Judicial Mechanism Design

Ron Siegel and Bruno Strulovici*

August 2019

Abstract

This paper studies the design of welfare-maximizing criminal judicial processes. We identify properties of the generically unique optimal processes 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 extreme, binary verdicts in conjunction with plea bargains to the use of jury instructions and an adversarial system, all of which emerge as the result of informational, commitment, and incentive arguments.

1 Introduction

Many criminal justice systems have features that may seem puzzling from an economic perspective. It is not clear, for example, why criminal trials usually have only two verdicts, “guilty” and “not guilty,” and why the “not guilty” verdict carries no punishment. After all, 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 also suggests that some punishment following an acquittal may sometimes be optimal. Moreover, while

---

¹We thank Daron Acemoglu, Robert Burns, Andy Daughety, Eddie Dekel, Richard Holden, Ian Jewitt, Louis Kaplow, Fuhito Kojima, Adi Leibovitz, Paul Milgrom, Wojciech Olszewski, Debraj Ray, Jennifer Reinganum, Marciano Siniscalchi, Kathy Spier, Jean Tirole, Abe Wickelgren, and Leeat Yariv for their comments. We thank Bentley MacLeod for his discussion at the NBER Summer Institute Law and Economics Workshop (2018). We thank participants at various seminars for their comments. David Rodina provided excellent research assistance. Strulovici acknowledges financial support from an NSF CAREER Award (Grant No. 1151410) and a fellowship from the Alfred P. Sloan Foundation. 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.

¹A recent study by Gross et al. (2014) of 7,482 death row convictions from 1973 to 2004 in the United States estimates that 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. In an influential review of all capital sentences given in the United States between 1973 and 1995, Liebman, Fagan, and West (2000) find that 68% of capital sentences
American sentencing guidelines that determine the length of a sentence following a conviction refer to the circumstances of the crime and the defendant’s criminal history, they do not refer to the strength of evidence as a relevant factor. This is puzzling since the length of the sentence is a continuous variable that can easily be adjusted.

Binary verdicts, no punishment following an acquittal, and other features of criminal justice systems, such as plea bargaining, are taken as given by much of the law and economics literature. Early work, pioneered by Becker (1966) and Stigler (1970), employed a (partial or general) equilibrium framework to study various aspects of law enforcement and criminal justice. More recent work, including Grossman and Katz (1983), Reinganum (1988), Baker and Mezzetti (2001), Kaplow (2011, 2017), and Daughety and Reinganum (2015a,b), takes a game theoretic approach. This literature has led to many important insights, but does not provide an economic justification for the aforementioned features of many criminal justice systems, because these features are assumed as part of the models.

This paper takes a different approach. Instead of studying a particular game, we take a step back and focus on the goals and constraints common to many existing criminal justice systems. We postulate a social planner who wants to deter potential criminals, suitably punish guilty defendants, and not punish innocent ones. We focus on a particular crime and a single defendant suspected of committing the crime. The main constraint faced by the social planner is that the defendant privately knows whether he is guilty. This constraint is common to all criminal justice systems; after all, if it was known whether the defendant committed the crime, there would be no need for a lengthy judicial process—the appropriate punishment could always be meted out. Faced with this constraint, the planner designs a system to generate evidence regarding the defendant’s guilt and determine the sentence.

In reality, defendants may have additional private information: they may be informed about the strength of evidence that may be uncovered, and may also differ in terms of their disutility from various sentencing schemes and their risk attitudes. We rule this out in our model to focus the analysis on the basic challenge every criminal justice system must contend with: the defendant knows whether he is guilty or not, and the judicial process must determine the defendant’s guilt and the appropriate punishment. We refer to this system as a judicial mechanism (or simply a mechanism), which may involve

---

2See 18 U.S.C. §3553 for a list of mitigating and aggravating factors recognized by the U.S. code. Mitigating factors include impaired capacity and duress. Aggravating factors include vulnerability of the victim, pecuniary gain, and previous offenses. None of the factors listed in the code reflect the strength of evidence found in the case.
many agents (including police, judge, jury, prosecutors, etc.) with differing abilities and information.\textsuperscript{3} The mechanism ends with a sentence, which can depend on the generated evidence. The process may be multi staged, with evidence collected at various points, the case may be “dropped,” there may be any number of “verdicts,” and the sentence could depend on the “strength of the case.”

Our goal is to identify properties of optimal judicial mechanisms and compare them with existing criminal justice systems. A central challenge is to identify a tractable set of mechanisms over which to optimize. The difficulty is that the evidence generation processes available to the planner may depend on the forensic technology, which we do not want to restrict, and the incentives of the various agents (including the defendant), which may be affected by the possible sentences. We deal with this by imposing only the minimal constraint described above, namely, that the defendant privately knows whether he is guilty. More precisely, we assume that given an available judicial process, changing the sentence (as a function of the generated evidence) in a way that does not change the defendant’s behavior results in another available judicial process with the same evidence generation process. This assumption means that the incentives of all agents other than the defendant are not affected by the sentence function, that is, their incentives are not a constraint the planner has to contend with. For example, this assumption says that given the same evidence jurors will reach the same verdict regardless of the severity of the punishment associated with the verdict.

While strong, the assumption is not completely unfounded. First, if the incentives of certain agents other than the defendant are a major concern, the planner may be able to replace them with other agents whose interests are more aligned with those of the planner, or provide more detailed and forceful instructions to agents such as jurors.\textsuperscript{4} Second, while the defendant’s private information poses a constraint common to all judicial systems, the other agents’ constraints could vary across systems (for example, some systems do not have juries). We are interested in properties of optimal mechanisms that may be common to various judicial systems. Third, to the extent that additional constraints arise in reality, we are likely considering a superset of the mechanisms available in practice. Since the set of mechanisms we consider is likely a superset of those available in practice, to the extent that features of

\textsuperscript{3}For mechanism design approaches to tort law, see Spier (1994), Klement and Neeman (2005), and Demougin and Fluet (2006), who analyze settlement and fee-shifting rules between plaintiffs and defendants.

\textsuperscript{4}The benefit of separating the incentives to find incriminating and exculpatory evidence in the American adversarial system has been noted in Dewatripont and Tirole (1999) and is consistent with early legal doctrine. In particular, Lord Eldon, Lord Chancellor in the United Kingdom, said in 1822 that "truth is best discovered by powerful statement on both sides of the quest." See also Shin (1998), Daughety and Reinganum (2000) and Deffains and Demougin (2008). The U.S system also features the prevailing judicial practice of keeping jurors uninformed of the precise consequences of a conviction. See, e.g., (Sauer (1995)). 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.
optimal mechanisms mirror those of existing criminal justice systems, such features may indicate that any additional constraints are not binding in practice.

We find that optimal mechanisms display striking similarities to existing judicial systems but also important differences. Optimal mechanisms can be thought of as implementing the following procedure. The defendant is first offered a plea bargain.\(^5\) If he accepts it, the process ends. If he rejects the plea bargain, a trial takes place, which ends with one of two verdicts, “guilty” or “not guilty.” The defendant is found guilty if the evidence regarding his guilt is sufficiently strong. A guilty verdict carries a sentence that is more severe than the one associated with a plea bargain; a “not guilty” verdict carries no punishment. Evidence brought to bear during the trial is used only to determine the verdict and not the sentence associated with the verdict. All these realistic features arise as part of any optimal mechanism and are not assumed features of the mechanisms.\(^6\)

The intuition for the structure of optimal mechanisms is that the plea bargain and conviction threshold are optimally set to fully screen defendants, so that guilty defendants take the plea bargain and innocent defendants go to trial. Once defendants are screened, evidence is no longer needed to determine their actual guilt—it is used solely to provide the proper incentives for screening. That perfectly screening mechanisms are available relies on our assumption that changing the sentence function does not affect the availability of the mechanism, provided that the defendant’s incentives are maintained. Indeed, without this assumption the various agents participating in the mechanism may decide to acquit defendants who reject the plea, since such defendants are correctly expected to be innocent. However, guilty defendants would then reject the plea as well, rendering the mechanism ineffective. The purely incentivizing role that evidence plays in the optimal mechanisms is one important difference from actual judicial systems, in which evidence also appears to be used to determine defendants’ actual guilt. The role that evidence plays in the optimal mechanisms highlights the fact that in practice too evidence may be used to incentivize guilty defendants to plead guilty. The discrepancy between the role that evidence plays in the optimal mechanisms and its apparent role in reality can be mitigated as follows. In the optimal mechanisms guilty defendants are indifferent between taking the plea and going to trial (and they take the plea). If a small fraction of them instead goes to trial, then the mechanism is nearly optimal and evidence regains its role in determining actual guilt in addition to its incentivizing role.

Another important difference between the optimal mechanisms and actual criminal trials is how the conviction threshold and the severity of the sentence following a conviction vary across crimes and

\(^5\)The use of a plea bargain is not due to cost savings: plea bargains are part of optimal mechanisms even when producing evidence is costless.

\(^6\)For example, if the class of mechanisms is restricted to prohibit plea bargains, it becomes optimal to use many verdicts, with the sentence increasing in the strength of the evidence that the defendant is guilty.
circumstances. In the United States, for example, the conviction criterion is “beyond a reasonable doubt” (BARD) for all criminal cases, but the punishment increases with the severity of the crime. In the optimal mechanisms the conviction threshold can optimally vary across crimes, and a conviction carries the maximal allowable punishment for the crime. This finding suggests that allowing the conviction criterion to vary could improve existing criminal trials.\(^7\)

Finally, the optimal plea bargain may sometimes involve a random sentence. This can happen, for example, when society and the defendant are risk averse, society is sufficiently less risk averse than the defendant, and deterrence is an important consideration. Such randomness is consistent with real-world plea bargains in which the judge has discretion over the sentence after the defendant irrevocably accepts the plea bargain,\(^8\) and with the institution of parole, which reduces the sentence and is stochastic at the time of sentencing. Random plea bargains are valuable owing to their deterrence effect. On the other hand, if the planner is not concerned with deterrence (which may be the case with crimes of passion), and society and the defendant are risk averse, then the plea sentence is optimally deterministic.

2 Judicial Mechanisms

Suppose that a crime has been committed and a suspect is arrested. We refer to the suspect as the defendant. The defendant is guilty with prior probability \(\lambda \in (0, 1)\), innocent with probability \(1 - \lambda\), and privately informed about his guilt \(\theta \in \Theta = \{i, g\}\).\(^9\) A judicial process begins that produces evidence regarding the defendant’s guilt and ends with the defendant’s sentence. The process may be multi-staged and involve other agents who may have private information (police, prosecutors, attorneys, jurors, etc.). The distribution of generated evidence may depend on the defendant’s type (for example, an innocent defendant is less likely to have been at the crime scene and therefore less likely to be identified by eye witnesses) and on the defendant’s actions during the judicial process (for example, admitting guilt may lead to a different evidence collection process than claiming innocence). The sentence may depend on the generated evidence and the defendant’s actions during the judicial process.

