Judicial Mechanism Design

Ron Siegel and Bruno Strulovici*

March 2018

Abstract

This paper proposes a modern mechanism design approach to study welfare-maximizing criminal judicial processes. We provide a framework for reducing a complex judicial process to a single-agent, direct-revelation mechanism focused on the defendant, and identify a commitment assumption that justifies this reduction. We identify properties of a generically unique class of optimal mechanisms for two notions of welfare distinguished by their treatment of deterrence. These mechanisms 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

From the time of his arrest to the adjudication of his case, a criminal defendant is at the center of a complex process aimed at determining his guilt and the appropriate sentence. This process involves many stages and actors, and typically includes plea bargaining with a prosecutor, search for evidence by investigators, examination and cross-examination of the defendant and witnesses, and deliberations

*We thank Daron Acemoglu, Robert Burns, Andy Daughety, Eddie Dekel, Richard Holden, Fuhito Kojima, Adi Leibovitz, Paul Milgrom, Wojciech Olszewski, Jennifer Reinganum, Marciano Siniscalchi, Kathy Spier, Jean Tirole, Abe Wickelgren, Leeat Yariv for their comments and Jennifer Reinganum for her discussion at the NBER Summer Institute Law and Economics Workshop (2015). The paper benefited from the reactions of seminar participants at UC Berkeley, Seoul National University, the NBER, the World Congress of the Econometric Society, the Harvard/MIT Theory workshop, Caltech's NDAM conference, the 2017 HKUST Workshop on Law and Economics, the University of Warwick, Oxford University, and Bocconi University. David Rodina provided excellent research assistance. Strulovici gratefully acknowledges financial support from an NSF CAREER Award (Grant No. 1151410) and a fellowship form the Alfred P. Sloan Foundation. Siegel: Department of Economics, The Pennsylvania State University, University Park, PA 16802, rus41@psu.edu. Strulovici: Department of Economics, Northwestern University, Evanston, IL 60208, b-strulovici@northwestern.edu.

that lead to a verdict and a sentence. Much of the existing literature focuses on different aspects of this process, especially on plea bargaining, the standard for conviction, and the severity of punishment. Grossman and Katz (1983) study the informational value of plea bargaining, and assume that rejecting a plea automatically leads to a lottery over two verdicts, whose associated sentences and probabilities as a function of the defendant's guilt are given exogenously. Reinganum (1988) considers a prosecutor who privately knows the strength of the case (the probability of a guilty verdict at trial), and analyzes the signaling game that results from plea bargaining. Baker and Mezzetti (2001) focus on the possibility of evidence gathering by the prosecutor if the plea bargain is rejected. Kaplow (2011) endogenizes the probabilities of the two possible verdicts in a setting without plea bargaining. Daughety and Reinganum (2015a,b) note that introducing a third, intermediate verdict can affect the social stigma experienced by a defendant and improve welfare.[1] Kaplow (2017) considers the optimal timing for dropping a prosecuted case in a multi-stage process, each stage of which is costly and partially informative regarding the defendant's guilt.

This paper investigates from a welfare perspective a broad class of judicial processes that determine the defendant's guilt and appropriate punishment, and identifies properties of the optimal ones. The analysis provides novel insights into a number of features of existing judicial systems, including plea bargaining, binary verdicts, and *beyond a reasonable doubt* as the conviction criterion, without assuming any of these features at the outset. Like the aforementioned works, our analysis focuses on the defendant, but it follows a modern mechanism design approach to identify properties of the welfare-maximizing mechanisms among the mechanisms that focus solely on the defendant.

Our contribution is threefold. First, we describe how to reduce complex multi-actor, multi-stage judicial processes to direct revelation mechanisms that involve only the defendant, and we identify precise commitment and informational assumptions that justify this reduction. Second, we identify properties of the optimal mechanisms for interim and ex-ante welfare objective functions, which differ by their treatment of deterrence. Finally, we compare our findings to existing features of the criminal justice system in the United States, such as the use of binary verdicts and plea bargains, and also to more subtle features such as the requirement that jurors ignore information not presented at trial and the use of an adversarial system to seek and present evidence.

Our first contribution of reducing the judicial process to a single-agent mechanism is necessary to clarify the scope of our mechanism design approach, which is entirely focused on the defendant. The reduction aims to answer a simple question: what class of mechanisms should we consider in our search for the optimal ones? Our starting point here is to assume that a defendant enters the mechanism

[1] In Siegel and Strulovici (2015), we propose a systematic study of multi-verdict systems in the absence of plea bargaining.

at the time of his arrest (so individual rationality is not required at this stage).[2] Next, to reduce the judicial process to a simple mechanism, we note that standard mechanism simplifications based on the revelation principle and mediation (Myerson 1979, 1983, 1986) would deliver a mechanism that involves *all* actors of the process. Such a mechanism would be too complex for our purpose. To further simplify the analysis and obtain a tractable set of feasible mechanisms, our fundamental assumption is that any information about the defendant's guilt that is generated by a given mechanism can also be generated by other mechanisms that differ only in the sentences they impose on the defendant. Versions of this assumption often appear in the law and economics literature without a micro foundation. We interpret this assumption as a commitment assumption concerning the other actors of the judicial system, and explain how existing features of the criminal justice system in the United States are consistent with this assumption. The reduction to single-agent mechanisms focused on the defendant, which is explained in Section 4 and performed in Appendix A, helps clarify, for instance, that allowing for asymmetric information on the part of the prosecution and other actors (as in Reinganum 1988), breaking the information acquisition process into multiple, possibly endogenously determined steps (as in Kaplow 2017), or allowing the prosecutor to drop the case in an intermediate step (as in Daughety and Reinganum 2015b) does *not* affect the qualitative properties of the welfare-maximizing sentencing schemes. In particular, our commitment assumption does not rule out multi-stage judicial processes or private information held by various actors of the judicial process. The reduction also achieves a secondary objective, which is to show that potentially complex evidence regarding the defendant's guilt can be reduced to a one-dimensional signal representing the likelihood of guilt of the defendant.[3]

