$).*
- *There is a prior probability $\lambda \in (0, 1)$ that the defendant is guilty.*
- *The type distribution μ has* full support *over $\mathcal{T}$.*

The full support assumption implies that it is *impossible to perfectly infer the defendant's guilt even if one knows the types of all other players.* This inference can be arbitrarily precise as long as it is imperfect. Players' types are typically correlated. For instance, the type of a witness (what he has observed) is correlated with the type (guilt or innocence) of the defendant. Likewise, Nature may turn up some evidence (e.g., fingerprints) with a probability that depends on the defendant's guilt.

4.1.2 Judicial Process

A judicial case gives rise to a judicial process, whose objective is to generate and aggregate information about the defendant's guilt and provide an adequate punishment to the defendant.

Definition 2 *Given a judicial case, a* ***judicial process*** P *is defined by:*

(i) An extensive-form game with incomplete information and a finite horizon, whose player set and type structures are given by the judicial case.

- *For each player j, let Σ_j denote the set of mixed strategies available to j in the extensive-form game. Each type τ_j chooses a mixed strategy $\sigma_j(\tau_j) \in \Sigma_j$. A strategy for player j is a collection $\sigma_j = \{\sigma_j(\tau_j) : \tau_j \in T_j\}$.*
- *We let $\mathcal{H}$ denote the set of possible histories $h \in \mathcal{H}$. Each history describes a realized play of all players in the game. A mixed-strategy profile $(\sigma_j : j \in \mathcal{I})$ generates a probability distribution over $\mathcal{H}$.*

(ii) A specific strategy profile $\sigma^P = \{\sigma_j^P(\tau_j)\}_{j \in \mathcal{I}, \tau_j \in T_j}$.

- *The strategy profile σ^P must satisfy some solution concept, explained in the next subsection ("Payoffs and Incentives").*

(iii) An ***outcome*** *function $h \mapsto (s(h), c(h)) \in \mathbb{R}_+^2$ that associates to each history a* ***sentence*** *and a* ***cost****.*

- *The sentence function s takes values in an interval of the form $[0, \bar{s}_J]$. The maximal punishment $\bar{s}_J > 0$ can depend arbitrarily on the case J.*

Players' strategy sets, the specific profile, and the outcome function depend on the judicial process P.

The outcome is interpreted as a sentence for the defendant and a social cost for running the process. By definition, the outcome depends only on the history of play, h, not on the defendant's type. In particular, a defendant's treatment by the judicial process does not depend directly on whether the defendant is guilty.

Remark 1 *The definition of a judicial process assumes that the strategy set Σ_j of each player j is the same for all types $\tau_j \in T_j$. In particular, at each information set, each player, including Nature, has the same action set regardless of his type.*

From an interim perspective, the objective of the social planner is to choose a judicial process that maximizes a welfare objective that will be defined shortly.

Let $\mathcal{P}(J)$ denote the set of judicial processes corresponding to judicial case J. This is the set of judicial processes that the social planner can choose from.

4.2 Payoffs and Incentives (Interim Perspective)

Utility: To any sentence $s \in [0, \bar{s}_J]$ corresponds a utility level $u_J(s)$ for the defendant. The defendant's utility can depend on the judicial case but is independent of his type.[35]

Reduction Assumption 2 *The utility function u_J is independent of the defendant's type.*

Since the judicial case J is fixed in the interim perspective, we omit the subscript J from the notation. Given a judicial process P and a strategy σ_d for the defendant, let $U^P(\theta, \sigma_d)$ denote the defendant's expected utility when his type is $\tau_d = \theta \in \{g, i\}$ and he plays strategy $\sigma_d \in \Sigma_d$ in the judicial process P while other players follow the strategy σ^P prescribed by P. Formally,

$$U^P(\theta, \sigma_d) = E[u(s(h))|\sigma^P, \sigma_d, \tau_d = \theta].$$

A strategy profile σ^P is *defendant-optimal* if

$$U^P(\theta, \sigma_d^P(\theta)) \geq U^P(\theta, \sigma_d)$$

for all $\sigma_d \in \Sigma_d$ and $\theta \in \{g, i\}$.

Defendant-optimality means that playing according to σ^P is incentive compatible from the perspective of the defendant. It is a minimum requirement imposed on the solution concept that the strategy profile σ^P must satisfy.[36]

Reduction Assumption 3 *For any judicial process $P \in \mathcal{P}(J)$, the strategy profile σ^P is defendant-optimal.*

Realized and Interim Welfare Given a judicial process $P \in \mathcal{P}(J)$ and history h, the realized interim welfare is equal to

$$w_J(h, \theta) = W_J(s(h), \theta) - c(h)$$

where $W_J : [0, \bar{s}] \times \{g, i\} \to \mathbb{R}_-$ can depend on the judicial case J and captures the Type I/II error considerations coming from giving a sentence s to a defendant of type θ. Assumption 2 requires that $W_J(s, i)$ be decreasing in s: a longer sentence to an innocent defendant decreases welfare.

[35]This assumption is common in the literature on hard evidence (see, e.g., Ben-Porath, Dekel, and Lipman 2014).

[36]One could additionally require that σ^P be a Perfect Bayesian equilibrium or obey a different solution concept. Such a requirement does not affect our analysis.

Given a judicial case J, the planner's objective is to maximize

$$\mathcal{W}_J(P) = E[w_J(h, \theta)]$$

over all judicial processes in $P \in \mathcal{P}(J)$, where the expectation is taken with respect to the random vector (h, θ) given the prior distribution μ.

4.3 Informativeness of Judicial Processes

Overview: A judicial process reveals information about the defendant: given the strategy profile σ^P implemented by P, the likelihood of each history h depends on the defendant's type. To each history h corresponds a signal, which captures the likelihood ratio of guilt associated with h.

In order to analyze a defendant's incentives, it is useful to define this likelihood ratio of guilt *conditional on each strategy* σ_d that the defendant may choose. The result is a *conditional likelihood ratio* for each history, which we denote $\ell(h, \sigma_d)$: it is the likelihood ratio of the defendant's guilt given that (i) the defendant plays strategy σ_d, (ii) all other players play the prescribed profile σ^P, and (iii) history h was realized.