The defendant’s optimal behavior in the process induces a single-player truthful (or incentive com-

\(^7\)This finding is consistent with Kaplow (2011), who argues that the conviction threshold should be endogenously determined.

\(^8\)A recent, widely publicized example of this case concerns Jared Fogle, a former Subway spokesman who accepted a plea bargain with 5 years in prison and subsequently received a sentence that far exceeded the one outlined in the plea bargain. “Jared Fogle, Former Subway Pitchman, Gets 15-Year Prison Term,” New York Times, November 19, 2015.

\(^9\)This also captures crimes for which the issue is not whether it was the defendant or someone else who committed the crime, but rather whether a crime was committed at all, and the defendant privately knows this. For example, whether a homicide was a murder or committed in self defense. We thank Abe Wickelgren for this interpretation.
compatible) direct judicial mechanism or, simply, a mechanism, which allows us to focus on the defendant’s private information regarding his guilt. The player is the defendant and his two possible actions are to report that he is innocent or to report that he is guilty. For notational clarity, we use hats to denote the reports: ı and ı. By a logic similar to that of the revelation principle (Myerson 1979), the defendant reports his guilt truthfully. 10

The mechanism is a pair \((F, S)\), where \(F = (F^ı, F^ı, F^ı, F^ı)\) is a tuple of signal distributions and \(S\) is a sentencing scheme. The distribution \(F^ı\) over signals \(t \in [0, 1]\) corresponds to the distribution of evidence generated when the defendant of type \(θ\) reports type \(ı\). We assume that distributions \(F^ı\) and \(F^ı\) have positive densities \(f^ı(t)\) and \(f^ı(t)\) on the support \(T = [0, 1]\). We order the signals according to their likelihood ratio following a report of ı, and assume the strict monotone likelihood ratio property (MLRP): the density ratio \(f^ı(t)/f^ı(t)\) is strictly increasing in \(t\). 11

The sentencing scheme \(S\) is a measurable mapping from the defendant’s report \(ı\) and realized signal \(t\) to a (possibly degenerate) sentence lottery \(S(t, ı) \in Δ([0, \bar{s}])\), where \(s\) is the highest allowable sentence for the crime and \(Δ([0, \bar{s}])\) is the set of lotteries over allowable sentences. Our analysis focuses on a particular crime, so \(s\) can vary across crimes. 12 The bound on punishment is taken as exogenous and reflects moral considerations that are beyond the scope of this paper. In particular, the Eighth Amendment of the United States Constitution bans “cruel and unusual” punishments and “excessive fines” (under its interpretation by the Supreme Court, see United States v. Bajakajian (1998)) and may be viewed as capturing ex post welfare considerations. 13 As is standard in mechanism design analysis, we assume that the designer can commit to any mapping from \(S\). Appendix H provides a rigorous micro-foundation for the use of direct judicial mechanisms by describing a rich class of multi-player

---

10 If it is profitable for the defendant to misreport his type, then the corresponding deviation is profitable for the defendant in the original judicial process.

11 The assumed existence of strictly positive densities rules out atoms, i.e., a positive measure of signals with the same likelihood ratio, and implies that the set of achievable likelihood ratios forms an interval. This assumption is used in the construction of welfare-improving schemes that keep the defendant’s expected utility unchanged. If there are atoms in the distribution of the likelihood ratio, any such atom can be decomposed into an interval of signals corresponding to a constant likelihood ratio, with the signal distribution having a density over the interval. The constructions used in the proofs of Theorem 1 and 3 go through, but if the signal threshold arising in the construction lies in such an interval, the optimal scheme will involve randomization over two extreme sentences when the signal observed has the corresponding likelihood ratio.

12 The same is true of the welfare function and prior \(λ\) introduced later in this section.

13 Indeed, another way of imposing the bound on punishment is to assume that the ex-post welfare functions introduced later in this section become infinitely negative beyond the threshold \(s\). The compactness of the sentence space is used to guarantee the existence of an optimal mechanism. We further discuss the restriction on the maximal possible sentence in Appendix F.
games in extensive form with incomplete information that generate evidence and a sentence and shows how to reduce them to single-agent judicial mechanisms. Appendix H also shows that taking the signal to be one dimensional, a common assumption in the law and economics literature, is without loss in our setting, even though in reality evidence may be multi-dimensional.

Our objective is to study the social welfare of different truthful mechanisms. We consider two notions of welfare, interim and ex ante. The interim notion captures social welfare after the crime has been committed, and is similar to the minimization of Type I and Type II errors (convicting an innocent defendant and acquitting a guilty ones) when facing a binary-verdict decision.\textsuperscript{14} Ex-ante welfare also takes into account the number of crimes committed and the deterrent effect of the mechanism.

From an interim perspective, once the crime has been committed, society wishes to punish guilty individuals and avoid punishing innocent ones, and takes into account the cost of producing evidence. We denote by $W(s, \theta)$ the social welfare of imposing a sentence $s$ on a defendant of type $\theta$.\textsuperscript{15} Any monetary cost of imposing the sentence, such as the cost of incarceration, is included in $W$. We assume that the welfare functions are non-positive (the highest possible welfare is 0), that $W(s, i)$ strictly decreases in $s$, and that $W(s, g)$ is continuous.\textsuperscript{16} The expected welfare of imposing a sentence $s$ on a defendant who is guilty with probability $\lambda$ is $\lambda W(s, g) + (1 - \lambda) W(s, i)$. Thus, the more likely the defendant is to be guilty, the more important it is to adequately punish him if he is in fact guilty; the less likely the defendant is to be guilty, the more important it is to avoid punishing him if he is in fact innocent. With a slight abuse of notation we denote by $W(\tilde{s}, \theta)$ the expected welfare of imposing a (possibly) random sentence $\tilde{s}$ on a defendant of type $\theta$.\textsuperscript{17} We denote by $C(F^g_\theta) \geq 0$ the expected cost of the judicial process associated with the signal distribution $F^g_\theta$, and assume for simplicity that the cost is additively separable from $W$.\textsuperscript{18} Thus, given a mechanism $(F, S)$ in which the defendant reports his guilt truthfully, the resulting interim welfare is

$$\lambda \left( \int_0^1 W(S(t, \hat{g}), g) f^g_{\hat{g}}(t) dt \right) - C \left( F^g_{\hat{g}} \right) + (1 - \lambda) \left( \int_0^1 W(S(t, \hat{i}), i) f^i_{\hat{i}}(t) dt \right) - C \left( F^i_{\hat{i}} \right). \quad (1)$$

\textsuperscript{14} In our setting, the decision need not be binary, since the designer can a priori choose any sentence on a continuum. The welfare functions used in our analysis and other works in law and economics may be viewed as a generalization of Type I and Type II errors to this more general environment.

\textsuperscript{15} These functions can be thought of as capturing ex-post welfare, since the crime has been committed and the defendant’s type enters as an argument.

\textsuperscript{16} Continuity is assumed for expositional simplicity. 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.

\textsuperscript{17} Formally, if $\tilde{s}$ represents a probability distribution over sentences in $[0, \bar{s}]$, then $W(\tilde{s}, \theta) = \int W(s, \theta) d\tilde{s}(s)$.

\textsuperscript{18} In general, the cost could depend on the entire mechanism. Here we omit this dependence for simplicity and refer the reader to Appendix A for a general treatment.
From an ex-ante perspective, society also wishes to deter crime. We focus on a specific crime, which entails a particular harm, \( h \), 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 \( \pi_g \) as exogenous.  

We assume that at most one individual is prosecuted for the crime.  

Letting \( u(s) \) denote the defendant’s utility from sentence \( s \), we assume that the social preferences over sentences, conditional on facing an innocent defendant, agree with those of the defendant. Thus, \( W(\cdot, i) = u(\cdot) \), where \( u(\cdot) \) is strictly decreasing, continuous, and normalized such that \( u(0) = 0 \). All of our results continue to hold if instead \( W(s, i) \) is an increasing, convex transformation of \( u \), which means that the social preferences are aligned with the defendant but exhibit weakly less risk aversion (see Appendix G). With a slight abuse of notation, we denote by \( u(\tilde{s}) \) the expected utility of the individual from the (possibly) random sentence \( \tilde{s} \).

Thus, given a mechanism \((F, S)\) in which the defendant reports his guilt truthfully, an individual commits the crime if

\[
b + \pi_g \left( \int_0^1 u(S(t, \hat{g})) f_\hat{g}(t) dt \right) > 0.
\]

In general, the mechanism faced by the defendant could depend on the amount of evidence gathered at the time of arrest. If this evidence is stochastic from the viewpoint of someone committing a crime, the second term of \((2)\) 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 general environment because the welfare-improving mechanisms constructed in Section 3 keep the expected utility of a guilty defendant unchanged and could be performed to each mechanism: they would improve welfare without affecting deterrence. For expositional simplicity, we focus here on a single belief at the time of arrest, which means that deterrence is evaluated by a single mechanism. Alternatively, we could interpret the analysis below as obeying the constraint that the same mechanism has to be applied regardless of the information acquired about the defendant at the time of arrest.

---

19 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.

20 This allows us to abstract from interdependencies between multiple defendants, an issue that is tangential to the focus of this paper. See Silva (2016) for an analysis of this issue.

21 This assumption appears in Grossman and Katz’ (1983) analysis of plea bargaining.

22 For the same reason, our results continued to hold if i) 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, since we would again perform the improvement for each realization of this evidence. Likewise, the results continue to hold if the mechanisms and the welfare and utility functions are allowed to depend on some observable characteristics of the defendant, such as the defendant’s age or social condition, as it the improvements would then be made separately for each observable characteristic of the defendant, without affecting deterrence while improving interim welfare.
The benefit 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

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

(3)

In a large society, the probability that any particular innocent individual will be prosecuted for that particular crime is infinitesimal. Moreover, a guilty defendant could also be wrongfully prosecuted for another crime. Therefore, the risk of being wrongfully convicted does not affect the incentive to commit crime, since it may equally arise whether an individual committed a crime or not, and is omitted from the incentive equation (2).\(^{23}\)

We denote this probability by $\pi_i > 0$. \(^{24}\) The ex-ante social welfare given a mechanism $(F,S)$ in which the defendant reports his guilt truthfully is

$$H((F,S)) \left[ \pi_g \left( \int_0^1 W(S(t,\hat{g}),g) f^\hat{g}_g(t) dt \right) - C(F^\hat{g}_g) \right] + \pi_i \left( \int_0^1 W(S(t,\hat{i}),i) f^\hat{i}_g(t) dt \right) - C(F^\hat{i}_i) - h].$$

We allow for $\pi_i + \pi_g < 1$, so it is possible that for some crimes no individual is prosecuted. For expositional simplicity, this formulation of welfare does not include the individual’s benefit from committing the crime. This benefit can be considered explicitly without affecting any of the results.\(^{25}\)

Equation (4) includes the mechanism’s deterrent effect. To compare this to our formulation of interim welfare, notice that by the time an individual is prosecuted the crime has already been committed, so from an interim perspective the social harm $h$ from the crime is “sunk.” The individual’s probability of guilt is then $\lambda = \pi_g / (\pi_g + \pi_i)$, which gives (1).