Our second contribution is to characterize welfare-maximizing sentencing schemes that are part of the optimal single-agent mechanisms focused on the defendant. We consider two notions of welfare. The simpler one is *interim welfare*, which describes society's trade-off between Type I and Type II errors, i.e., convicting an innocent defendant vs. acquitting a guilty one, and how the severity of these errors depends on the sentence given to the defendant.[4] The second notion is *ex-ante welfare*, which

[2] The properties of optimal mechanisms uncovered in this paper hold even if the designer can influence or optimize over pre-arrest stages, such as the level of law enforcement effort, as in Becker (1968), because these properties must hold for any given design of the pre-arrest stages. Since the consideration of these early stages would complicate the exposition of the analysis without affecting its results, we omit these stages from our analysis.

[3] The issue is to show that, under general conditions, welfare-maximizing sentencing schemes do not depend on the evidence per se, but only on the likelihood ratio that it implies. The issue is not trivial when multiple evidence outcomes imply the same likelihood ratio. The key is to show that the distribution of evidence conditional on achieving a particular likelihood ratio is independent of the defendant's true type. This result is established in Appendix A.

[4] We do not take a stand on the considerations that underlie the weight put on these errors. Our analysis applies regardless of how society weights retribution and incapacitation motives for jailing guilty defendants and how abhorrent

takes into account the number of crimes committed by adding to the previous considerations the effect of sentencing schemes on deterrence.

Formally, we consider direct, single-agent mechanisms, which map signals acquired about the defendant's guilt (arising from the actions of the defendant and other actors that produce evidence) into lotteries over sentences within a fixed interval $[0, \bar{s}]$, where $\bar{s}$ is the highest admissible sentence. To identify characteristics of the optimal mechanisms, we exploit distinctive features of judicial systems. First, transfers are not allowed: "allocations" (i.e., sentences) constitute the designer's only instrument to distinguish between guilty and innocent defendants. Second, the objectives of the defendant and the social planner may be aligned or misaligned, depending on whether the defendant is innocent or guilty. In particular, one may view the social welfare associated with an innocent defendant as being proportional to the utility of such a defendant. Intuitively, the best way to evaluate the social loss incurred from a positive sentence given to an innocent defendant is to look at how the defendant experiences it.[5] More generally, arguments relying on quasi-linear preferences do not apply: for example, the interim welfare associated with a guilty defendant may increase in the sentence up to some "ideal" punishment point, and then decrease in the sentence, since the sentence is then excessive. Similarly, the welfare loss from punishing an innocent defendant may be convex and increasing in the sentence, reflecting the view that it is abhorrent to impose even a short sentence to an innocent defendant. Third, the state of the world is binary: the defendant is either guilty or innocent.[6] In particular, one may without loss of generality order signals about the defendant's guilt according to their likelihood ratios, i.e., how likely they are to be produced by an innocent versus a guilty defendant, which can be interpreted as how incriminating these signals are.

Under these assumptions, welfare-maximizing sentencing schemes have strikingly simple features. First, the optimal mechanism when a defendant does not admit guilt is as follows: If the evidence is weak enough—below some likelihood-ratio threshold—the defendant is acquitted (i.e., receives a null sentence).[7] If it exceeds the threshold, the defendant receives the largest admissible sentence. These

the jailing of innocent defendants is perceived to be. Deterrence is a different consideration, and we examine it explicitly in our analysis of ex-ante welfare.

[5] This assumption is commonplace since Grossman and Katz (1983), who base it on the constitutional mandate to protect the innocent. We later relax the assumption and show that our results hold as long as the welfare function exhibits less risk aversion than the defendant's utility function.

[6] This is a common assumption in the literature, which we also make here. In reality, the defendant's private information need not be binary: he may face multiple counts or have private information at the time of his arrest about the evidence that may be uncovered subsequent to his arrest. We abstract from such complications in the present paper, which provides a relatively general mechanism design analysis when the state of the world is binary.

[7] A null sentence also arises when the prosecutor drops the case, as in Daughety and Reinganum (2015b).

features of the optimal mechanisms hold for both interim and ex-ante welfare objectives. In particular, these extreme, binary sentences are optimal even when deterrence is excluded from the social objective.[8] Second, we find that the optimal sentencing scheme for a defendant who admits guilt is either a fixed sentence, reminiscent of the plea sentence studied by Grossman and Katz (1983), or a lottery over two sentences, whose distribution is *independent* of the defendant's actual guilt.[9] Such a two-point lottery is never optimal in the setting of Grossman and Katz (1983).[10] Moreover, when the welfare associated with a guilty defendant is single peaked, with an ideal (from an interim perspective) sentence of $\hat{s}$, the two points of the optimal lottery lie on the same side of $\hat{s}$. Thus, for instance, if deterrence is a major concern, a defendant admitting guilt either receives a very severe (above the ideal) sentence, or a more moderate, but still severe, sentence between the ideal and the very severe sentence. Finally, when the welfare function pertaining to a guilty defendant is concave, we show that the interim-optimal sentencing scheme for a defendant who admits guilt is always a single sentence.[11]

These results may be thought of as implementing the following procedure. First, the defendant is offered a plea bargain, which is a punishment that is independent of any additional evidence regarding his guilt. If the defendant accepts the plea bargain, the case is adjudicated. If he rejects the plea bargain, he goes to trial, during which evidence regarding his guilt is generated.[12] At the conclusion of the trial he is either acquitted or convicted. This outcome is determined by an evidence threshold, so that he is convicted if and only if the evidence is sufficiently incriminating. The punishment following a conviction is severe relative to the plea bargain, whereas an acquittal leads to no punishment.