This section shows that the following result: if the defendant chooses strategy σ_d, then for any $l \in (0, \infty)$, the distribution of the sentence $s(h)$ conditional on $\ell(h; \sigma_d) = l$ is *independent of the defendant's type*. Put differently, the only thing that matters about history h and the defendant's type θ, as far as sentencing is concerned, is the sufficient statistic $\ell(h; \sigma_d)$.

This result will be used later to simplify the sentence function $h \mapsto s(h)$ of any judicial process P by replacing it by a sentence lottery $S(l) \in \Delta([0, \bar{s}])$ that depends on the history h only through the information $l = \ell(h, \sigma_d)$ that h contains given the defendant's strategy σ_d.

Likelihood Ratio: Formally, if there were finitely many histories, $\ell(h; \sigma_d)$ would be defined by:

$$\ell(h; \sigma_d) = \frac{Pr(h|\theta = g, \sigma_d)}{Pr(h|\theta = i, \sigma_d)}.$$

However, we wish to allow for a continuum of signals and, hence, histories. To define ℓ in the general case, let $\Delta(\theta; \sigma_d)$ denote the probability distribution over the set $\mathcal{H}$ of histories if the defendant has type θ and follows policy σ_d. Then, $\ell(h; \sigma_d)$ is the Radon-Nikodym derivative of $\Delta(g; \sigma_d)$ with respect to $\Delta(i; \sigma_d)$. Our assumptions that players' type distributions have full support and that players' strategies spaces are type-independent guarantee that $\Delta(g; \sigma_d)$ is absolutely continuous with respect to $\Delta(i; \sigma_d)$ and, hence, that $\ell(h; \sigma_d)$ is well defined.

The quantity $\ell(h; \sigma_d)$ is the likelihood ratio of guilt of the defendant when (i) history h is observed and (ii) it is known that the defendant followed strategy σ_d, (iii) other players play σ^P. It describes

how much information the judicial process reveals about the defendant's type if history h occurs and the defendant is known to have followe strategy σ_d.

Information-Based Sentencing: The next result shows that, for any strategy σ_d followed by the defendant, the sentence lottery faced by a defendant conditional on a given information (likelihood ratio) revealed by the judicial process is statistically *independent* of the defendant's type.

Proposition 1 *Fix a judicial process P and defendant strategy σ_d. For any $l \in (0, \infty)$,*[37] *the random variables $s(h)$ and $c(h)$ are distributed independently of θ conditional on $\ell(h, \sigma_d) = l$ and on the defendant playing strategy σ_d.*

Proof. Fix some $l \in (0, \infty)$ and let $\mathcal{H}(l)$ denote the set of histories h for which $\ell(h; \sigma_d) = l$. If no such history exists, the claim of Proposition 1 is true by default, so suppose that $\mathcal{H}(l) \neq \emptyset$. For $\theta \in \{g, i\}$, let $F_l(\cdot|\theta)$ denote the probability distribution of h over $\mathcal{H}(l)$ conditional on strategy σ_d and defendant type θ.

We will show that $F_l(\cdot|\theta)$ is independent of θ. Since $\ell(\cdot; \sigma_d)$ is the Radon-Nikodym derivative of $\Delta(g; \sigma_d)$ with respect to $\Delta(i; \sigma_d)$, we have, for any measurable subset $\mathcal{B}$ of $\mathcal{H}(l)$,

$$\Delta(g, \sigma_d)(\mathcal{B}) = \int_{\mathcal{B}} \ell(h; \sigma_d) d\Delta(i, \sigma_d)(h).$$

Since $\ell(h; \sigma_d)$ is by construction constant and equal to l over $h \in \mathcal{B}$, we can take it out of the integral, and obtain

$$\Delta(g, \sigma_d)(\mathcal{B}) = l\Delta(i, \sigma_d)(\mathcal{B}).$$

Since this is true in particular for $\mathcal{B} = \mathcal{H}(l)$, we have

$$\Delta(g, \sigma_d)(\mathcal{H}(l)) = l\Delta(i, \sigma_d)(\mathcal{H}(l)).$$

By definition of F_l, we have $F_l(\mathcal{B}|\theta) = \frac{\Delta(\theta, \sigma_d)(\mathcal{B})}{\Delta(\theta, \sigma_d)(\mathcal{H}(l))}$ for each $\theta \in \{g, i\}$. The previous two equations then imply that

$$F_l(\mathcal{B}|g) = F_l(\mathcal{B}|i)$$

for any measurable subset $\mathcal{B}$ of $\mathcal{H}(l)$, which shows that $F_l(\cdot|\theta)$ is independent of θ, as desired. Since $s(h)$ and $c(h)$ are deterministic functions of h, this implies that their distributions are also independent of θ, conditional on l and σ_d. ■

4.4 Direct-Revelation Mechanism

Fix a judicial case J. From any judicial process $P \in \mathcal{P}(J)$, we may construct a simpler game with the following properties:

[37] Note that

- There is only one player, the defendant. The social planner receives a payoff from the game but does not take any action.
- The defendant has a binary strategy space: report $\hat{\theta} \in \{\hat{g}, \hat{\imath}\}$.
- The outcome of the game is a lottery over sentences and costs.
- Given a defendant type θ and report $\hat{\theta}$, the game assigns a distribution of the sentence s and the cost c that is the same as in the judicial process P if the defendant has type θ and follows strategy $\sigma_d^P(\hat{\theta})$ in P and other players follow the strategies prescribed by σ^P.
- The defendant's realized utility $u(s)$ and the planner's realized welfare $W(s,\theta) - c$ are defined as in P.

Let $F_\theta^{\hat{\theta}}$ denote the distribution of the likelihood ratio $\ell(h, \sigma_d^P(\hat{\theta}))$ if the defendant has type θ and follows strategy $\sigma_d^P(\hat{\theta})$. This defines a 4-tuple of distributions. Likewise, we can associate a cost $C_\theta^{\hat{\theta}}$ that corresponds to the expected cost in the judicial process P when the defendant has type θ and follows strategy $\sigma_d^P(\hat{\theta})$. This defines a 4-tuple of expected costs.