### 3 Optimal judicial mechanism

The standard approach to optimal mechanism design is to optimize over all truthful direct mechanisms. But in our setting it is not clear which truthful mechanisms $(F,S)$ are available. This is because the

\(^{23}\)Our results would go through 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 thus have the additional benefit of increasing deterrence by increasing the utility differential between an innocent defendant and a guilty one.

\(^{24}\)As with $\pi_g$, for expositional simplicity we will take $\pi_i$ to be exogenous.

\(^{25}\)Most of our analysis proceeds by modifying sentencing schemes without affecting the expected utility of a guilty defendant. Under such modifications, the set of defendants committing the crime, and their benefit from doing so, is unchanged.
signal distributions $F_{\theta}^\hat{\theta}$ regarding the defendant’s guilt are determined by the underlying evidence-gathering technology and the actions of various actors in the judicial system (the strategies of the players in the underlying multi-player games). These actors may respond to different incentives, which means that the signals available to a mechanism designer could a priori depend on the sentencing scheme, making the set of mechanisms over which the designer can optimize is potentially very complex to describe.\(^{26}\)

We now introduce our main assumption, which allows us to overcome this difficulty without restricting the underlying evidence-gathering technology. The assumption states that given an available (truthful) mechanism $(F,S)$, any mechanism $(F,\tilde{S})$ that results from changing the sentencing scheme to $\tilde{S}$ is also available, provided that it is truthful. This assumption is in line with our focus on the defendant’s private information regarding his guilt.

To formalize the assumption, denote by $F$ the set of all distribution profiles $F = (F_i^i,F_i^g,F_{\theta}^i,F_{\theta}^g)$ with positive densities on $T = [0,1]$ such that the density ratio $f_{\theta}^g(t)/f_i^i(t)$ is strictly increasing in $t$ (as specified in Section 2) and by $S$ the set of all measurable functions from $T = [0,1]$ (signals) and $\{i,\hat{\theta}\}$ (reported type) to $\Delta([0,\bar{s}])$ (lotteries over sentences). We say that a mechanism $(F,S) \in F \times S$ is truthful if both types find it optimal to report truthfully, that is,

$$\int_0^1 u(S(t,\hat{\theta}))f_{\theta}^g(t)dt \geq \int_0^1 u(S(t,i))f_i^i(t)dt$$

and

$$\int_0^1 u(S(t,\hat{\theta}))f_i^i(t)dt \geq \int_0^1 u(S(t,\hat{\theta}))f_{\theta}^g(t)dt.$$ \(^{(5)}\)

We denote the set of truthful mechanisms by $M^{IC} \subseteq F \times S$. We denote by $M \subseteq M^{IC}$ the set of available mechanisms, over which we will optimize. We maintain the following assumption throughout our analysis.

**Assumption 1** If $(F,S) \in M$, $\tilde{S} \in S$, and $(F,\tilde{S}) \in M^{IC}$, then $(F,\tilde{S}) \in M$.

Assumption 1 formalizes the idea that, provided that the defendant’s behavior is unchanged, modifying the sentence function does not affect the unmodeled players’ (prosecutor, jurors, etc.) behavior in such a way as to prevent the generation of the signal distributions $F$.\(^{27}\)

---

\(^{26}\)Another difference from the standard setting is that the defendant’s type does not determine the defendant’s preferences over outcome, but rather the distributions of evidence different actions will generate. This is similar to the literature on hard information (for a recent contribution see, for example, Ben-Porath, Dekel, and Lipman 2014), but unlike many papers in that literature, in the present setting there is no a priori obvious set of mechanisms over which to optimize.

\(^{27}\)As discussed in the introduction, additional realistic constraints will likely make the set of available mechanisms smaller.
assumption is that if the incentives of these players are affected, the planner may be able to replace them with other players or provide them with better instructions so as to generate distributions $F$. As will become clear once properties of the optimal mechanisms are identified, less restrictive versions of Assumption 1 suffice for our results. Versions of Assumption 1 appear (explicitly or implicitly) in many law and economics papers.\textsuperscript{28} As discussed in the Introduction, Assumption 1 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. Appendix H provides a novel micro-foundation for the assumption. Section 4 interprets Assumption 1 in light of existing features of the US criminal justice system.

We now identify properties of the optimal truthful mechanisms among the available ones, first for interim welfare and then for ex-ante welfare. We compare and discuss the features of these mechanisms in Section 4.

### 3.1 Interim welfare

A judicial mechanism $(F, S)$ is interim optimal if, given the prior probability $\lambda$ that the defendant is guilty, the mechanism maximizes interim welfare (1) among all available mechanisms in $\mathcal{M}$. Considering interim-optimal mechanisms first allows us to disentangle deterrence from other welfare considerations and makes the arguments of 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.”\textsuperscript{29} We also assume that the defendant and society when facing a guilty defendant are risk averse (their risk attitudes are otherwise arbitrary and, in particular, need not be related).

**Assumption 2** The function $W(\cdot, g)$ is single-peaked with peak $\hat{s} \leq \bar{s}$, where $\bar{s}$ is the maximal allowable sentence, the functions $W(\cdot, g)$ and $u(\cdot)$ are concave, and at least one of them is strictly concave.\textsuperscript{30}

\textsuperscript{28}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.

\textsuperscript{29}See 18 U.S.C §3553. These guidelines also state that another goal is “to protect the public from further crimes of the defendant.” This incapacitation reasonably 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.

\textsuperscript{30}Strict concavity of $W(\cdot, g)$ means that $W(\mu s + (1 - \mu)s', g) > \mu W(s, g) + (1 - \mu)W(s', g)$ for all $\mu$ in $(0, 1)$ and $s \neq s'$. 
Theorem 1  If Assumption 2 holds, then any interim optimal mechanism has the following properties.\textsuperscript{31}

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

(ii) The guilty defendant’s sentence is constant.

(iii) The guilty defendant is indifferent between reporting truthfully and misreporting, that is (5) holds as an equality.

Moreover, any mechanism that fails to have the above properties can be improved for all priors $\lambda$ by a single mechanism with these properties.

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, where 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.\textsuperscript{32}

We note that the signal is not used by the mechanism to determine the sentence if the defendant pleads guilty, even if signal distributions $F^g_i$ and $F^g_g$ are informative about the defendant’s guilt. This highlights that what underlies the optimality of extreme sentences here is unrelated to justifications for extreme sentences based on law enforcement costs and deterrence (Becker (1968)). It is the screening value of pleas, emphasized by Grossman and Katz (1983), that makes pleas optimal. While Grossman and Katz (1983) noted this benefit of pleas, they did not show their optimality among other mechanisms: they studied the optimal two-verdict system with a plea sentence, whereas we show that such a system is in fact globally optimal, at least from an interim perspective (and under Assumption 2). As shown below, this result generally fails without Assumption 2 or when deterrence is taken into account.\textsuperscript{33}

The last statement in Theorem 1 shows that the argument underlying the result is non-Bayesian: starting from any mechanism, there is another mechanism with the properties stated in the theorem that improves upon the initial mechanism state by state (i.e., conditional on each of the defendant’s

\textsuperscript{31}All statements are required to hold except on a set of measure zero. For instance, the optimal sentence for an innocent defendant could take arbitrary values over a set of signals that has zero Lebesgue measure. Since these signals arise with probability zero, such a change is irrelevant. The same observation holds for Theorem 3.

\textsuperscript{32}In fact, the interim-optimal sentences following an acquittal are strictly positive when pleas are not allowed. See Bray (2005), Lando (2005), Fisher (2012), and Siegel and Strulovici (2019) for this point and a generalization to multi-verdict trials.

\textsuperscript{33}Grossman and Katz (1983) focused on interim welfare and did not consider deterrence.
type). 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)).

The idea underlying the proof of Theorem 1 is to improve social welfare conditional on facing an innocent defendant and conditional on facing a guilty defendant. Since in equilibrium the defendant reports his type truthfully, the signal is not needed to determine his guilt. Instead, the signal is used to devise a sentencing scheme that induces the defendant to report truthfully, and the sentencing scheme is such that social welfare is maximized subject to truthful reporting. The relevant incentive constraint is dissuading a guilty defendant from pretending to be an innocent one. Thus, given a level of utility for the innocent defendant, we would like to choose the sentencing scheme that is the least attractive for a guilty defendant. The MLRP of the signal distribution (which, we recall, is without loss of generality) shows that this is the two-step sentence function in part (i). This step does not rely on the defendant or society being risk averse. Risk aversion then implies that the sentence for the guilty defendant must be constant; moving from a random sentence to its certainty equivalent constant sentence for the guilty defendant relaxes the incentive constraint and increases social welfare, as long as the constant sentence does not exceed \( \hat{s} \). If it does, then we can decrease it to \( \hat{s} \), which gives the highest possible social welfare.

The proof of Theorem 1, as well as all other proofs, are in the Appendix.

The optimality of a constant sentence following an admission of guilt relied heavily on Assumption 2. Without Assumption 2, it is not immediately clear how an optimal sentencing function \( S(\cdot, \hat{g}) \) following an admission of guilt should depend on the signal, whether it should involve random sentences, and if so, what properties those sentences should satisfy. The following result answers these questions.

**Theorem 2** Regardless of whether Assumption 2 holds:

(i) In any interim optimal mechanism, the innocent defendant’s sentence is a step function of \( t \), which jumps from 0 to \( \bar{s} \) at some cutoff \( \bar{t} \).

(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.

---

34The result is also 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. Here, the instrument is a whole sentence scheme and the objective concerns 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.

35This last point is not generally true for ex-ante optimal mechanisms, because decreasing the sentence for the guilty defendant leads to more crime.

36The necessity of this property follows from the uniqueness proof for Theorem 1. See Appendix D.

37Here, “generically” is in the sense of prevalent sets over the vector space of welfare functions. See Appendix E for a
then it can be chosen to be 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 \([0, \hat{s}]\). If, in addition, \( W(\cdot, g) \) and \( u(\cdot) \) are concave and at least one of them is strictly concave, then a random plea cannot be optimal.