These features are reminiscent of criminal trials in the United States. Plea bargains are a frequent outcome of criminal proceedings; trials usually end in an acquittal or a conviction, with the criterion for a conviction being "beyond a reasonable doubt;" acquittals carry no punishment, whereas a conviction typically leads to a punishment more severe than a plea bargain would. We emphasize that *our analysis does not assume binary verdicts, an evidentiary conviction threshold, or no punishment following an*

[8] This underlines the fact that the optimality of extreme sentencing schemes here is completely different from the optimality of extreme schemes pointed out by Becker (1968), which are due to deterrence and enforcement cost considerations.

[9] We show that the mechanism described here is generically uniquely optimal for a concept of genericity applied to the set of all possible welfare functions. Studying genericity in this setting is nontrivial because the space of welfare objectives that we consider is infinite dimensional. From this perspective, our analysis contributes the mechanism design analysis over infinite dimensional spaces, and relates to notions of genericity studied by Anderson and Zame (2001) and Jehiel et al. (2006).

[10] Their analysis can be interpreted as optimizing over a restricted class of mechanisms to maximize interim welfare.

[11] The optimality of a sentence that is independent of the signal is not due to cost savings or limited resources on the part of the prosecutor. Such considerations would further reinforce our findings.

[12] At this point the case may also be dropped if the evidence is sufficiently weak, as in Daughety and Reinganum (2015b).

acquittal. These features, as well as plea bargains, emerge as features of the optimal mechanism.

Interim and ex-ante optimal mechanisms thus have qualitatively similar features. When utility and welfare functions are concave, one difference concerns the punishment associated with a plea bargain. This punishment is deterministic in any interim-optimal mechanism, but may be random in an ex-ante optimal mechanism. When the punishment is random, it takes one of two values. Moreover, a random plea is more likely to be optimal for more serious crimes.[13] 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,[14] 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.

While the optimal mechanisms share many features with existing criminal justice systems, two important and related differences emerge from our analysis. First, the optimal mechanisms are fully separating: guilty defendants accept the plea bargain and innocent defendants reject the plea bargain and go to trial.[15] In reality, most of the defendants who go to trial are in fact guilty. Second, in real trials evidence is used to determine the defendant's guilt, but evidence in the optimal mechanisms serves this purpose only off the equilibrium path, since in equilibrium only innocent defendants go to trial. The role of evidence is instead to incentivize guilty defendants to accept the plea bargain. This requires commitment on the part of the designer, since in equilibrium all defendants who are convicted are in fact innocent.

This feature of the optimal mechanisms and the difference in the role that evidence plays are mitigated by considering mechanisms that are "close" to the optimal ones and achieve similar welfare, as we explain in Section 5. In the optimal mechanisms, guilty defendants are indifferent between accepting the plea bargain and going to trial. Suppose instead that a small fraction of them goes to trial. Then, evidence recovers its role in determining the defendant's guilt, as well as incentivizing most guilty defendants to accept the plea bargain. If the prior that the defendant is guilty is relatively high, it suffices that a small fraction of guilty defendants go to trial for most convicted defendants to

[13] More precisely, we show that if the interim welfare assigned to a guilty defendant and the defendant's utility are both concave in the sentence given to the defendant, then a two-point lottery is optimal in terms of ex-ante welfare only if the support of this lottery lies above the ex post optimal sentence, i.e., the ideal punishment of a guilty defendant, absent any deterrence consideration. In general, the optimality of two-point lotteries follows from a concavification argument reminiscent of the Bayesian persuasion literature (Kamenica and Gentzkow 2011) and the dynamic contracting literature (Spear and Srivastava 1987).

[14] A recent, widely publicized example of this case concerns Jared Fogle, a former Subway spokesman who accepted a plea bargain and subsequently received a sentence that exceeded the one outlined in the plea bargain.

[15] Such separation arises in many existing papers, including in Grossman and Katz's (1983) main model.

be guilty. Thus, our analysis suggests that the combination of plea bargains and trials with binary verdicts can generate high welfare, and also shows that evidence plays an important part in making plea bargains attractive, in addition to its role in determining the defendant's guilt during a trial.

Our results also shed light on features of the criminal justice system that may be viewed as implementing a form of commitment. Indeed, the commitment required to implement the optimal mechanism can be mapped back to the reduction performed in the first stage of our analysis. This reduction assumed that whatever signals are produced in one mechanism can also be produced in other mechanisms that impose different sentences on the defendant. Thus, the actors of the judicial process who generate the signals in one mechanism must also be incentivized to generate these signals in other mechanisms. This is consistent with an adversarial system, in which regardless of the sentencing scheme different parties are incentivized to look for incriminating and exculpatory evidence. Similarly, the instruction given to jurors to focus on the evidence presented at trial and ignore, in particular, any inference from the fact that a plea bargaining procedure may have preceded the trial, as well as the rule that prevents any part of this procedure from being disclosed during the trial, are consistent with the commitment assumption made in this paper: the signals generated during the trial regarding the defendant's guilt guide the verdict independently of any signal that the defendant sent about his guilt during the plea bargaining procedure. The analysis thus provides a justification for these important features of the criminal justice system.