Proposition 1 implies that conditional on defendant strategy $\sigma_d^P(\hat{\theta})$ for $\hat{\theta} \in \{\hat{g}, \hat{\imath}\}$ and likelihood ratio l, the distribution of s is independent of θ. Let $S(l, \hat{\theta}) \in \Delta([0, \bar{s}])$ denote a lottery with this distribution. We call the mapping $S : (l, \hat{\theta}) \mapsto S(l, \hat{\theta})$ the *sentencing scheme* associated with P.

In summary, given a judicial case J, we associate to any judicial process P a 4-tuple F of signal distributions, a 4-tuple C of expected costs, and a sentencing scheme S.

Definition 3 *Given a judicial case J and a judicial process P, the direct-revelation mechanism (or "DRM") $M(P)$ associated with P is a single-agent mechanism focused on the defendant in which:*

- *The agent's strategy is to report a type $\hat{\theta} \in \{\hat{g}, \hat{\imath}\}$.*
- *Given a defendant type θ and report $\hat{\theta}$, a signal l is generated according to distribution $F_\theta^{\hat{\theta}}$.*
- *A sentence s is realized according to the lottery $S(\ell, \hat{\theta})$.*
- *The defendant's expected utility if he reports $\hat{\theta}$ and his type is θ is $U_\theta^{\hat{\theta}} = E[u(S(\ell, \hat{\theta}))|F_\theta^{\hat{\theta}}]$.*
- *Expected welfare if the defendant has type θ and reports $\hat{\theta}$ is $W_\theta^{\hat{\theta}} = E[W(S(\ell, \hat{\theta}), \theta)|F_\theta^{\hat{\theta}}] - C_\theta^{\hat{\theta}}$.*

Formally, given a judicial case and a judicial process, a DRM can be reduced to a triplet (F, C, S) where F is the 4-tuple distribution, C is a 4-dimensional vector, and S is a sentencing scheme. The sentencing scheme S does not require the planner to know the defendant's true type: the scheme depends only on the information made available through the mechanism: namely, the agent's report and the information (likelihood ratio) l.

Proposition 2 *Given a judicial case J, a judicial process P, and associated DRM $M(P)$, truth-telling is an optimal strategy for the defendant in $M(P)$ and expected welfare is the same as under P.*

Proof. If a defendant with type θ reports $\hat{\theta}$, he generates a signal l according to $F_\theta^{\hat{\theta}}$. By construction of the sentencing scheme S, the sentence lottery faced by the defendant is the same as in the initial judicial process if he follows $\sigma_d^P(\hat{\theta})$. Since the process P is defendant-optimal, reporting truthfully is optimal. Given this, realized welfare is distributed as in the initial judicial process. ■

4.5 Invariance Assumption and Statement of the Mechanism Design Problem (Interim Perspective)

Given a judicial case J, let $\mathcal{M}(J) = \{M(P) : P \in \mathcal{P}(J)\}$ denote the set of all direct-revelation mechanisms **available** to the planner. Given a signal-distribution tuple F and sentencing scheme $S : (l, \hat{\theta}) \mapsto S(l, \hat{\theta}) \in \Delta([0, \bar{s}])$, we recall that

$$U_\theta^{\hat{\theta}} = E[u(S(l, \hat{\theta}))|F_\theta^{\hat{\theta}}],$$

where the expectation is taken with respect to the signal l, which has distribution $F_\theta^{\hat{\theta}}$, and the lottery S. A sentencing scheme S is said to be F**-truthful** if for each $\theta, \theta' \in \{g, i\}$,

$$U_\theta^\theta \geq U_\theta^{\theta'}.$$

We now state the key assumption, which allows us to perform mechanism design analysis.

Reduction Assumption 4 (Invariance) *If $(F, C, S) \in \mathcal{M}(J)$ and S' is F-truthful, then $(F, C, S') \in \mathcal{M}(J)$.*

When this assumption holds, we say that $\mathcal{M}(J)$ *is invariant.*

Invariance means that if a judicial process that generates a particular signal distribution F is available to the planner, then any sentencing scheme that induces truth-telling by the defendant given the signal distribution F is also available to the planner at the same expected costs. Put differently, the social planner can choose any truthful sentencing scheme given F without affecting the availability of the signal distribution F or its expected cost.

This property allows us to focus on the minimal incentive compatibility constraint that any judicial process must satisfy, which is that a guilty defendant should not strictly benefit from mimicking the strategy of an innocent defendant, and vice versa. We are now ready to state the mechanism design problem (interim perspective):

Problem 1 (Interim-Global) *Given a judicial case J and invariant set $\mathcal{M}(J)$, solve*

$$\max_{(F,C,S)\in\mathcal{M}(J)} \lambda_J \left(E[W(S(l, \hat{g}), g)|F_g^{\hat{g}}] - C_g^{\hat{g}}\right) + (1 - \lambda_J) \left(E[W(S(l, \hat{\imath}), i)|F_i^{\hat{\imath}}] - C_i^{\hat{\imath}}\right).$$

Theorems 1 and 2 do not fully characterize the optimum (F^*, C^*, S^*) of this problem. Rather, they provide properties of the sentencing scheme S^* that must be satisfied by at the optimum. They concern a simpler mechanism design problem:

Problem 2 (Interim-Sentencing) *Given a judicial case J, invariant class $\mathcal{M}(J)$, and tuples F and C, solve*

$$\max_{\{S:(F,C,S)\in\mathcal{M}(J)\}} \lambda_J \left(E[W(S(l,\hat{g}),g)|F_g^{\hat{g}}] - C_g^{\hat{g}} \right) + (1-\lambda_J) \left(E[W(S(l,\hat{\imath}),i)|F_i^{\hat{\imath}}] - C_i^{\hat{\imath}} \right).$$

4.6 Judicial System: Ex-Ante Perspective

This section provides a more general ex-ante formulation than the one studied in Sections 2 and 3. This formulation allows heterogeneity in the types of crimes committed and in the judicial cases that defendants may face. This generalization allows us to describe an improvement for the entire justice system, which treats all kinds of crimes and judicial cases together. The proof of Theorem 3 readily extends to this more general setting, because the improvement constructed in the proof can be applied judicial case by judicial case.

From an ex-ante perspective, we consider the entire set $\mathcal{J}$ of possible judicial cases that may be generated after crimes have occurred..