(iv) The guilty defendant is indifferent between reporting truthfully and misreporting, that is (5) holds as an equality.

Theorem 2 shows that relaxing Assumption 2 optimally leads to the possibility of having the guilty defendant face a lottery over two sentences, which are different from the ones faced by the innocent defendant and can be chosen to be independent of the signal.\(^{38}\) The proof modifies the sentence for an innocent defendant similarly to the proof of Theorem 1. The new idea in the proof of Theorem 2 (relative to Theorem 1) is to consider the guilty defendant’s expected utility from his sentence, rather than the sentence itself, and find the utility distribution that maximizes social welfare and maintains the same expected utility for the guilty. Thus, the proof improves social welfare conditional on facing an innocent defendant and conditional on facing a guilty defendant. The proof is based on a concavification argument reminiscent of arguments used in contract theory and strategic communication. This argument does not require Assumption 2.\(^{39}\)

### 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. Thus, any modification of a given mechanism must take into account the modification’s impact on deterrence. The proof of Theorem 1 suggests that under Assumption 2 this consideration need not necessarily lead to a radically different analysis of the optimal sentencing scheme. In the 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), then 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

\(^{38}\) Alternatively, it could be based on the signal \( t \) in any way that maintains (6).

\(^{39}\) The proof of Theorem 1 requires Assumption 2, but provides sharper predictions. As stated in part (iii) of the Theorem 2, under Assumption 2 these sharper predictions can also be obtained from the proof of Theorem 2, but the proof of Theorem 1 in this case is more transparent.
commit the crime.\footnote{Recall our assumption that the ex-ante probability that an individual would be arrested for a crime that he did not commit is exceedingly low. Therefore, only changes in the expected utility of a guilty defendant affect the incentives to commit crime. In fact, the improvements in the proofs of 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, the improvements would reduce these incentives. In this case, our results continue to hold provided that the expression in the square brackets of (4) is negative, i.e., society is better off when a crime is not committed even if the perpetrator is caught and punished optimally.} 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 2 optimal deterrence may lead to sentences that exceed $\hat{s}$. In this case, the improvements constructed in Theorem 1, while increasing interim welfare, also increase the utility of guilty defendants. This increases the set of individuals who commit the crime, and may therefore lower ex-ante welfare.

Our next result identifies properties of the ex-ante optimal mechanisms, which maximize ex-ante welfare (4) among all available mechanisms. The result shows that ex-ante optimal mechanisms (with or without Assumption 2) are similar to interim optimal mechanisms without Assumption 2, as described in Theorem 2. This is because, as stated above, modifying the sentence function is a way that does not change the guilty defendant’s utility, as we do in the proof of Theorem 2, 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. On the other hand, part (iii) of Theorem 2 must be modified for ex-ante 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 $t$, which jumps from 0 to $\tilde{s}$ at some cutoff $\tilde{t}$.\footnote{The necessity of this property follows from the uniqueness proof for Theorem 1. See Appendix D.}

(ii) There is an ex-ante 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 must generically hold for any ex-ante optimal mechanism.\footnote{The notion of genericity is the same as in Theorem 2.} If the defendant’s sentence is random, then it can be chosen to be 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 (5) 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 get a sense for when a random sentence is optimal, suppose that Assumption 2 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, referring to the notation from the proof of Theorem 3, \(\hat{\bar{W}}\) is not concave below \(u(\hat{s})\), and in addition \(\hat{\bar{W}}(U^g) < \bar{W}(U^g)\).

4 Discussion

Many features of the optimal judicial mechanisms identified by Theorems 1, 2, and 3 are familiar from the American legal system. The first is the fixed sentence given to a defendant who reports he is guilty. This is always interim optimal, and often ex-ante optimal, and is similar to the plea bargaining procedure in the United States. A plea bargain makes a trial unnecessary, so the sentence cannot depend on a trial’s outcome or on evidence that would have been produced in the course of a trial. Alternatively, the optimal scheme for a guilty defendant may involve a lottery over two sentences, which is independent of the signal regarding his guilt. This is in contrast to the sentencing scheme for a defendant who claims to be innocent, which optimally depends on the signal (but also involves two sentences). A lottery that disregards the signal is similar to plea bargains 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. 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.

\[43\] 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 \(\hat{s}\).

\[44\] See Federal Rules of Criminal Procedure 11(c)(1)(C) and 11(c)(1)(B).
Intuitively, the stochastic element that may optimally follow a guilty plea captures the fact that the welfare function conditional on facing a guilty defendant may be locally convex in the defendant's utility, i.e., social preferences may be risk loving in a guilty defendant's utility. This feature can arise at sentence levels at which the ex-post welfare function $W(\cdot, g)$ is decreasing, which creates the possibility that the function $U \rightarrow W(u^{-1}(U), g)$ is convex, even when both $u$ and $W(\cdot, g)$ are concave.\textsuperscript{45} Put differently, lotteries become an efficient way of punishing a guilty defendant. Because under Assumption 2 lotteries arise optimally only when deterrence is a significant consideration in sentencing, 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.

Another feature of the optimal mechanisms that is familiar from the American legal system is that a defendant who reports he is innocent receives either a sentence of 0 or some fixed higher sentence, and this depends on whether the signal is higher than some threshold. This can be interpreted as a trial with two possible outcomes, an acquittal or a 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 produce such evidence. The evidence threshold is high if it is much more important to acquit innocent defendants than it is to punish guilty ones, as reflected by the welfare function.

An acquittal carries no punishment. It is worth emphasizing that we did not assume that the lowest sentence had to be zero. Instead, acquittals emerge as a feature of the optimal mechanisms. We also did not assume binary verdicts. This ubiquitous feature of trial systems also emerges as a feature of the 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. We point out that this justification for using the harshest possible sentence is completely different from Becker's (1968) enforcement cost justification.\textsuperscript{46} In practice there may be ethical or practical consideration that limit the maximal sentence available in a given trial. This feature is built into our model through the assumed maximal sentence $\bar{s}$, which is taken to be exogenous.\textsuperscript{47}

\textsuperscript{45}Indeed, note that the composition $g \circ f$ of two concave functions is guaranteed to be concave only if $g$ is increasing.

\textsuperscript{46}Becker's (1968) argument is that iff a guilty defendant is apprehended with probability $\gamma(e)$, where $e$ is the amount spent on law enforcement and $\gamma$ increases in $e$, and receives a sentence $s$ in this case, his expected punishment in $\gamma(e)s$. If $s$ can be increased, the same level of deterrence can be achieved by reducing $\gamma(e)$ and hence the cost of law enforcement.

\textsuperscript{47}In practice, the sentence associated with a "guilty" verdict may also depend on many factors we do not explicitly
The properties of optimal mechanisms identified in Theorems 1, 2, and 3 can be used to construct optimal mechanisms. To see this, consider a notion of welfare (interim or ex-ante), choose a distribution $F$ that is part of some available mechanism, and specify a utility level $U^g$ for the guilty defendant. Then the results identify a generically unique sentence function $S$ such that $(F, S)$ maximizes welfare among all available mechanisms with distributions $F$. This is because in an optimal mechanism a guilty defendant is indifferent between reporting truthfully and misreporting, which pins down the signal cutoff $\bar{t}$ at which the innocent’s sentence jumps from 0 to $\bar{s}$, and the proofs show how to easily identify the single sentence or lottery over two sentences for the guilty. Thus, given a set of available distributions $F$ (which is determined by the underlying evidence-generation technology and agent’s incentives), identifying optimal mechanisms becomes relatively simple.

This observation points to an important difference between the optimal mechanisms and actual criminal trials. In the United States, for example, the punishment following a conviction increases with the severity of the crime, but the conviction criterion is “beyond a reasonable doubt” (BARD) for all criminal cases. In the optimal mechanisms a conviction carries the maximal allowable punishment for the crime and circumstance but the conviction threshold could optimally vary across crimes. This is because the conviction threshold $\bar{t}$ is determined by the optimal level of utility for the guilty defendant, which in turn society’s welfare function. And this welfare function could well depend on the specific crime. For example, the sentence $\hat{s}$ from Assumption 2, which 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).

Another important difference is in 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 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 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 $\bar{s}$.

\[48\] In reality, it may be that jurors interpret BARD differently according to 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.
the optimal mechanism is modified slightly.

The importance of Assumption 1

As discussed in the previous section, we can interpret the binary sentencing scheme for innocent defendants as the outcome of a trial. We then have that in the optimal mechanisms only innocent defendants go to trial.\textsuperscript{49} This feature relies on Assumption 1 because the proofs of Theorems 1, 2, and 3 require that changing the punishment mapping $S$ does not affect the signal distributions $F$. With a binary verdict, the signal distribution following a trial can be viewed as summarized by the binary verdict, so Assumption 1 means that jurors’ verdict decisions are not affected by the punishment scheme. 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 removes the punishment aspect from the decision of the jurors, which may make it easier for them to consider the evidence presented without dwelling on the punishment that a conviction would bring. In addition, 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.\textsuperscript{50} There is also compelling evidence that jurors have a limited understanding of the sentences faced by defendants. For example, the Capital Jury Project found that most jurors “grossly underestimated” the amount of jail time associated with a guilty verdict. There is also empirical evidence that harsher sentences do not result in lower conviction rates.

\textsuperscript{49}Complete separation also arises in other papers, including Grossman and Katz (1983). In an extension, they showed that when defendants are heterogeneous in their degree of risk aversion, partial pooling can arise. Below we show that partial pooling also arises in “nearly optimal” mechanisms that achieve close to optimal welfare.

\textsuperscript{50}For example, in \textit{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.
In a study of non-homicide violent case-level data of North Carolina Superior Courts, Da Silveira (2015) finds that the probability of conviction of defendants going to trial in fact increases with the sentence that they face.51 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.52

But regardless of whether jurors are unduly influenced in their conviction decisions, in reality most defendants who are convicted in trial are guilty. One way to square this with our characterization of the 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 go 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 relaxes Assumption 1.

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

$$\hat{\lambda} = \frac{\lambda \alpha}{\lambda \alpha + (1 - \lambda)}. \tag{7}$$

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 (7), we have

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

Thus, for any fraction $\alpha > 0$ and conviction threshold $\hat{t}$ there corresponds a posterior belief $\hat{p}(\hat{t})$ of guilt. To get a rough sense of this threshold, suppose that the likelihood ratio at the optimal threshold

---

51Da Silveira’s analysis excludes the most and least severe cases to focus on a relatively homogeneous pool of cases.