2 Judicial Mechanisms

Suppose a crime has been committed, and a suspect is arrested and charged. The criminal justice machinery is then set in motion, leading to a judicial decision and a sentence. In reality, this processes does not give the full information outcome, which would be punishing only the guilty, and at the ex-post optimal level. This is because, at a minimum, the defendant knows whether he is guilty but the judicial system does not. The judicial process produces evidence that can be used to determine the defendant's guilt, and the technology for producing such evidence is limited and given exogenously. We model this process as a game with incomplete information. The defendant is guilty with prior probability $\lambda \in (0,1)$ and innocent with probability $1-\lambda$, and is privately informed about his guilt $\theta \in \Theta = \{i, g\}$.[16] The game may involve additional players (the police, prosecutors, attorneys, jurors, etc.) who may have private information and take various actions that produce evidence. The game

[16] 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.

concludes with a sentence that is a function of the history of players' actions and produced evidence and, possibly, exogenous random shocks. The game may be quite complex, but several properties of the judicial process can be studied by focusing on the strategic behavior of the defendant while summarizing the produced evidence by a signal $t \in [0,1]$ regarding the defendant's guilt.

More precisely, given a profile of strategies of the players, we write another, single-player game, which we refer to as a *direct judicial mechanism* or, simply, a *mechanism.* The player is the defendant, the set of actions for each of his types is the set of types, and the outcome resulting from action $\hat{\theta}$ is the same as the outcome of the original game if the defendant played as if he were of type $\hat{\theta}$. By a logic similar to that of the revelation principle (Myerson 1979), the defendant reports his type truthfully, leading to a truthful mechanism.[17] The reduction of the multi-player game to a (single-player) truthful mechanism involves several subtleties. A detailed description is given in Section 4 and Appendix A. The definitions and results in Appendix A are not necessary for understanding the analysis of the optimal mechanisms in Section 3.

Thus, the mechanism is characterized by distributions $F_{\theta}^{\hat{\theta}}$ of signals t, where $\theta \in \Theta$ is the defendant's type and $\hat{\theta}$ is his reported type, and a (possibly degenerate) sentence lottery $S\left(t, \hat{\theta}\right) \in \Delta([0, \bar{s}])$, where $\bar{s}$ is the highest allowable sentence for the crime and $\Delta([0, \bar{s}])$ is the set of lotteries over possible sentences.[18] For notational clarity, we use hats to denote the reported types: $\hat{\imath}$ and $\hat{g}$. The sentencing scheme S assigns a (possibly random) sentence based on the signal and the defendant's reported type. The possible dependence of distributions $F_{\theta}^{\hat{\theta}}$ on the defendant's type and reported type captures the possibility that the defendant's strategy in the original game can affect the distribution of evidence produced regarding his guilt both directly and by affecting the actions of the other players, as well as the possibility that his actual guilt affects the distribution of evidence (for example, an innocent defendant is less likely to have been at the crime scene and therefore less likely to have been seen by eye witnesses).

Given the defendant's report $\hat{\theta}$, the relevance of the signal t for determining the defendant's guilt is entirely captured by the likelihood ratio associated with it, that is, the probability that t has been generated by a guilty versus an innocent defendant. We thus assume without loss of generality that the signal is one dimensional and ordered by its likelihood ratio. More precisely, we assume that distributions $F_{g}^{\hat{\theta}}$ and $F_{i}^{\hat{\theta}}$ have strictly positive densities $f_{g}^{\hat{\theta}}(t)$ and $f_{i}^{\hat{\theta}}(t)$ over the support $T = [0,1]$ that

[17] If it is profitable for the defendant to misreport his type, then the corresponding deviation is profitable for the defendant in the original game.

[18] Our analysis focuses on a particular crime, so $\bar{s}$ can vary across crimes. In addition, $\bar{s}$ can depend on the evidence and information collected up to the defendant's arrest. The same is true of the welfare function and prior λ introduced below.

satisfy the strict monotone likelihood ratio property (MLRP): the density ratio $f_g^{\hat{\theta}}(t)/f_i^{\hat{\theta}}(t)$ is strictly increasing in t.[19]

To summarize, a mechanism M is a pair (F, S), where $F = \left(F_i^{\hat{i}}, F_g^{\hat{i}}, F_i^{\hat{g}}, F_g^{\hat{g}}\right)$ is a tuple of signal distributions and S is a sentencing scheme. The distribution $F_\theta^{\hat{\theta}}$ of signals, which corresponds to a defendant of type θ who reports $\hat{\theta}$, has density $f_\theta^{\hat{\theta}}$. Distributions $F_i^{\hat{\theta}}$ and $F_g^{\hat{\theta}}$ satisfy the MLRP. The mechanism is truthful if the defendant (optimally) reports his type truthfully.

Our objective is to study the social welfare of different truthful mechanisms. We will consider two notions of welfare, interim and ex ante. The interim notion captures social welfare after the crime has been committed and a defendant whose guilt is uncertain is apprehended. Ex-ante welfare also takes into account the number of crimes committed and the possibility that for any particular crime no suspect is apprehended.

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 θ.[20] Any monetary cost of imposing the sentence, such as the cost of incarceration, is included in W. We assume that $W(s, i)$ strictly decreases in s, and that $W(s, g)$ is continuous.[21] For some results (though not the main one, Theorem 2), we will assume that $W(s, g)$ is also single peaked in s with peak $\hat{s} \in (0, \bar{s})$. The sentence $\hat{s}$ is the socially optimal one when the defendant is guilty. 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."[22] The welfare of imposing a sentence s on a defendant who is guilty with probability λ 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

[19] The existence of strictly positive densities does entail some loss of generality. In particular, we rule out atoms, i.e., a positive measure of signals with the same likelihood ratio, and assume 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.