Definition 4 *Given a correspondence $J \rightrightarrows \mathcal{P}(J)$ that assigns to each judicial case $J \in \mathcal{J}$ a set of judicial processes available to the social planner, a **judicial system** Y is a selection from this correspondence:*

$$Y : J \mapsto P(J) \in \mathcal{P}(J).$$

From an ex-ante perspective, a judicial system is the object that social planner optimizes over (keeping fixed the level of law enforcement other possible instruments for the social planner). When choosing a judicial system, the social planner takes into account its effect on deterrence, as explained next.

4.7 Deterrence and Welfare: Ex-Ante Perspective

We consider a society in which each individual has an opportunity to commit a single crime and runs the risk of being prosecuted for a crime. The timing of the game is as follows:

- **Step 0:** The planner chooses a judicial system Y.
- **Step 1:** Each individual decides whether to commit crime.

- **Step 2:** For each committed crime, at most one individual is arrested and prosecuted. Each crime generates at most one judicial case, in which the defendant may or may not be the individual who committed the crime.[38]

We assume that an individual is prosecuted for at most one crime. Each individual faces a lottery over judicial cases that depends on whether he committed a crime and, likewise, each crime gives rise to a distribution over judicial cases. These lotteries are related in potentially complex ways, which we do not specify explicitly. **Crime incentives:** When an individual considers whether to commit crime, he trades off his privately known (and possibly negative) benefit from committing the crime with the expected cost of punishment.

Given a judicial system Y, consider an individual m with a benefit $b_m \in \mathbb{R}$ from committing the crime. If the individual commits the crime, he is uncertain about the judicial case $\tilde{J}$ that he may face (if at all), because the judicial case will depend on the information that is accumulated after he commits the crime and before he gets arrested (if he gets arrested). This distribution must satisfy properties that are described below.

If individual m commits a crime, his expected utility is

$$U_m(g, Y) = b_m + E[U(\tilde{J}, g)|g]$$

where $U(\tilde{J}, \theta)$ is the expected utility of a defendant of type θ conditional on judicial case $\tilde{J}$, as defined earlier, and the expectation is taken over cases $\tilde{J}$ conditional on m committing the crime.

If m does not commit a crime, his expected utility is

$$U_m(i, Y) = E[U(\tilde{J}, i)|i],$$

because he may still be arrested for a case committed by someone else. [39]

Therefore, m commits the crime if and only if

$$b_m > \beta_m(Y) = U_m(i, Y) - U_m(g, Y).$$

Reduction Assumption 5 *(i) The benefits from crime $\{b_m\}_m$ is independently distributed in the population. (ii) For each m, the distribution of b_m does not have any atoms.*

[38] We could consider crimeless judicial cases, in which a non-criminal act was committed and some individual is arrested and prosecuted to determine whether the act was indeed criminal. For example, it may be unclear at the time of arrest whether the individual who committed the act intended to commit it.

[39] Another interpretation is that m may commit a benign act that is wrongly interpreted as a crime and leads to m's arrest. See Kaplow (2011).

Both guilty and innocent individuals have a chance *not* to be arrested. This possibility is formally captured by including a special judicial case, "$J_\emptyset$," that corresponds to getting a zero sentence with probability 1 (i.e., the upper bound on the sentence is $\bar{s}_{J_\emptyset} = 0$). The distribution of the judicial case faced by an individual depends on whether this individual is guilty. In particular, it is natural to assume that $Pr(J_\emptyset) \approx 1$ for an innocent individual and $Pr(J_\emptyset) < 1$ for a guilty one.

A judicial system Y determines the set $\mathcal{N}$ of citizens who commit the crime:

$$\mathcal{N}(Y) = \{m : U_m(g, Y) > U_m(i, Y)\}.$$

The more deterring the system Y, the smaller this set.

We make the following independence assumptions:

Reduction Assumption 6 *Let $\mathcal{Y}$ denote the set of judicial systems available to the judicial planner.*

1. *For each individual m, the distribution of the case $\tilde{J}$ that m will face conditional on his type $\theta_m \in \{g, i\}$ is independent of $Y \in \mathcal{Y}$.*

2. *For each individual m, the distribution of the case $\tilde{J}$ that m will face conditional on his type $\theta_m \in \{g, i\}$ is independent of b_m.*

3. *Conditional on a judicial case J in which m is the defendant, m's utility function u_J from the sentence is independent of θ_m and b_m.*

These assumptions are made for tractability. The first assumption means that, while the judicial system can affect the set of individuals who commit crime, it does not affect the judicial case generated by an individual conditional on his decision of whether to commit crime. This rules out the possibility that the number of crimes committed affects the ability of the justice system to arrest the correct defendant. It may be a reasonable assumption if the number of committed crimes represents a small fraction of the population. If we allowed the probability of wrongful arrests to increase in the number of committed crimes, there could be multiple equilibria: high-crime-frequency equilibria in which judicial mistakes are frequent and individuals have less incentive to abstain from crime because they are more likely to be convicted when they are innocent, and low-crime-frequency equilibria in which wrongful arrests are rare and the incentives to abstain from crime are high.

The third assumption rules out the possibility that there is some correlation between a defendant's guilt and, say, his risk aversion over a stochastic sentence.

4.8 Invariance Assumption and Statement of the Mechanism Design Problem: Ex-Ante Perspective

In principle, the social planner could choose a level of law enforcement that would affect the distributions of $\tilde{J}$ conditional on θ. We ignore this aspect for simplicity and treat the level of law enforcement as given.[40]

For each judicial case J, let $q_J < 0$ denote the social harm caused by the crime that was committed. When an individual commits a crime, the harm from the crime may be uncertain, and q_J represents the realized harm at the time of the arrest or, more generally, the expectation of the realized harm given the information available of the time of the arrest.

A judicial system Y specifies a DRM (F_J, C_J, S_J) for each case J. The ex-ante welfare associated with a judicial system $\mathcal{Y}$ is:

$$\begin{aligned}\mathcal{W}(Y) \quad &= \int_m Pr(b_m > \beta_m(Y)) E_m \Big[\lambda_{\tilde{J}} \Big(E[W(S_{\tilde{J}}(l,\hat{g}),g)|(F^{\hat{g}}_g)_{\tilde{J}}] - (C^{\hat{g}}_g)_{\tilde{J}}\Big) \\ &\quad + (1-\lambda_{\tilde{J}}) \Big(E[W(S_{\tilde{J}}(l,\hat{i}),i)|(F^{\hat{i}}_i)_{\tilde{J}}] - (C^{\hat{i}}_i)_{\tilde{J}}\Big) - q_{\tilde{J}}\Big]\end{aligned} \tag{7}$$

where the distribution of $\tilde{J}$ is taken conditional on a crime being committed by m.
The set $\mathcal{Y}$ of judicial systems available to the planner is **invariant** if, for any judicial case J, the class $\mathcal{M}(J)$ of DRM given J is invariant.