52Elder (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.
$\bar{t}$ 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.\textsuperscript{53} Also suppose that, consistent with criminal data in the United States, 90% of defendants are in fact guilty.\textsuperscript{54} 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 $\lambda$ of guilt, our modified mechanism remains consistent with BARD and the observation that most defendants are guilty. With a fraction $\alpha$ of guilty defendants going to trial, we incur a welfare loss relative to the optimal mechanism since these guilty defendants are sometimes acquitted and sometimes punished too severely. But this loss concerns only a small fraction of guilty defendants. In addition, once some guilty defendants go to trial, evidence regains its role in determining the defendant’s guilt, in addition to its role in incentivizing some guilty defendants to accept the plea bargain.

5 Conclusion

This paper uses modern mechanism design to identify some properties of optimal judicial systems. We reduce dynamic, multi-player judicial processes to single-player revelation mechanisms focused on the defendant, and perform a mechanism design analysis in this setting. We then show that optimal mechanisms have features that parallel many of those in the American criminal justice system, including binary verdicts, a conviction following sufficiently incriminating evidence, no punishment following an acquittal, and plea bargains with certain and uncertain punishments. One difference between our results and actual trials is that in optimal mechanisms the conviction threshold, and not only the punishment,


\textsuperscript{54}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.
may vary across crimes and circumstances. Another difference is that in optimal mechanisms only innocent defendants go to trial, and the role of evidence in a trial is to incentivize guilty defendants to take a plea bargain and not to determine whether they are actually guilty. However, mechanisms in which a fraction of guilty defendants go to trial achieve close to maximal welfare and recover the role of evidence in actual trials. This suggests that the combination used in practice of plea bargains and trials with binary verdicts can generate high welfare, and that evidence in actual trials may also play a role in facilitating the institution of plea bargaining, in addition to its use in determining defendants’ guilt.

One of our key assumptions is that the defendant’s only private information is whether he is guilty. This may be appropriate for a first step in the analysis of optimal criminal justice systems, because the defendant’s private knowledge of his guilt is an issue all criminal justice systems must contend with. In reality, 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 and their risk attitudes. 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 with such defendants would be an interesting direction for future work.

---

55 Our analysis does, however, allow for private information by other actors in the system, such as the prosecutor. See Appendix H.

56 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 \((F,S) \in \mathcal{M}\). We modify the mechanism in a way that maintains truthfulness and increases interim welfare. We modify only the sentence function without changing the signal distributions \(F\). Assumption 1 then ensures that the resulting mechanisms is available.

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

\[
\int_0^1 u(\tilde{S}(t, i)) f^i_g(t) dt = u(\hat{s}) F^i_g([0, \tilde{t}]) + u(\hat{s}) F^i_g([\tilde{t}, 1]) = \int_0^1 u(S(t, i)) f^i_g(t) dt. \tag{8}
\]

The cutoff \(\tilde{t}\) exists because distribution \(F^i_g\) has no atoms.\(^{57}\) Rearranging 8 yields

\[
\int_0^1 [u(S(t, i)) - u(\tilde{S}(t, i))] f^i_g(t) dt = 0. \tag{9}
\]

The function \(t \mapsto u(S(t, i)) - u(\tilde{S}(t, i))\) crosses 0 once from below, since \(u(S(t, i))\) lies in the interval \([u(\hat{s}), u(0)]\) for all \(t\) and any sentence function \(S(\cdot, i)\), while \(u(\tilde{S}(t, i))\) equals \(u(0)\) for \(t \leq \tilde{t}\) and jumps down to \(u(\hat{s})\) at \(t = \tilde{t}\). This increases social welfare, provided that truthfulness is maintained. Truthfulness is maintained because (10) and the fact that (6) holds for mechanism \((F, S)\) shows that (6) continues to hold when \(S(\cdot, i)\) is replaced with \(\tilde{S}(\cdot, i)\).

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

\[
u(s^{ce}) = \int_0^1 u(S(t, \hat{g})) f^\hat{g}_g(t) dt.
\]

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

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 \(s^{ce} \geq s^g\) and \(W(\cdot, g)\) is increasing up to \(\hat{s}\), we have \(W(s^{ce}, g) \geq W(s^g, g) \geq \int_0^1 W(S(t, \hat{g}), g) f^\hat{g}_g(t) dt\), so \(s^g\) increases welfare conditional on facing a guilty defendant. If instead \(s^{ce} > \hat{s}\), we set \(s^g = \hat{s}\). This sentence yields the highest possible social welfare conditional on facing a guilty defendant.

By construction the guilty defendant is indifferent between \(s^{ce}\) and reporting truthfully with the sentence function \(S(\cdot, \hat{g})\). Since \(s^g \leq s^{ce}\), he thus prefers \(s^g\) to reporting truthfully with \(S(\cdot, \hat{g})\). By construction of \(\tilde{S}(\cdot, i)\)

\(^{57}\)If there is an atom at the relevant signal, randomizing between 0 and \(\hat{s}\) with the correct probability generates the requisite indifference.

\(^{58}\)The result is proved by a simple integration by parts, and follows from a result initially proved by Karlin (1968). See Athey (2002) for a statement of the result and Friedman and Holden (2008) for a recent example of its use in economics.
and the fact that (5) holds for mechanism \((F, S)\), he prefers reporting truthfully with sentence function \(S(\cdot, \hat{g})\) to misreporting with sentence function \(\tilde{S}(\cdot, \hat{i})\). Thus, he prefers sentence \(s^g\) to misreporting with sentence function \(\tilde{S}(\cdot, \hat{i})\), so truthfulness is maintained for the guilty defendant, that is, (5) continues to hold when \(S(\cdot, \hat{g})\) is replaced with \(s^g\).

If (5) 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 \(\tilde{S}(\cdot, \hat{i})\). 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

\[
  u(s^g) = \int_0^1 u(\tilde{S}(t, \hat{i})) f^g_1(t) dt = \int_0^1 [u(s^g) - u(\tilde{S}(t, \hat{i}))] f^g_1(t) dt = 0,
\]

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

\[
  \int_0^1 [u(s^g) - u(\tilde{S}(t, \hat{i}))] f^g_1(t) dt \leq 0 \Rightarrow \int_0^1 u(\tilde{S}(t, \hat{i})) f^g_1(t) dt \geq \int_0^1 u(s^g) f^g_1(t) dt,
\]

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 \(S(\cdot, \hat{i})\) with a step function \(\tilde{S}(\cdot, \hat{i})\) that is equal to zero below \(\bar{t}\) and equal to \(\check{s}\) above it, with \(\bar{t}\) chosen to make a guilty defendant indifferent between \(\tilde{S}(\cdot, \hat{i})\) and \(S(\cdot, \hat{i})\) when misreporting his type, so an innocent defendant prefers \(\tilde{S}(\cdot, \hat{i})\) to \(S(\cdot, \hat{i})\) when reporting truthfully. The cutoff \(\bar{t}\) is now increased until the guilty defendant is indifferent between \(S(\cdot, \hat{g})\) and \(\tilde{S}(\cdot, \hat{i})\). 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

\[
  \max_{s^g(\cdot) \in \Delta[\check{s}, \bar{s}]} \int_0^1 W(s(t), g) f^g_2(t) dt
\]

subject to

\[
  \int_0^1 u(s(t)) f^g_2(t) dt = U^g.
\]

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

\[
  \max_{u(\cdot) \in \Delta[u(\check{s}), u(0)]} \int_0^1 E[\bar{W}(\hat{u}(t))] f^g_2(t) dt
\]

subject to

\[
  \int_0^1 E[\bar{u}(t)] f^g_2(t) dt = U^g,
\]

where \(\bar{W}(U) = W(u^{-1}(U), \hat{g})\) for any \(U \in [u(\check{s}), 0]\). The two formulations are equivalent because the defendant’s utility \(u(\cdot)\) is strictly decreasing in the sentence.
To characterize the solution of (11) subject to (12), it is useful to consider a simpler optimization problem:

\[
\max_{\hat{u} \in \Delta[u(s), u(0)]} E[\hat{W}(\hat{u})] \tag{13}
\]

subject to

\[
E[\hat{u}] = U^g \tag{14}
\]

Consider a stochastic process \( \hat{u} : T \to \Delta[u(s), u(0)] \) whose sample paths are Lebesgue measurable and that satisfies (12). This process induces a measure \( F^u \) over \([u(s), u(0)]\) such that for any Borel subset \( B \) of \([u(s), u(0)]\), \( F^u(B) = \int_0^1 Pr(\{\hat{u}(t) \in B\}) f^\hat{u}_g(t) dt \). Intuitively, \( F^u(B) \) is the probability that the defendant receives a utility level in \( B \) given the utility process \( \hat{u} \) and given that the signal \( t \) is distributed according to \( F^g \). Let \( \hat{u} \) denote a random variable distributed according to \( F^u \). By construction, \( \hat{u} \) satisfies (14) and

\[
\int_0^1 E[\hat{W}(\hat{u}(t))] f^\hat{u}_g(t) dt = E[\hat{W}(\hat{u})]. \tag{15}
\]

Therefore, \( \hat{u} \) is a solution of (11) subject to (12) if and only if \( \hat{u} \) is a solution of (13) subject to (14).

We now solve for (13) subject to (14). For any \( U \) in the interval \([u(s), u(0)]\), let

\[
\hat{W}(U) = \sup\{x : (U, x) \in co(\hat{W})\},
\]

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

It is well-known that \( \hat{W}(U^g) \) is the solution of (13) subject to (14):\(^{59}\) If \( \hat{W}(U^g) = \hat{W}(U^g) \), the maximal value is achieved by the constant sentence \( u^{-1}(U^g) \). In this case, by (15), an optimal \( \hat{u} \) is achieved by the sentence function \( \hat{S}(\cdot, \hat{g}) \equiv u^{-1}(U^g) \), which is constant in the signal \( t \). If \( \hat{W}(U^g) < \hat{W}(U^g) \), the maximal value is achieved by randomizing between \( u^{-1}(U) \) and \( u^{-1}(U) \), where \( U = \max \{v < U^g : \hat{W}(v) = \hat{W}(v)\} \) and \( \hat{U} = \min \{v > U^g : \hat{W}(v) = \hat{W}(v)\} \), with probabilities \( \alpha \) and \( 1 - \alpha \) such that \( \alpha \hat{U} + (1 - \alpha) \hat{U} = U^g \). In this case, again by (15), the constant stochastic sentence function \( \hat{S}(\cdot, \hat{g}) \) (which is independent of the signal) that for every signal \( t \) assigns sentence \( u^{-1}(U) \) with probability \( \alpha \) and sentence \( u^{-1}(U) \) 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 \( \hat{u}(\hat{s}) \), which proves that if \( \hat{W}(U^g) < \hat{W}(U^g) \), then the two-point support lies in \([0, \hat{s}]\).\(^{60}\) 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 \( W \) 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 \( \hat{S}(\cdot, \hat{g}) \) and \( \hat{S}(\cdot, \hat{i}) \), that is,

\[
\int_0^1 u(\hat{S}(t, \hat{g})) f^\hat{g}_g(t) dt - \int_0^1 u(\hat{S}(t, \hat{i})) f^\hat{g}_g(t) dt = 0,
\]

so (5) holds when \( S \) is replaced with \( \hat{S} \). Moreover, since function \( \hat{S}(\cdot, \hat{g}) \) is independent of the signal, the last equality can be written as

\[
\int_0^1 [u(\hat{S}(t, \hat{g})) - u(\hat{S}(t, \hat{i}))] f^\hat{g}_g(t) dt = 0.
\]

\(^{59}\) A more detailed discussion of a similar use of concavification appears in Aumann et al. (1995) and Kamenica and Gentzkow (2011). These papers concern concavification with respect to beliefs. Concavification has been used in other contexts, particularly 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).