[20] This can be thought of as ex post welfare, since the crime has been committed and the defendant's type enters as an argument.

[21] Continuity is assumed for expositional simplicity.

[22] 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.

a (possibly) random sentence $\tilde{s}$ on a defendant of type θ.[23] We denote by $C\left(F_{\theta}^{\hat{\theta}}\right) \geq 0$ the expected cost of the judicial process associated with the signal distribution $F_{\theta}^{\hat{\theta}}$, and assume for simplicity that the cost is additively separable from W. Thus, given a truthful mechanism $M = (F, S)$, the resulting interim welfare is

$$\lambda\left(\left(\int_0^1 W(S(t,\hat{g}),g) f_g^{\hat{g}}(t)dt\right) - C\left(F_g^{\hat{g}}\right)\right) + (1-\lambda)\left(\left(\int_0^1 W(S(t,\hat{\imath}),i) f_i^{\hat{\imath}}(t)dt\right) - C\left(F_i^{\hat{\imath}}\right)\right). \quad (1)$$

This formulation of welfare does not take into account the effect of the mechanism on the number of crimes committed. We study this effect by considering ex-ante welfare, which includes the deterrent effect of the mechanism. Deterrence plays a key role in the seminal economic analyses of criminal justice systems of Becker (1966) and Stigler (1970), as well as in recent ones (Kaplow 2011). We assume that if a crime is committed at most one individual is prosecuted for it.[24] Given a crime, the probability that the individual who committed the crime is prosecuted for it is non-negligible. The probability that the individual who is prosecuted for the crime is in fact innocent is also non-negligible. But in a large society, the probability that any *particular* innocent individual will be prosecuted for that particular crime is infinitesimal. For expositional convenience,[25] we assume that individuals treat this latter probability as 0 when they contemplate whether to commit a 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. Again for expositional convenience, we treat π_g as exogenous.[26] 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.[27] Thus, $W(\cdot, i) = u(\cdot)$, where $u(\cdot)$ is strictly decreasing, continuous, and normalized such that $u(0) = 0$.[28] 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 D).

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

[24] 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.

[25] The welfare-improving mechanisms constructed in Section 3 keeps the expected utility of a guilty defendant unchanged and increases the expected utility of an innocent defendant. If the probability that any given innocent individual is prosecuted is treated as strictly positive, the constructed mechanism would thus have the additional benefit of increasing deterrence by increasing the utility differential between an innocent defendant and a guilty one.

[26] 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.

[27] This assumption appears in Grossman and Katz' (1983) analysis of plea bargaining.

[28] Continuity is assumed for expositional simplicity.

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 truthful mechanism $M = (F, S)$, an individual commits the crime if

$$b + \pi_g \left(\int_0^1 u(S(t, \hat{g})) f_g^{\hat{g}}(t) dt \right) > 0. \tag{2}$$

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

$$H(M) = 1 - G_b \left(-\pi_g \left(\int_0^1 u(S(t, \hat{g})) f_g^{\hat{g}}(t) dt \right) \right). \tag{3}$$

Given that each realized crime entails a social harm h, the ex-ante social welfare is

$$H(M) \left[\pi_g \left(\left(\int_0^1 W(S(t, \hat{g}), g) f_g^{\hat{g}}(t) dt \right) - C\left(F_g^{\hat{g}}\right) \right) + \pi_i \left(\left(\int_0^1 W(S(t, \hat{\imath}), i) f_i^{\hat{\imath}}(t) dt \right) - C\left(F_i^{\hat{\imath}}\right) \right) - h \right], \tag{4}$$

where $\pi_i > 0$ is the probability that the prosecuted individual is innocent.[29] 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.[30]

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 allows us to recover (1).

3 Optimal judicial mechanism

Our objective is to identify properties of optimal judicial mechanisms when the judicial authority has full commitment, and then compare these mechanisms to existing judicial procedures. As we will see, despite its strength, the full-commitment assumption delivers optimal judicial mechanisms that resemble existing judicial procedures.

The standard approach to optimal mechanism design is to consider all direct mechanisms, that is, all mappings from reported types to possible outcomes, and to optimize over the subset of all truthful (incentive compatible) direct mechanisms. A difficulty in our setting is that not all direct mechanisms

[29] As with π_g, for expositional simplicity we will take π_i to be exogenous.

[30] 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.

are feasible. This is because the possible distributions $F_\theta^{\hat{\theta}}$ are determined by the (unmodeled) available evidence-gathering technology and the equilibrium strategies of the players in the possible (unspecified) original games.[31] We overcome this difficulty by focusing the optimization on the sentencing scheme S, given a distribution tuple F. Our main assumption is that if a distribution tuple F is part of a truthful feasible mechanism, then the designer can choose any sentence mapping as a function of the defendant's report and the generated signal and, provided that incentive compatibility is maintained, obtain another truthful feasible mechanism.

Assumption 1 *If mechanism (F, S) is feasible and truthful, then for any sentencing scheme $(t, \hat{\theta}) \mapsto \tilde{S}\left(t, \hat{\theta}\right) \in \Delta([0, \bar{s}])$ that maintains truthfulness, the mechanism $\left(F, \tilde{S}\right)$ is also feasible (and truthful).*

Assumption 1 formalizes our notion of full commitment. It captures the idea that changing the sentencing 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. 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 without being justified. Section 4 and Appendix A provide a novel micro foundation for the assumption. Section 5 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 feasible ones, first for interim welfare and then for ex-ante welfare. We compare and discuss the features of these mechanisms in Section 5.