Problem 3 (Ex-Ante) *Given an invariant set $\mathcal{Y}$ of judicial systems, choose a judicial system Y : $J \mapsto P(J) \in \mathcal{P}(J)$ to maximize $\mathcal{W}(Y)$ over $\mathcal{Y}$.*

The problem is analyzed using the same reduction to DRMs as in the interim case.

Symmetric Version: Sections 2 and 3 consider a simpler version of the ex-ante mechanism design problem, in which there is only one type of crime, and only one judicial case—except for the identity defendant and other actors—for each crime committed. It is straightforward to check that the formula (7) reduces to (6) for the symmetric case and, reciprocally, that the argument used to prove Theorem 3 extends to the asymmetric case, because the improvement constructed in Theorem 3 can be applied judicial case by judicial case and does not affect deterrence even in the asymmetric case.

5 Comparing Optimal Mechanisms with Existing Legal Systems

Many features identified by Theorems 1, 2, and 3 are found in existing legal systems, but other features are unrealistic. This section compares our findings to existing legal systems and discusses key differences.

[40] Treating law enforcement as a decision variable would not change our results, which would still apply by doing a partial optimization conditional on the optimal level of law enforcement.

Signal-Independent Plea Sentence: The first feature is the fixed sentence given to a defendant who reports he is guilty. This feature is similar to the plea bargaining procedure in the United States and in other countries. A plea bargain makes a trial unnecessary, and the plea sentence does not depend on a trial's outcome or on evidence that would have been produced in the course of a trial. Even when the optimal scheme for a guilty defendant involves a lottery over two sentences, this lottery is independent of the signal about the defendant's guilt. By contrast, a defendant who claims to be innocent faces a sentence that optimally depends on the signal (but which also takes two possible values). A lottery that disregards the signal is similar to a plea bargain with uncertain punishments, as is the case when the plea bargain does not specify a particular sentence or when the judge can decide on a sentence other than the one specified, without allowing the defendant to withdraw his plea.[41] Since the punishment in such pleas is determined without a trial, it does not depend on the evidence that a trial would have generated. Such a lottery is also consistent with the institution of parole, which introduces a stochastic element to guilty plea sentences, and with the discretion of a judge of whether to accept a guilty plea agreement between the defendant and the prosecution.

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 $U \to W(u^{-1}(U), g)$ is convex, even when both u and $W(\cdot, g)$ are concave.[42] Put differently, lotteries become an efficient way of a punishing a guilty defendant. Under Assumption 3 lotteries arise optimally only when deterrence is a significant consideration in sentencing. Therefore, an empirical prediction of the theory is that stochastic sentences following a guilty plea are more likely to arise for crimes whose commission depends more elastically on deterrence.

Binary Verdicts with Acquittal or Severe Punishment: Another feature that is familiar in the American legal system is that a defendant who reports he is innocent receives either a null sentence or some fixed higher sentence, depending on whether the signal is higher than some threshold. This feature can be interpreted as a trial with two possible outcomes, an acquittal or a conviction, in which the punishment following a conviction is independent of the strength of evidence that led to the conviction. The outcome is determined by an evidence threshold criterion: based only on the evidence (signal), the defendant is convicted if and only if a guilty defendant is sufficiently more likely than an innocent defendant to have produced such evidence. The evidence threshold is high if it is more important to

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

[42] Indeed, note that the composition $g \circ f$ of two concave functions is guaranteed to be concave only if g is increasing.

acquit innocent defendants than to punish guilty ones, a preference that will be captured by the welfare functions $W(\cdot, g)$ and $W(\cdot, i)$.

An acquittal carries no punishment. We did not assume that judicial mechanisms must include a zero sentence. Instead, acquittals emerge as a feature of the optimal mechanisms. We also did not assume binary verdicts. This ubiquitous feature of trial systems emerges as a feature of optimal mechanisms. Intuitively, binary verdicts are optimal because they provide the optimal separation power between guilty and innocent defendants: to make the sentencing scheme the least attractive possible to a guilty defendant (and hence relax his incentive compatibility constraint), it is optimal to give the harshest possible sentence for evidence that is most likely to have been generated by a guilty defendant, and the most lenient sentence for evidence most likely to have been generated by an innocent defendant. As noted in Section 3, this justification for using a maximal sentence is unrelated to Becker's (1968) justification, which is based on enforcement costs. In practice, there may be ethical or practical considerations that limit the maximal allowable sentence in a given trial or for a given crime.[43]

Conviction Threshold: An important difference between the optimal mechanisms and criminal trials in practice concerns the conviction threshold used to convict defendants. In many countries, the punishment following a conviction increases with the severity of the crime, but the conviction criterion is fixed: For example, the United States guilt must be established "beyond a reasonable doubt" (BARD) regardless of the crime. In optimal mechanisms, by contrast, the conviction threshold could optimally vary across crimes. This threshold is determined by the optimal level of utility for the guilty defendant, which in turn depend on society's welfare function. This welfare function could depend on the specific crime. For example, the sentence $\hat{s}$ from Assumption 3 that maximizes social welfare when facing a guilty defendant is likely higher for more serious crimes. This finding suggests that existing judicial systems could be improved by incorporating more nuanced conviction criteria, which vary with the crime (and possibly other factors).[44]

Role of Evidence As noted in previous sections, another important difference is between the role that evidence plays in optimal mechanisms and the one it appears to play in actual criminal trials. In a trial, evidence is used to determine whether the defendant is guilty; in an optimal mechanism, evidence is used to incentivize guilty defendants to admit their guilt. Defendants who claim to be

[43]The sentence associated with a "guilty" verdict may depend on many factors we do not explicitly model. The effect of many of these factors, such as the defendant's criminal history or aggravating circumstances, can be captured by varying the maximal sentence $\bar{s}$.