\(^{60}\) Sentences higher than \( \hat{s} \) can be replaced by \( \hat{s} \), which increases interim welfare and relaxes the incentive constraint.
This is equivalent to
\[
\int_0^1 [u(s^{ce}) - u(\tilde{S}(t, \hat{\tilde{\iota}}))] f_{g}(t) dt = 0,
\]
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
\[
\int_0^1 [u(s^{ce}) - u(\tilde{S}(t, \hat{\tilde{\iota}}))] f_{\hat{\iota}}(t) dt \leq 0,
\]
which is equivalent to
\[
\int_0^1 [u(\tilde{S}(t, \hat{g})) - u(\tilde{S}(t, \hat{\iota}))] f_{\hat{\iota}}(t) dt \leq 0.
\]
This can be written as
\[
\int_0^1 u(\tilde{S}(t, \hat{g})) f_{\hat{\iota}}(t) dt - \int_0^1 u(\tilde{S}(t, \hat{\iota})) f_{\hat{\iota}}(t) dt \leq 0
\]
because \(\tilde{S}(\cdot, \hat{g})\) is independent of the signal. This shows that (6) 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 (4). 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 \(\tilde{S}(\cdot, \hat{\iota})\) increases the utility of innocent defendants relative to mechanism \(S(\cdot, \hat{\iota})\), 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 (4). 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 \([0, \bar{s}]\), then \(\hat{W}\) is concave on the utility interval \([u(\hat{s}), u(0)]\). In this case, \(\hat{W}\) coincides with \(W\) for \(U_{g} \geq u(\hat{s})\), 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.](image)

D Proof of uniqueness in Theorem 1

Take a truthful mechanism \((F, S)\). Suppose first that \(S\) violates i) over a positive measure of signals. In this case, the step function \(\tilde{S}(t, \hat{i})\) constructed in the first part of the proof is such that the difference \(S(t, \hat{i}) - \tilde{S}(t, \hat{i})\)
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.\(^{61}\) Since $u$ is strictly decreasing, this implies that the single-crossing function $\delta: t \mapsto \delta(t) = u(S(t, \hat{i})) - u(\hat{S}(t, \hat{i}))$ is strictly negative over $T_1$ and strictly positive over $T_2$. Let $H(t) = \int_{\hat{t}}^{1} \delta(\tau)f'_{\hat{g}}(\tau)d\tau$. 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 $T_1 \cup T_2$.\(^{62}\) Let $\gamma(t) = f'_{\hat{g}}(t)/f_{\hat{g}}(t)$. By strict MLRP, $\gamma$ is a strictly increasing function and thus almost everywhere differentiable. Therefore,

$$
\int_{[0,1]} \delta(t)f'_{\hat{g}}(t)dt = \int_{[0,1]} \delta(t)f'_{\hat{g}}(t)\gamma(t)dt = \int_{[0,1]} -H'(t)\gamma(t)dt = \int_{[0,1]} H(t)\gamma'(t)dt < 0
$$

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

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

Suppose now that $S$ violates ii), i.e., $S(t, g)$ is non-constant. There are two cases to consider. If $u$ is strictly concave, then the certainty equivalent $s^ce$ is strictly higher than $s^g$: it is possible to increase a guilty defendant’s expected punishment without violating incentive compatibility. If $s^ce \leq \hat{s}$, then because $W(s, g)$ 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 $\hat{\hat{W}}$ are such that whenever $\hat{W}$ is linear over some maximal interval $I$ (i.e., there is no interval strictly containing $I$ over $\hat{W}$ 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 $\hat{W}$ is to randomize over the endpoints of $I$, i.e., to use a two-point lottery. Over the remaining domain of $\hat{W}$, $\hat{W}$ and $\hat{\hat{W}}$ coincide, and because $\hat{W}$ is locally strictly concave (since it is always

---

\(^{61}\)Indeed, the difference must be non-zero over a set of positive measure. Since $t \mapsto S(t, \hat{i}) - \hat{S}(t, \hat{i})$ is single crossing from positive to negative, this implies that the existence of one of the two sets mentioned. Finally, since $S(t, \hat{i})$ and $\hat{S}(t, \hat{i})$ 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 $S(t, \hat{i})$ strictly exceeds $\hat{S}(t, \hat{i})$ over a set of positive measure, it must also be exceeded by it over a set of positive measure.

\(^{62}\)The fact that $H(0) = 0$ is simply a restatement of (0). Nonnegativity of $H$ comes from the fact that the integrand of $H$, $\delta(t)f'_{\hat{g}}(t)$, 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 $T_2$. Theorem.
is linear coincides with $w$ domain is always continuous.

The notion of “almost all” that we choose is the standard mathematical notion of “prevalence,” which is used to conceptualized genericity for infinite-dimensional sets, like the set of functions that we consider here.

Given a topological vector space $W$, a subset $\mathcal{G}$ is said to be prevalent if there exists a finite dimensional subspace $V$ of $W$ such that for all $w \in W$, we have $w + v \in \mathcal{G}$ for all $v \in V$ except for a set of $v$ that has Lebesgue measure zero in $V$. Intuitively, it means that almost all translations of $w$ by elements in $V$ 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 $U \mapsto W(U) = W(u^{-1}(U), g)$. Since $u^{-1}$ is continuous 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 $W$ denote the vector space of all real-valued, continuous functions over $[0, 1]$ and $\mathcal{G}$ be the subset of $W$ consisting of all functions $w$ whose concavification $\tilde{w}$ 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 $W$.

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

Given a function $w \in W$, 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 $\{I^a_k\}_k$ denote the collection of maximal intervals of $[0, 1]$ over which the concavification $\tilde{w}_a$ of $w_a$ is linear and coincides with $w_a$ at three or more points of these intervals. Since these intervals are maximal, they are closed. Moreover, if $a$ is increased slightly, it is straightforward to see, by strict convexity of $f$, that there are at most two points of contact over $I^a_k$ for all $a' > a$: all interior points $x$ of $I^a_k$ are such that $w_{a'}(x) < \tilde{w}_a(x)$.

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 $I^a_{k'}$ that are either in the closure of the complement of $\cup_k \{I^a_k\}$, or consist of intervals that strictly contain some $I^a_k$. In particular, one may associate to each new interval a rational number $r_{a', k'}$ that belongs to $I^a_{k'}$, but not to any other interval $I^a_k$.

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 $r_{a', k}$ 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.

---

63 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.

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

65 Any strictly convex (or strictly concave) function would work equally well.

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

---
The restriction on the maximal possible sentence

We consider the impact of the assumption that the sentence is restricted to the interval \([0, \bar{s}]\), 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.

Risk-attitude discrepancy between social and individual preferences

While social preferences may be broadly aligned with those of the defendant when he is innocent, they need not be identical. We now relax the assumption, common in the literature (Grossman and Katz (1983)), that \(W(\cdot, i) = u(\cdot)\), and assume instead that there exists an increasing transformation \(\phi : \mathbb{R} \to \mathbb{R}\) such that \(W(s, i) = \phi(u(s))\).

According to this representation, harsher sentences decrease welfare when the defendant is innocent, thus preserving the ordinal alignment of social and defendant preferences. However, the representation allows different perceptions of risk between the social preference and the defendant. Our result in this section concerns the case in which \(\phi\) is convex, which means that the social preferences when facing an innocent defendant exhibit less risk aversion than the defendant’s preferences. This assumption captures the idea that society may not internalize the full extent of an innocent defendant’s exposure to the judicial process.

**Proposition 1** Suppose that \(\phi\) 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 \(S(\cdot, i)\) with a step function \(\tilde{S}(\cdot, i)\) that is equal to zero below \(\bar{\ell}\) and equal to \(\bar{s}\) above it, with \(\bar{\ell}\) chosen to make an innocent defendant indifferent between \(S(\cdot, i)\) and \(\tilde{S}(\cdot, i)\).
For expositional simplicity, let us normalize the utility functions as follows: \( u(0) = 0, u(\hat{s}) = -1, \phi(0) = 0 \) and \( \phi(-1) = -M \). This normalization is without loss of generality, as is easily checked.\(^{67}\) We wish to show that

\[
\int_0^1 W(S(t, i))f_i^1(t)dt \leq \int_0^1 W(\hat{S}(t, i))f_i^1(t)dt = -MF_i^1([\bar{t}, 1]),
\]

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

\[
\int_0^1 \phi(u(S(t, i)))f_i^1(t)dt \leq -MF_i^1([\bar{t}, 1]), \tag{16}
\]

It follows from the indifference equation (8) and the above normalization that the cutoff \( \bar{t} \) satisfies

\[
-F_i^1([\bar{t}, 1]) = \int_0^1 u(S(t, i))f_i^1(t)dt. \tag{17}
\]

Since \( u(\cdot) \) takes values in \([-1, 0]\) we can view \(-u(\hat{s})\) as a weight in a convex combination. Since also \( u(0) = \phi(0) = 0, u(\hat{s}) = -1, \) and \( \phi(-1) = -M \), we have\(^{68}\)

\[
\phi(u(S(t, i))) = \phi \left[ (-u(S(t, \hat{i})))(-1) + (1 - (-u(S(t, i))))0 \right] \\
\leq (-u(S(t, i)))\phi(-1) + (1 - (-u(S(t, i))))\phi(0) \\
= Mu(S(t, i)).
\]

Integrating this equation for \( t = 0 \) to \( 1 \) with respect to the density \( f_i^1 \) yields

\[
\int_0^1 \phi(u(S(t, i)))f_i^1(t)dt \leq M \int_0^1 u(S(t, i))f_i^1(t)dt.
\]

Combining this with (17) then yields (16). \( \blacksquare \)

## H From Judicial Processes to Direct Revelation Mechanisms: A Formalization

We describe a multi-agent environment and a reduction to a single-agent environment focused on the defendant that lead to Assumption 1. Intuitively, Assumption 1 entails two distinct features:

1. Given any judicial mechanism and equilibrium strategy profile for all players, there is another “direct” judicial mechanism for which i) the defendant has only one action, which is to report his type, ii) there is an equilibrium of the direct mechanism in which players other than the defendant follow the same strategy profile as in the initial equilibrium and the defendant’s equilibrium strategy is to truthfully report his type;

\(^{67}\)The utility of the defendant can always be translated and scaled without affect the defendant’s incentives. Likewise translating the welfare function has not impact on the optimization of ex-ante welfare.