3.1 Interim welfare

A judicial mechanism is interim optimal if it maximizes interim welfare (1) among all truthful feasible mechanisms, given the prior probability λ that the defendant is guilty. Although our main objective is to identify properties of ex-ante optimal mechanisms, considering first the case of interim-optimal mechanisms allows us to disentangle deterrence from other welfare considerations and makes the arguments of the proof easier to follow.

For our first result, which describes some properties of any interim-optimal mechanism, we assume that he welfare function $W(\cdot, g)$ is single-peaked. The peak $\hat{s} \leq \bar{s}$ is the socially optimal sentence conditional on facing a guilty defendant, where $\bar{s}$ is the maximal allowable sentence. This implies that

[31] 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.

it is never interim optimal to assign a sentence higher than $\hat{s}$. We also assume that the defendant and society when facing a guilty defendant are risk averse. Taken together, these assumptions lead to sharp predictions and simplify the analysis of interim-optimal mechanisms. The same assumptions do not achieve this for ex-ante optimal mechanisms. This is because deterrence may optimally leads to sentences higher than $\hat{s}$, in which case risk aversion does not simplify the analysis.[32] We drop all these assumptions in Theorem 2, which identifies properties of ex-ante optimal mechanisms and requires a different proof. Theorem 2 also applies to interim-optimal mechanisms, but the characterization is less sharp than the one in Theorem 1.

Theorem 1 *Suppose that $W(\cdot, g)$ is single-peaked at $\hat{s}$, that $W(\cdot, g)$ and $u(\cdot)$ are concave, and that at least one of them is strictly concave.*[33] *Then, any interim optimal mechanism has the following properties:*[34]

(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.

Moreover, any mechanism that fails to have the above properties can be improved for all priors λ 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. While it may seem realistic, intuitive, and perhaps comforting to give a null sentence to defendants against whom little evidence was produced, the interim optimality of such a sentence is by no means obvious, and generally fails to hold.[35]

[32] Most notably, the fact that the punishment of a guilty defendant is deterministic in any interim-optimal mechanisms does not generally hold for ex-ante optimal mechanisms even when the defendant and society are risk averse.

[33] 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 μ in $(0, 1)$ and $s \neq s'$.

[34] 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 2.

[35] In fact, the interim-optimal sentences following an acquittal are strictly positive when pleas are not allowed. See Siegel and Strulovici (2017) for this point and a generalization to multi-verdict trials.

We also note that the signal is not used by the mechanism to determine the sentence if the defendant pleads guilty. Intuitively, this is as if pleading guilty prevents a trial and the evidence that it generates. More formally, a signal-independent sentence is used even if signal distributions $F_i^{\hat{g}}$ and $F_g^{\hat{g}}$ are informative about the defendant's guilt. 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 the concavity and single-peak assumptions). As shown in the next section, this result generally fails when deterrence is taken into account.[36]

Finally, the last statement of Theorem 1 shows that the argument 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 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)).[37]

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 the defendant reports his type truthfully, the signal is not needed to determine guilt in equilibrium. 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 preventing 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 being risk averse. The sentence for the guilty defendant must be constant because both he and society are risk averse; moving from a random to a constant sentence for the guilty defendant thus 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.[38]

[36] Grossman and Katz (1983) focused on interim welfare and did not consider deterrence.

[37] The 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.

[38] This last point is not generally true for ex-ante optimal mechanisms, because decreasing the sentence for the guilty leads to more crime. The proof of Theorem 2 avoids this issue by maintaining the same utility for the guilty and using a concavification argument.

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

Consider a truthful feasible mechanism $M = (F, S)$. We modify M in a way that maintains feasibility and incentive compatibility and increases interim welfare. To guarantee feasibility of our modification, we do not change the signal distributions F, and instead construct an improvement that concerns only the sentencing scheme. Assumption 1 ensures the feasibility of such an improvement.

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

$$\int_0^1 u(\tilde{S}(t, \hat{\imath})) f_i^{\hat{\imath}}(t) dt = u(0) F_i^{\hat{\imath}}([0, \bar{t}]) + u(\bar{s}) F_i^{\hat{\imath}}([\bar{t}, 1]) = \int_0^1 u(S(t, \hat{\imath})) f_i^{\hat{\imath}}(t) dt. \tag{5}$$

Because the signal's distribution has no atoms, such a cutoff always exists. Rearranging (5) yields

$$\int_0^1 [u(S(t, \hat{\imath})) - u(\tilde{S}(t, \hat{\imath}))] f_i^{\hat{\imath}}(t) dt = 0. \tag{6}$$

The function $t \mapsto u(S(t, \hat{\imath})) - u(\tilde{S}(t, \hat{\imath}))$ crosses 0 once from below, since $u(S(t, \hat{\imath}))$ lies in the interval $[u(\bar{s}), u(0)]$ for all t, while $u(\tilde{S}(t, \hat{\imath}))$ equals $u(0)$ for $t \leq \bar{t}$ and jumps down to $u(\bar{s})$ at $t = \bar{t}$. Finally, the density ratio $f_i^{\hat{\imath}}(t)/f_g^{\hat{\imath}}(t)$ is decreasing in t, by MLRP. A standard result in comparative statics analysis[39] then implies that

$$\int_0^1 [u(S(t, \hat{\imath})) - u(\tilde{S}(t, \hat{\imath}))] f_g^{\hat{\imath}}(t) dt \geq 0. \tag{7}$$

Therefore, given the signal distribution $F_g^{\hat{\imath}}$, a guilty defendant weakly prefers the initial sentence function $S(\cdot, \hat{\imath})$ to the new one, $\tilde{S}(\cdot, \hat{\imath})$. By incentive compatibility of mechanism M, a guilty defendant weakly prefers signal distribution $F_g^{\hat{g}}$ with sentence function $S(\cdot, \hat{g})$ to signal distribution $F_g^{\hat{\imath}}$ with sentence function $S(\cdot, \hat{\imath})$. Thus, incentive compatibility continues to hold when $S(\cdot, \hat{\imath})$ is replaced with $\tilde{S}(\cdot, \hat{\imath})$.