[44]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.

innocent are either set free or severely punished, based on the evidence. "Incriminating evidence," that is, evidence sufficiently more likely to be produced by a guilty defendant than an innocent one, leads to punishment. But since all guilty (and only guilty) defendants admit their guilt, the informational content of the evidence plays no role in determining the defendant's actual guilt. This role of evidence in the optimal mechanisms is tightly linked to Assumption 1. We now discuss this connection and also discuss how evidence regains the role of determining guilt when optimal mechanisms are modified slightly.

5.1 Invariance Property and Commitment

If we interpret the binary sentencing scheme for innocent defendants as the outcome of a trial, optimal mechanisms have the property that only innocent defendants go to trial.[45] This feature relies on Assumption 1 because the proofs of Theorems 1, 2, and 3 require that changing the sentencing scheme S does not affect the distribution of the signal. In particular, Assumption 1 implies that even if jurors are convinced that only innocent defendants go to trial, and even though the punishment following the conviction is severe, they would still reach a "guilty" verdict if the evidence is sufficiently incriminating.

The importance of minimizing the influence of the punishment severity on the verdict determination has been recognized in criminal trials in the United States. One relevant feature is the separation between the fact-finding stage, in which jurors play a decisive role, and the sentencing stage, in which the judge or judges play a more important role. This separate allows jurors to entirely focus on the evidence presented to them to assess guilt without dwelling on the punishment that a conviction would bring. 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)). Instructions to the jury entirely focus on describing the procedure for finding facts. 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)). In many cases, jurors are unaware of the minimum-punishment guidelines relevant for the case.[46] There is compelling evidence that jurors have a limited understanding of the sentences faced by defendants. For example, the Capital

[45] Complete separation also arises in numerous other papers in law and economics, including Grossman and Katz (1983). In an extension, these authors show that when defendants are heterogeneous in their degree of risk aversion, partial pooling can arise.

[46] For example, in *State v. May* (Arizona Superior Court, 2007) a thirty-five-year-old defendant was sentenced to 75 years in jail after being found guilty of touching, in a residential swimming pool, the clothing of four children in the vicinity of their genitals (Nelson, 2013). Jurors had doubts about the guilt of the defendant: they were twice unable to reach a verdict within the first three days of deliberation. It is very likely that they were surprised by the extreme punishment handed down after the very narrow conviction.

Jury Project found that most jurors "grossly underestimated" the amount of jail time associated with a guilty verdict. There is also evidence that harsher sentences do not result in lower conviction rates. In a study of non-homicide violent case-level data of North Carolina Superior Courts, Da Silveira (2017) finds that the probability of conviction of defendants going to trial in fact increases with the sentence that they face.[47] This correlation cannot be easily explained away by prosecutor behavior. For example, if prosecutors attached more importance to obtaining a conviction when the case is more severe, they would send to trial defendants who are more likely to be found guilty and obtain a guilty plea from the other ones, and one would expect the probability of plea settlements to increase with the severity of the sentence associated with a conviction. This relation seems contradicted by the data.[48]

5.2 Trials with Guilty Defendants

In reality, most defendants convicted in trial are guilty. One way to reconcile this with our characterization of optimal mechanisms without concluding that existing trials are very far from optimal is to notice that in the optimal mechanisms guilty defendants are indifferent between taking a plea and going to trial. 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, which may provide a more applied justification for Assumption 1.

Suppose that under the optimal sentencing scheme a fraction α 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 λ 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

$$\hat{\lambda} = \frac{\lambda\alpha}{\lambda\alpha + (1-\lambda)}. \tag{8}$$

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 (8), we have

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

[47] Da Silveira's analysis excludes the most and least severe cases to focus on a relatively homogeneous pool of cases.

[48] Elder (1989) finds evidence that circumstances that may aggravate punishment *reduce* the probability of settlement. Similarly, Boylan (2012) finds that a 10-month increase in prison sentences raises trial rates by 1 percent.

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. That is, 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.[49] Also suppose that, consistent with criminal data in the United States, 90% of defendants are in fact guilty.[50] 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 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 λ of guilt, our modified mechanism remains consistent with BARD and the observation that most defendants are guilty. With a fraction α 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 regains its role in determining the defendant's guilt, in addition to its role in incentivizing most guilty defendants to accept the plea bargain.

5.3 Defendants with Additional Private Information

One of our key modeling assumptions is that the defendant's only private information is whether he is guilty. This assumption is in line with our objective, which is to provide a first step in the analysis of optimal criminal justice systems, because the defendant's private knowledge of his guilt is the universal issue that all criminal justice systems must contend with. In practice, of course, defendants may have additional private information: they may be informed about the strength of evidence that may be uncovered in the case, and may also differ in terms of their disutility from various sentencing schemes

[49] 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.

[50] 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.

and their risk attitudes.[51] Such heterogeneity provides a different explanation for why guilty defendants may go trial. Faced with the judicial mechanism described in this paper, guilty defendants who are even slightly less risk averse than the defendants in the baseline model, or slightly more optimistic about the evidence generated against them (because they were more cautious in the commission of the crime or for behavioral reasons), would strictly prefer to go to trial. Analyzing the optimal mechanism in such settings would be an interesting direction for future work.[52]

[51] Our analysis does, however, allow for private information by other actors in the system, such as the prosecutor. See Section 4.

[52] We considered an extension of the model in which there are three types of defendants: guilty, innocent, and "innocent-looking guilty." Defendants of the last type are indistinguishable from innocent defendants in that they generate the same signal distribution (they committed the "perfect crime") and have the same utility function. Therefore, the analysis reduces to two types: guilty defendants and innocent-looking defendants (the latter type pools innocent and innocent-looking guilty defendants). The only difference with respect to our baseline model is that the ex-post welfare function conditional on facing an innocent-looking defendant becomes a convex combination of the ex post welfare functions $W(\cdot, i)$ and $W(\cdot, g)$ and, in particular, is no longer an increasing transformation of the utility function u. The argument used to prove that a two-step sentence increases welfare no longer works.

A Proof of Theorem 1

We show that any available mechanism can be improved upon (weakly) by another available mechanism that satisfies (i) and (ii) in the statement of Theorem 1. Appendix D shows that the improvement is strict if the original mechanism does not satisfy (i) and (ii).