\(^{68}\)The inequality is a direct application of the definition of \( \phi \)'s convexity if \( t \mapsto S(t, \hat{i}) \) is deterministic. If \( S(t, i) \) is a lottery, the proof is equally straightforward. For example, fixing some \( t \), suppose that \( S(t, \hat{i}) \) is a lottery with distribution \( g \). Then \( \phi(u(S(t, i))) = \int_{[0,1]} \phi(u(\hat{s}))g(\hat{s})d\hat{s} \). For each \( \hat{s} \), the convexity of \( \phi \) and together with \( u(\hat{s}) \in [-1,0], u(\hat{s}) = -1, u(0) = 0, \phi(0) = 0, \) and \( \phi(-1) = -M, \) imply \( \phi(u(\hat{s})) = \phi((-u(\hat{s})(-1) + (1 - (-u(\hat{s}))))(0)) = (-u(\hat{s}))(\phi(-1) + (1 - (-u(\hat{s}))))\phi(0) = Mu(\hat{s}) \). Integrating over \( \hat{s} \) then yields \( \phi(u(S(t, i))) \leq M \int_{[0,1]} u(\hat{s})g(\hat{s})d\hat{s} = Mu(S(t, i)) \).
2. Given i) any direct judicial mechanism and equilibrium strategy profile in which the defendant reveals his type truthfully, and ii) any change of the mechanism’s sentencing scheme for which truth-telling is still optimal for the defendant keeping other actors’ strategy profile unchanged, the same equilibrium strategy profile is an equilibrium under the new sentencing scheme.

To implement the second feature, one could assume direct and full control of the designer over the actors other than the defendant, i.e., that the designer can choose their strategies. The microfoundation we provide is more general as it requires only that if a strategy profile is an equilibrium in the original judicial process, then it remains optimal for the other actors to behave in the same way provided the defendant does, regardless of the sentencing scheme. For the first feature, we postulate the existence of a mediator (Myerson (1983, 1986)).

Formally, a judicial mechanism as an extensive-form game with incomplete information and a finite horizon. The players of this game, indexed by $i \in I$, represent all actors of the judicial mechanism. At the beginning of the game, nature draws players’ types, which lie in some probability space $(T = \times_{i \in I} T_i, \Sigma_T, \mu)$ and may be correlated. The outcome of the mechanism consists of i) some evidence $E$ taking value in some probability space $(E, \Sigma_E, \nu)$, ii) a sentence $s \in \mathbb{R}_{+}$, iii) the realized utility $u_i$ of each player $i \in I$, iv) a realized gross social welfare $w$, and a social cost $c \geq 0$, which captures the time, money, and effort invested in the mechanism.

Throughout the analysis, we maintain the assumptions that, i) the defendant’s type is binary (guilty or innocent), ii) the realized sentence $s$ lies in some exogenously given interval $[0, s]$, iii) the defendant’s realized utility is an exogenously given function $u(s)$ of the realized sentence $s$, iv) the realized gross welfare is an exogenously given function $w(s, \theta)$ of the realized sentence $s$ and the defendant’s type $\theta$.

We let $C$ denote the class of all judicial mechanisms that are available to the designer. In the analysis to follow, various equilibrium concepts may be used without affecting any of the results. For concreteness, we focus on sequential equilibrium.

**Definition 1** A judicial mechanism is regular if

1) the order of moves is independent of the defendant’s type, and the set of action nodes and actions of all players (including moves of nature following nature’s determination of the defendant’s type at the root) is independent of the defendant’s type;

2) the defendant learns his type immediately after types are drawn;

3) the realized social cost is a function of the sequence of actions taken by players other than the defendant.

4) there exists an equilibrium $\sigma$.

Part 1) means that the two subtrees that follow nature’s determination of the defendant type are symmetric. This property is used in the construction of a mediated mechanism below. A judicial process is a pair consisting of a regular judicial mechanism and an equilibrium profile $\sigma$.

**Mediated mechanism.** Given a judicial mechanism in $C$, we consider a modification that reduces the defendant’s decision problem to reporting his type. This reduction is formalized as follows: starting from a judicial mechanism, a new node is inserted immediately after nature has drawn players’ types, at which the defendant privately reports his type to a mediator. The rest of the game tree is unchanged except that the

---

69 For instance, the type of a witness is correlated with the type (guilty or innocence) of the defendant.

70 The separation between welfare and cost parallels their separation in Section 2. In particular, welfare captures the trade off between Type I and Type II errors, as well as any expenditures associated with the sentence, while costs include any information acquisition and administrative costs of the judicial process.
mediator now replaces the defendant in every one of the defendant’s action nodes, and the mediator’s information sets are different. The information sets of the mediator reflect the fact that instead of observing the defendant’s true type, he observes the defendant’s report of his type. The mediator moves at all nodes in which the defendant would have moved, observing the same history of actions by other players as the defendant would. The mediator takes the same actions as the defendant would at the corresponding information set (mixing with the same probabilities whenever the defendant randomizes over actions) if his type were equal to the type reported by the defendant. Part 1) of the definition of a regular judicial mechanism guarantees that this modification is well defined. The utility of the players is as in the initial judicial process, except that their dependence on the defendant’s actions is now replaced by an identical dependence on the mediator’s actions.\textsuperscript{71} The realized welfare and cost are the same as in the initial judicial process. The defendant’s utility is the same function of the final sentence as in the initial judicial process.

Assumption 3 (Feasible Mediation) Given a judicial mechanism in $\mathcal{C}$, the mediated mechanism associated to it also belongs to $\mathcal{C}$.

Lemma 1 Given a regular judicial process, the associated mediated mechanism has the following properties:

1) It is regular;

2) Letting $\sigma$ denote the equilibrium of the initial judicial process, the augmented strategy profile in which the defendant reports his type truthfully and other players (including the mediator) follow $\sigma$ is an equilibrium of the mediated mechanism;

3) Under this augmented strategy profile, the realized welfare and cost have the same probability distributions as those arising under the initial judicial process

Proof. Suppose that the defendant truthfully reveals his type. Then, by construction, the mediator follows the correct strategy. Given this, other players’ incentives are identical to the initial judicial mechanism, in which $\sigma$ is an equilibrium. Since there is no off-path report by the defendant (he uses both reports with positive probability in equilibrium), there is no issue of off-path belief for the mediator. Finally, given that the mediator and other players are playing $\sigma$, truth-telling is clearly optimal for the defendant: if it were not, following the other type’s strategy would be a strict improvement in the initial judicial mechanism, contradicting the premise that $\sigma$ was an equilibrium in that mechanism.

A mediated mechanism together with the equilibrium described in the previous lemma is called a mediated process. In a mediated process, the equilibrium is truthful in the sense that the defendant reports his type truthfully.

We now turn to our main invariance assumption. Given a mediated mechanism, a sentencing scheme is a map $\tilde{s} : (E, \tilde{\theta}) \rightarrow \Delta([0, \bar{s}])$ where $E$ is the evidence generated by the mechanism and $\tilde{\theta}$ is the type reported by the defendant. A mediated mechanism is an $\tilde{s}$-modification of the initial mediated mechanism if it is identical to that mechanism except for the following changes: 1) For each leaf of the game, the outcome sentence $s$ is replaced by $\tilde{s}$. That is, for any leaf of the mediated mechanism, we replace the outcome sentence $s$ by a lottery $\tilde{s}$ whose distribution depends only on the realized evidence $E$ and the defendant’s report $\tilde{\theta}$. If the lottery is degenerate, this entails no modification of the game tree. In particular, note that each leaf of the game depends on all past play, including the defendant’s report $\tilde{\theta}$, so that the sentence can depend on $\tilde{\theta}$. If the lottery is non-degenerate,
the leaf is replaced by a move of nature determining the realization of the lottery followed by an outcome giving
the realized sentence together with all other outcomes of the game.

An \( \tilde{s} \)-modification thus preserves the game tree of the initial mediated mechanism, except possibly for
the leaves, which may be replaced by a lottery followed by new leaves, but changes the sentences of the initial
mediated mechanism in a particular way that depends only on the evidence \( E \) and the report \( \hat{\theta} \).

Given a mediated process with equilibrium strategy profile \( \sigma \) and sentencing scheme \( \tilde{s} \), let \( u(\hat{\theta}; \theta, \tilde{s}) \) denote
the expected utility of the defendant of type \( \theta \) who reports \( \hat{\theta} \) when other players—with the mediator replacing
the defendant as explained above—follow the equilibrium strategy profile \( \sigma \) in any \( \tilde{s} \)-modification of the mediated
mechanism.\(^{72} \) A sentencing scheme \( \tilde{s} \) is \textbf{truthful} if \( u(\hat{\theta}; \theta, \tilde{s}) \geq u(\hat{\theta}; \theta, \tilde{s}) \) for all pairs \( (\theta, \hat{\theta}) \).

\textbf{Assumption 4} Given any mediated mechanism with equilibrium profile \( \sigma \) and any truthful sentencing scheme
\( \tilde{s} \), there exists an \( \tilde{s} \)-modification of the mediated mechanism such that if the defendant reveals his type truthfully,
the strategy profile \( \sigma \) forms a continuation equilibrium of the \( \tilde{s} \)-modification. Such a modification, together with
the truthful equilibrium \( \sigma \), is called \textbf{admissible}.

Assumption 4 captures two forms of commitment. First, the designer can implement any sentencing scheme:
he can optimize over all mappings from signals and reports into sentences. This corresponds to the form of
commitment power that is usually assumed in mechanism design. Second, starting from a given mediated
process, the designer can modify the sentencing scheme without affecting the incentives of the actors in the
process, other than the defendant. If the new sentencing scheme is truthful, the assumption implies that i) the
same signals about the defendant’s guilty can be generated under the new scheme and, ii) the expected social
cost, which was assumed to depend only on the actions taken by players other than the defendant, is unchanged.

For instance, this assumption may impose that an investigator’s effort to look for evidence be unaffected
by the sentencing scheme faced by the defendant. Likewise, a defense lawyer or prosecutor’s effort to argue
their case should be unaffected by the sentencing scheme. An alternative interpretation is that the designer can
always find actors willing to perform their tasks regardless of the sentencing scheme. The US criminal justice
system has features that resemble this assumption: for example, in an adversarial system, each side has an
incentive to obtain an acquittal or, respectively, a conviction, regardless of the particular sentence associated
with a conviction. Likewise, jurors are tasked with a fact finding mission that is independent of the possible
sentence given to the defendant. Juror selection may also be viewed as an instrument of the justice system that
helps achieve this objective.