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

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

Since u is concave and decreasing, s^{ce} is greater than the average sentence $s^a = \int_0^1 E[S(t, g)] f_g^{\hat{g}}(t) dt$: The defendant, being risk averse, prefers a fixed sentence over any lottery with the same expectation. To achieve indifference, the fixed sentence must thus be weakly higher, since the defendant dislikes higher

[39] 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.

sentences. Since $W(\cdot, g)$ is also concave, $W(s^a, g) \geq \int_0^1 W(S(t, \hat{g}), g) f_g^{\hat{g}}(t)dt$. But because $W(\cdot, g)$ decreases above $\hat{s}$ (the socially optimal sentence conditional on facing a guilty defendant), if s^{ce} is sufficiently greater than s^a, it might be that $W(s^{ce}, g) < \int_0^1 W(S(t, \hat{g}), g) f_g^{\hat{g}}(t)dt$.

Thus, to set the improving constant sentence s^g for a guilty defendant, there are two cases to consider. If s^{ce} is less than the socially optimal sentence $\hat{s}$ conditional on facing a guilty defendant, we set $s^g = s^{ce}$. Since $s^{ce} \geq s^a$ and $W(\cdot, g)$ is increasing up to $\hat{s}$, we have $W(s^{ce}, g) \geq W(s^a, g) \geq \int_0^1 W(S(t, \hat{g}), g) f_g^{\hat{g}}(t)d$, so s^g increases welfare conditional on facing a guilty defendant. If instead $s^{ce} > \hat{s}$, we set $s^g = \hat{s}$. This sentence yields the highest possible social welfare conditional on facing a guilty defendant.

Given signal distribution $F_g^{\hat{g}}$, the guilty defendant is by construction indifferent between s^{ce} and the sentence function $S(\cdot, \hat{g})$, so because $s^g \leq s^{ce}$ he prefers s^g to $S(\cdot, \hat{g})$. We have already argued that a guilty defendant prefers signal distribution $F_g^{\hat{g}}$ with sentence function $S(\cdot, \hat{g})$ to signal distribution $F_g^{\hat{\imath}}$ with sentence function $\tilde{S}(\cdot, \hat{\imath})$. Thus, he prefers sentence s^g to signal distribution $F_g^{\hat{\imath}}$ with sentence function $\tilde{S}(\cdot, \hat{\imath})$, so incentive compatibility is maintained.

If this preference is strict, we increase the cutoff $\bar{t}$ until the guilty defendant becomes indifferent between s^g and signal distribution $F_g^{\hat{\imath}}$ with sentence function $\tilde{S}(\cdot, \hat{\imath})$. This modification also increases welfare since it increases the utility of an innocent defendant. It also guarantees that an innocent defendant prefers signal distribution $F_i^{\hat{\imath}}$ with sentence function $\tilde{S}(\cdot, \hat{\imath})$ to s^g, because of the guilty defendant's indifference and MLRP (as in the first part of the proof). ■

3.2 Ex-ante welfare and deterrence

Ex-ante welfare takes into account the number of committed crimes, so any modification of a given sentencing scheme must take into account the modification's impact on deterrence. The proof of Theorem 1 suggests that this consideration need not necessarily lead to a radically different analysis. 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 commit the crime.[40] In

[40] 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 2 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.

this case, therefore, ex-ante welfare also increases. In particular, Theorem 1 identifies properties of the mechanisms that maximize ex-ante welfare among all truthful feasible mechanisms in which the certainty equivalent of a guilty defendant does not exceed $\hat{s}$.

In general, however, 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.

The next result identifies properties of the ex-ante optimal mechanisms, which maximize ex-ante welfare (4) among all truthful feasible mechanisms. The result shows that adding deterrence as a consideration 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. 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 utility from his sentence, instead of the certainty equivalent sentence, and find the sentence that maximizes social welfare and maintains the same utility for the guilty. Thus, the proof improves social welfare conditional on facing an innocent defendant and conditional on facing a guilty defendant, without changing the number of crimes committed and the identity of the perpetrators. This allows us to avoid having to find the optimal level of deterrence explicitly, and also guarantees that we do not need to be concerned about effect of the optimization on the probability of apprehension, the benefit accrued to the perpetrator from the crime, etc. This also means that the result applies to both interim optimal mechanisms and ex-ante optimal mechanisms. The proof is based on a concavification argument reminiscent of arguments used in contract theory and strategic communication. A key advantage of this approach is that single-peakedness of $W(\cdot, g)$ and concavity of the utility and welfare functions are no longer assumed.[41]

Theorem 2 *(i) In any ex-ante optimal mechanism, the innocent defendant's sentence is a step function of t, which jumps from 0 to $\bar{s}$ at some cutoff $\bar{t}$.*[42]

(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.[43] *If the defendant's sentence*

[41] The proof of Theorem 1 requires these assumptions and applies only to interim welfare, but provides sharper predictions. As we point out in the proof of Theorem 2, under the same assumptions these sharper predictions for interim welfare can also be obtained from the proof of Theorem 2, but the proof of Theorem 1 in this case is more transparent.