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 G.

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:

$$\int_0^1 u(\tilde{S}(t, \hat{\imath}))f_g^{\hat{\imath}}(t)dt = u(0)F_g^{\hat{\imath}}([0, \bar{t}]) + u(\bar{s})F_g^{\hat{\imath}}([\bar{t}, 1]) = \int_0^1 u(S(t, \hat{\imath}))f_g^{\hat{\imath}}(t)dt. \tag{9}$$

The cutoff $\bar{t}$ exists because distribution $F_g^{\hat{\imath}}$ has no atoms.[53] Rearranging (9) yields

$$\int_0^1 [u(S(t, \hat{\imath})) - u(\tilde{S}(t, \hat{\imath}))]f_g^{\hat{\imath}}(t)dt = 0. \tag{10}$$

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 $t \leq \bar{t}$ 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[54] then implies that

$$\int_0^1 [u(S(t, \hat{\imath})) - u(\tilde{S}(t, \hat{\imath}))]f_i^{\hat{\imath}}(t)dt \leq 0. \tag{11}$$

This increases social welfare, provided that truthfulness is maintained. Truthfulness is maintained because (11) 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 $s^{ce} \geq s^a$ because u is concave and decreasing. Since $W(\cdot, g)$ is also concave, $W(s^a, g) \geq \int_0^1 W(S(t, \hat{g}), g)f_g^{\hat{g}}(t)dt$. Since $W(\cdot, g)$ is single-peaked at $\hat{s}$, it decreases in s for $s \geq \hat{s}$, 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 $s^{ce} \geq s^a$ and $W(\cdot, g)$ is increasing up to $\hat{s}$, we have $W(s^{ce}, g) \geq W(s^a, g) \geq \int_0^1 W(S(t, \hat{g}), g)f_g^{\hat{g}}(t)d$, 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.

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

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

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, that is, (1) continues to hold when $S(\cdot,\hat{g})$ is replaced with s^g.

If (1) when $S(\cdot,\hat{g})$ is replaced with s^g holds strictly, 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 also increases welfare since it increases the utility of an innocent defendant. It also maintains truthfulness of the innocent defendant, because the guilty defendant's indifference 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.

B Proof of Theorem 2

Consider an available mechanism (F,S). Similarly to the proof of Theorem 1, we will modify the mechanism by changing the sentence function in a way that maintains truthfulness and increases ex-ante welfare.

As in the proof of Theorem 1, 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

$$\max_{\hat{u}(\cdot)\in(\Delta([u(\bar{s}),u(0)]))^T} \int_0^1 E[\hat{W}(\hat{u}(t))]f_g^{\hat{g}}(t)dt \tag{12}$$

subject to

$$\int_0^1 E[\hat{u}(t)]f_g^{\hat{g}}(t)dt = U^g, \tag{13}$$

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 (12) subject to (13), it is useful to consider a simpler optimization problem:

$$\max_{\dot{u}\in\Delta([u(\bar{s}),u(0)])} E[\hat{W}(\dot{u})] \tag{14}$$

subject to

$$E[\dot{u}] = U^g. \tag{15}$$

Consider a stochastic process $\hat{u} : T \to \Delta[u(\bar{s}), u(0)]$ whose sample paths are Lebesgue measurable and that satisfies (13). 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 (15) and

$$\int_0^1 E[\hat{W}(\hat{u}(t))] f_g^{\hat{g}}(t)dt = E[\hat{W}(\dot{u})]. \tag{16}$$

Therefore, $\hat{u}$ is a solution of (12) subject to (13) if and only if $\dot{u}$ is a solution of (14) subject to (15).

We now solve for (14) subject to (15). 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 (14) subject to (15):[55] 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 (16), 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 $u^{-1}(\underline{U})$ and $u^{-1}(\overline{U})$, 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 α and $1-\alpha$ such that $\alpha\underline{U} + (1-\alpha)\overline{U} = U^g$. In this case, again by (16), the constant stochastic sentence function $\tilde{S}(\cdot, \hat{g})$ (which is independent of the signal) that for every signal t assigns sentence $u^{-1}(\underline{U})$ with probability α and sentence $u^{-1}(\overline{U})$ 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 $[0, \hat{s}]$.[56] 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 $\tilde{S}(\cdot, \hat{\imath})$, that is,

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

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

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

[55] 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).

[56] 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 E proves the genericity claim in part (ii).

C Proof of Theorem 3

Consider an available mechanism (F,S) and construct the same improving available mechanism $(F,\tilde{S})$ as in the proof of Theorem 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 the proof of Theorem 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 $\hat{s}$. Figure 1 illustrates this.

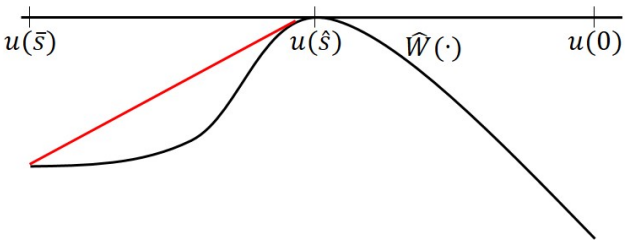

Figure 1: Lottery over sentences.

D Proof of uniqueness in Theorem 1

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.[57] Since u is strictly 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 $T_1 \cup T_2$. [58] Let $\gamma(t) = f_i^{\hat{\imath}}(t)/f_g^{\hat{\imath}}(t)$. By strict MLRP, γ 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 γ' 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.[59] 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 (ii) 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.

E Proof of generic uniqueness in Theorems 2 and 3

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

[57] 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.

[58] The fact that $H(0) = 0$ is simply a restatement of(10). 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.

[59] This is immediate if $W(\cdot,i) = u(s)$. The general case is explained in Appendix G. See Equation (19).

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.[60]

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[61] 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[62] 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 $A(w) = \{a \in \mathbb{R} : w_a$ 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,[63] 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.

---

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

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

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

[63] 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 $x \in (\underline{x}, \bar{x})$.