The proof of the following lemma is a straightforward consequence of Assumption 4 and the definition of an
admissible modification.

\textbf{Lemma 2} Given a mediated process, any admissible \( \tilde{s} \)-modification is also a mediated process, whose associated
equilibrium is identical to the initial one. In particular, the defendant reports his type truthfully.

\textbf{Definition 2} A mediated process is \textbf{acceptable} if the following properties hold in equilibrium:

1) The distribution of \( E \) has the same support over \( \mathcal{E} \) for \( (\theta, \hat{\theta}) \in \{g, i\}^2 \), with positive density functions \( f_\theta^\hat{\theta} \)
over the support;\(^{73} \)

2) For any report \( \hat{\theta} \), the likelihood ratio \( \ell(E, \hat{\theta}) = f_\theta^\hat{\theta}(E)/f_\theta^\hat{\theta}(E) \) is a random variable that is pinned down by the
realized evidence \( E \). The distribution of \( \ell(E, \hat{\theta}) \) is assumed to be \( \hat{\theta} \in \{g, i\} \) has a non-atomic distribution
with support \( \mathbb{R}_+ \);

\(^{72}\)The defendant’s expected utility does not depend on which \( \tilde{s} \)-modification is considered, conditional on \( \sigma \).

\(^{73}\)The density is taken with respect to measure \( \nu \) over \( \mathcal{E} \).
3) Conditional on $E$ and $\hat{\theta}$, $s$ is statistically independent of $\theta$.

Part 3) of this definition does not rule out sentences that depend on actors’ private information. Rather, it states that conditional on the admissible evidence $E$ and the defendant’s reported type, no other information about the defendant’s type can affect the sentence. Thus, if some actor initially had private information about the defendant’s type, either this information is incorporated in the evidence (through a witness report, for example), or it does not affect the sentence (e.g., a family member of the defendant is aware of incriminating evidence that is never revealed in the process). However, other forms of the actors’ private information (such as their ability to competently investigate the case or, in the case of jurors, hidden biases pro or against the defendant that are independent of the defendant’s true type) can affect the sentence in arbitrary ways.\(^{74}\)

Two judicial processes are said to be welfare equivalent if they induce the same distribution for i) the realized welfare conditional on each type of the defendant and ii) the social cost.

The purpose of the next result is to show that, while sentences a priori depend in arbitrary ways on the evidence uncovered in the case, one may for the purpose of maximizing welfare focus without loss of generality on sentencing schemes that depend only on the likelihood ratio of guilt implied by the evidence. The challenge is that several evidence collections can a priori lead to the same likelihood ratio, and the key step in the argument is to show that given any likelihood ratio level, the distribution of evidence that yields this level is independent of the defendant’s true type.

**Lemma 3** Given any acceptable mediated process, there exists a truthful sentencing scheme $\tilde{s}$ that depends on $(E, \hat{\theta})$ only through $\ell(E, \hat{\theta})$ and $\hat{\theta}$ such that any admissible $\tilde{s}$-modification is welfare equivalent to the initial mediated process. Moreover, under any such modification, the sentence $\tilde{s}$ is statistically independent of $\theta$ conditional on $\ell(E, \hat{\theta})$ and $\hat{\theta}$.

**Proof.** Fix any acceptable mediated process, $l \in (0, \infty)$, and $\hat{\theta} \in \{g, i\}$, and let $E(l, \hat{\theta})$ denote the set of $E \in E$ for which $\ell(E, \hat{\theta}) = l$. For $\theta \in \{g, i\}$, let $F^\theta_i(\cdot | \theta)$ denote the probability distribution of $E$ over $E(l, \hat{\theta})$ conditional on report $\hat{\theta}$, true type $\theta$. The key is to observe that $F^\theta_i(\cdot | \theta)$ is independent of $\theta$. Indeed, by construction we have $f^\theta_i(E) = l f^\theta_i(E)$ for all $E \in E(l, \hat{\theta})$. Integrating over any measurable subset $B$ of $E(l, \hat{\theta})$, we obtain $F^\theta_i(B) = l F^\theta_i(B)$ for any such $B$. Since this is true in particular for $B = E(l, \hat{\theta})$, we conclude that $F^\theta_i(B | g) = F^\theta_i(B | i)$, since the left-hand side equals $F^\theta_i(B) / F^\theta_i(E(l, \hat{\theta}))$ while the right-hand side equals $F^\theta_i(B) / F^\theta_i(E(l, \hat{\theta}))$, and these ratios are equal by the previous observations.

For any $l$ and $\hat{\theta}$, define $\tilde{s}$ to be the sentence lottery over $[0, \bar{s}]$ whose distribution is equal to the distribution of the initial sentence $s$ conditional on $\hat{\theta}$ and the event $\ell(E, \hat{\theta}) = l$. Definition 2 guarantees that the distribution of $s$ conditional on $E$ and $\hat{\theta}$ is independent of $\theta$, and we have just showed that the distribution of $E$ conditional on $l$ and $\hat{\theta}$ is also independent of $\theta$. Combining these observations, it follows that the distribution of $\tilde{s}$ is well defined independently of $\theta$.\(^{75}\)

\(^{74}\)Of course, any effect of biases on the sentence are suboptimal from a welfare perspective, and the optimal mechanism derived in Section 3 depends only on the likelihood ratio as shown by Lemma 3 below, and not on any other private information held by any of the other actors of the judicial mechanism.

\(^{75}\)More precisely, we need to show that $s$ is independent of $\theta$ conditional on $\hat{\theta}$ and $l$. Consider any bounded functions $g, h$ of $s$ and $\hat{\theta}$, respectively, and let $F^\theta_i$ denote the distribution of $E$ conditional on $E(l, \hat{\theta})$, which we have shown to be independent of $\theta$. We have $E[g(s)h(\theta)|l, \hat{\theta}] = \sum_{E \in E(l, \hat{\theta})} E[g(s)h(\theta)|E, \hat{\theta}]dF^\theta_i(E) = \sum_{E \in E(l, \theta)} E[g(s)|E, \hat{\theta}]E[h(\theta)|E, \hat{\theta}]dF^\theta_i(E) = E[g(s)|l, \hat{\theta}]E[h(\theta)|l, \hat{\theta}]$, where the first equality comes from the fact that any evidence $E \in E(l, \hat{\theta})$ implies likelihood ratio $l$, so $l$ can be removed from the conditioning variables; the second equality comes from independence of $s$ and $\theta$ conditional on $E$ and $\hat{\theta}$; and the third equality comes from the same logic as the first one. Comparing the first and last expressions
Now suppose that a defendant of type $\theta$ reports $\hat{\theta}$ and that other players follow the strategies prescribed by the equilibrium $\sigma$ of the initial mediated process. By construction, the defendant will face the same sentence lottery under the scheme $\hat{s}$ as he was facing under scheme $s$ by reporting $\hat{\theta}$ and, hence, the same utility distribution. Therefore, the sentencing scheme $\hat{s}$ is truthful, because truth telling was by assumption optimal in the mediated process with sentencing scheme $s$.

Conditional on the defendant telling the truth and other players following the initial equilibrium $\sigma$, the distribution of welfare is the same as under the initial scheme, because by assumption the realized welfare function depends only on the realized sentence and the defendant’s type, and we have showed that the sentence distribution conditional on the defendant’s type is the same as before. Similarly, the distribution of the social cost is the same as before, because the distribution of players’ realized actions is unchanged.

The fact that the $\hat{s}$-modification is feasible in the sense of Definition 2 is trivial since the distribution of $E$ has not been modified.

There remains to check that $\hat{s}$ is statistically independent of $\theta$ conditional on $\hat{\theta}$ and $\ell(E, \hat{\theta})$. This is a direct consequence of fact that $s$ is independent of $\theta$ conditional on $\hat{\theta}$ and $E$ and the earlier observation that the distribution of $E$ conditional on $\hat{\theta}$ and the event $\ell(E, \hat{\theta}) = l$ is independent of $\theta$ for any fixed $l$. \[\blacksquare\]

We now state the main definition and result of this section.

**Definition 3** A class $C$ of judicial mechanisms is **rich** if the following holds:

1) All judicial mechanisms in $C$ are regular;

2) All mediated processes are acceptable;

For any mediated process associated with $C$, type $\theta$, and report $\hat{\theta}$, let $F_{\hat{\theta}}^{\theta}$ denote the distribution of the variable $t = \frac{\ell(E, \hat{\theta})}{1 + \ell(E, \hat{\theta})}$ conditional on true type $\theta$ and report $\hat{\theta}$, and let $\mathcal{F}$ denote the set of all tuples $(F_{\hat{\theta}}^{\theta})_{\theta, \hat{\theta} \in \{g, i\}}$.

A **direct judicial mechanism** is a single-agent game in which the defendant reports a type $\hat{\theta}$, a signal $t \in [0, 1]$ is realized, and the signal leads to a sentence $s$ whose distribution is given by a sentencing scheme $s(\hat{\theta}, t) \in \Delta([0, \bar{s}])$. The signal $t$ has a distribution that depends on $\theta$ and $\hat{\theta}$. The outcome of a direct judicial mechanism consists of i) a realized utility for the defendant, which depends only on the realized sentence, ii) a realized welfare, which depends only on the realized sentence and $\theta$, and iii) a realized social cost, which depends on $\theta$, $\hat{\theta}$, and $t$. A direct judicial mechanism is **truthful** if truth telling is optimal for the defendant.

The results obtain the in this section imply the following theorem.

**Theorem 4** Suppose that the class $C$ of judicial mechanisms is rich and satisfies Assumptions 3 and 4. Then, the following holds:

1) Any judicial process is associated with a truthful, direct judicial mechanism, for which the distribution tuple of $t$ lies in $\mathcal{F}$;

2) The distributions of all tuples in $\mathcal{F}$ are nonatomic and have full support;

3) Given any direct judicial mechanism with distribution tuple $F$ associated with a judicial process and any truthful sentencing scheme, there is another direct judicial mechanism with this sentencing scheme and the same tuple $F$, and social cost function as in the original direct judicial mechanism;

Part 3) of Theorem 4 shows that optimizing over all judicial processes associated with $C$ is welfare equivalent to optimizing over all direct judicial mechanisms with signal tuples in $\mathcal{F}$. Theorem 4 thus provides a foundation for Assumption 1.

---

in this sequence of equalities then shows that $s$ and $\theta$ are independent conditional on $l$ and $\hat{\theta}$.
References