F Unbounded Sentences and Relation to Crémer-McLean

We revisit the assumption that the sentence is bounded and relate this restriction to the possibility of achieving the first best outcome and to Crémer and McLean's (1988) results. From an interim perspective, it would be socially optimal to give a null sentence to an innocent defendant and the sentence $\hat{s}$ to a guilty defendant. Under what conditions can this allocation be approximately implemented? Suppose that the sentence space is unbounded and $u(s) \to -\infty$ as $s \to +\infty$. This assumption alone does not guarantee that the first best can be implemented. Indeed, suppose that all the possible signals have a likelihood ratio that is bounded above by some constant $\bar{\ell} < \infty$ (recall that we did not assume that the likelihood ratio is unbounded). Then, if arbitrarily harsh punishments are used in the sentencing scheme for the innocent, they must be used with sufficiently low probability to guarantee that an innocent defendant's expected utility remains close to that in the first best. But with a bounded likelihood ratio, this implies that a guilty defendant's expected utility from these punishments is no worse than $\bar{\ell}$ times that of an innocent defendant. This limits the ability to screen the defendant. Thus, to guarantee that the first best is achievable, one generally needs that both the punishment and the likelihood ratio be unbounded.

If it were possible to reward the defendant, however, the first best would be achievable even when the likelihood ratio is bounded. More precisely, suppose that the sentence space is extended to $\mathbb{R}$ and that u decreases from $+\infty$ to $-\infty$, so that arbitrarily high rewards and punishments are both available to the social planner. Then, by using arbitrarily high rewards for low likelihood ratios and arbitrarily high punishments for high likelihood ratios, one can construct a sentencing scheme that achieves any given level of utility for an innocent defendant while providing an arbitrarily negative utility for a guilty defendant. The logic of the argument is very similar to Crémer and McLean (1988) and the details are omitted.

G Welfare vs. 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 increasing transformation $\phi : \mathbb{R}_- \to \mathbb{R}_-$ such that $W(s, i) = \phi(u(s))$. The weak convexity of ϕ 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, 2, and 3, we focus for simplicity on Theorem 3.

Proposition 3 *Suppose that ϕ is increasing and convex and that the assumptions of Theorem 3 are otherwise unchanged. Then, there exists a welfare-maximizing optimal mechanism that satisfies all the conclusions of Theorem 3.*

Proof. The construction is identical to the proof of Theorem 3. The welfare function $W(s, i)$ enters only the first step of the proof of Theorem 3, 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

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

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

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

It follows from the indifference equation (9) and the argument following it that

$$\int_0^1 u(\tilde{S}(t,\hat{\imath}))f_i^{\hat{\imath}}(t)dt \geq \int_0^1 u(S(t,\hat{\imath}))f_i^{\hat{\imath}}(t)dt \tag{18}$$

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

$$-F_i^{\hat{\imath}}([\bar{t},1]) \geq \int_0^1 u(S(t,\hat{\imath}))f_i^{\hat{\imath}}(t)dt \tag{19}$$

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[64]

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

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 (19) then yields (17). ■

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.

[64]The inequality is a direct application of the definition of ϕ'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 ϕ and together with $u(\tilde{s}) \in [-1,0]$, $u(\bar{s}) = -1$, $u(0) = 0$, $\phi(0) = 0$, and $\phi(-1) = -M$, imply $\phi(u(\tilde{s})) = \phi((-u(\tilde{s})(-1) + (1-(-u(\tilde{s})))(0))) \leq (-u(\tilde{s}))\phi(-1) + (1-(-u(\tilde{s})))\phi(0) = Mu(\tilde{s})$. Integrating over $\tilde{s}$ then yields $\phi(u(S(t,\hat{\imath}))) \leq M\int_{[0,\bar{s}]} u(\tilde{s})g(\tilde{s})d\tilde{s} = Mu(S(t,\hat{\imath}))$.

Boylan, R. (2012) "The Effect of Punishment Severity on Plea Bargaining," *Journal of Law and Economics*, Vol. 55, pp. 565–591.

Bray, S. (2005) "Not Proven: Introducing a Third Verdict," *University of Chicago Law Review*, Vol. 72, pp.1299–1329.

Crémer, J, McLean, R. (1988) "Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions," *Econometrica*, Vol. 56, pp. 1247–1257.

Da Silveira, B. (2017) Bargaining with Asymmetric Information: An Empirical Study of Plea Negotiations," *Econometrica*, Vol. 85, pp. 419–452.

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.

Doval, L., Skreta, V. (2020a) "Mechanism Design with Limited Commitment," *Working Paper*.

Doval, L., Skreta, V. (2020b) "Optimal Mechanism for the Sale of a Durable Good," *Working Paper*.

Elder, H. (1989) "Trials and Settlement in the Criminal Courts: an Empirical Analysis of Dispositions and Sentencing," *Journal of Legal Studies*, Vol. 18, pp. 191–208.

Fisher, T. (2011) "Conviction Without Conviction," *Minnesota Law Review*, Vol. 96, pp. 833–885.

Gerardi, D. and Yariv, L. (2008) "Costly Expertise," *American Economic Review*, Vol. 98, pp. 187–193.

Gershkov, A., Szentes, B. (2009) "Optimal Voting Schemes with Costly Information Acquisition," *Journal of Economic Theory*, Vol. 144, pp. 36–68.

Gross, S., O'Brien, B., Hu, C., and E. Kennedy (2014) "Rate of False Conviction of Criminal Defendants who are Sentenced to Death," *Proceedings of the National Academy of Sciences*, Vol. 111, pp. 7230–7235.

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.

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

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.

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.

Liebman, J.S., Fagan, J., West, V. and Lloyd, J. (1999) "Capital Attrition: Error Rates in Capital Cases, 1973-1995," *Texas Law Review*, Vol. 78, pp. 1839–1865.

Myerson, R. (1979) "Incentive Compatibility and the Bargaining Problem," *Econometrica*, Vol. 47, pp. 61–73.

Nelson, W. (2013) "Political Decision Making by Informed Juries." *William and Mary Law Review*, Vol. 55, pp. 1149–1166.

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

Shi, X. (2012) "Optimal Auctions with Information Acquisition," *Games and Economic Behavior*, Vol. 74, pp. 666–686.

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.

Skreta, V. (2006) "Sequentially Optimal Mechanisms," *Review of Economic Studies*, Vol. 73, pp. 1085–1111.

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.