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

[43] Here, "generically" is understood in the sense of prevalent sets over the vector space of welfare functions. See

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) Parts (i), (ii) also hold for interim optimal mechanisms, i.e., if "ex-ante" is replaced with "interim." For part (iii), if $W(\cdot, g)$ is single-peaked at $\hat{s}$, then the two-point support lies in $[\hat{s}, \bar{s}]$. 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.

We emphasize that when a guilty defendant receives a stochastic sentence, the sentence may be chosen to be statistically independent of the signal t. That is, the sentence can be determined by a purely exogenous randomization. Alternatively, it could be based on the signal t, e.g., with $S(t, \hat{g})$ taking the higher of the two points in the support when t exceeds some cutoff (or falls below a different cutoff).

Proof. Consider a truthful feasible mechanism $M = (F, S)$. Similarly to the proof of Theorem 1, we will modify M in a way that maintains feasibility and incentive compatibility and increases ex-ante welfare.

As in the proof of Theorem 1, we replace the sentence function $S(\cdot, \hat{\imath})$ with a step function $\tilde{S}(\cdot, \hat{\imath})$ that is equal to zero below $\bar{t}$ and equal to $\bar{s}$ above it, with $\bar{t}$ chosen to make an innocent defendant indifferent between $S(\cdot, \hat{\imath})$ and $\tilde{S}(\cdot, \hat{\imath})$, so a guilty defendant weakly prefers $S(\cdot, \hat{\imath})$ to $\tilde{S}(\cdot, \hat{\imath})$. The cutoff $\bar{t}$ is now increased until the guilty defendant is indifferent between $S(\cdot, \hat{g})$ and $\tilde{S}(\cdot, \hat{\imath})$. This change increases the utility of an innocent defendant, and therefore social welfare.

We now modify the sentence function $S(\cdot, \hat{g})$ in a way that keeps the guilty defendant's expected utility, U^g, unchanged. We wish to find a sentence function $\tilde{S}(\cdot, \hat{g})$ that maximizes ex-ante social welfare subject to giving the guilty defendant utility U^g. We denote by $\tilde{M}$ the mechanism that differs from M only by replacing sentencing scheme S with sentencing scheme $\tilde{S}$. By (3), the fraction of individuals who commit the crime depends only on the guilty defendant's expected utility U^g, so $H(M) = H\left(\tilde{M}\right)$. We therefore consider the following modification of (4), which captures the ex-ante social welfare from $\tilde{M}$ without the social harm of the crime h and without the cost associated with signal distributions F, as these are the same for mechanisms M and $\tilde{M}$:

$$H(\tilde{M})\left[\pi_g\left(\int_0^1 W(\tilde{S}(t,\hat{g}),g)f_g^{\hat{g}}(t)dt\right)+\pi_i\left(\int_0^1 W(\tilde{S}(t,\hat{\imath}),i)f_i^{\hat{\imath}}(t)dt\right)\right].$$

Appendix C for a precise definition of genericity in this context.

Thus, we are looking for a sentence function $\tilde{S}(\cdot,\hat{g})$ that solves

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

subject to

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

Notice that the sentence function $\tilde{S}(\cdot,\hat{g})$ that solves this problem also increases interim welfare. To solve this problem, it is convenient to reformulate it in terms of the defendant's utility, i.e., to find a mapping from types to lotteries over utilities that solves

$$\max_{\hat{u}(\cdot)\in(\Delta([u(\bar{s}),u(0)]))^T} \int_0^1 E[\hat{W}(\hat{u}(t))] f_g^{\hat{g}}(t)dt \tag{8}$$

subject to

$$\int_0^1 E[\hat{u}(t)] f_g^{\hat{g}}(t)dt = U^g, \tag{9}$$

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

To characterize the solution of (8) subject to (9), it is useful to consider a simpler optimization problem:

$$\max_{\dot{u}\in\Delta([u(\bar{s}),u(0)])} E[\hat{W}(\dot{u})] \tag{10}$$

subject to

$$E[\dot{u}] = U^g. \tag{11}$$

Consider a function $\hat{u} \in (\Delta([u(\bar{s}),u(0)]))^T$ that satisfies (9), and let $\dot{u} = \int_0^1 \hat{u}(t) f_g^{\hat{g}}(t)dt$ be the expected utility distribution. Then $\dot{u}$ satisfies (11), and

$$\int_0^1 E[\hat{W}(\hat{u}(t))] f_g^{\hat{g}}(t)dt = E[\hat{W}(\dot{u})]. \tag{12}$$

Therefore, $\hat{u}$ is a solution of (8) subject to (9) if and only if $\dot{u} = \int_0^1 \hat{u}(t) f_g^{\hat{g}}(t)dt$ is a solution of (10) subject to (11).

We now solve for (10) subject to (11). For any U in the interval $[u(\bar{s}), u(0)]$, let

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

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

It is well-known that $\bar{W}(U^g)$ is the solution of (10) subject to (11):[44] If $\hat{W}(U^g) = \bar{W}(U^g)$, the maximal value is achieved by the constant sentence $u^{-1}(U^g)$. In this case, by (12), the optimal $\hat{u}$ is unique and achieved by the sentence function $\tilde{S}(t, \hat{g}) = u^{-1}(U^g)$, which is constant in the signal t. If $\hat{W}(U^g) < \bar{W}(U^g)$, the maximal value is achieved by randomizing between $u^{-1}\left(\underline{U}\right)$ and $u^{-1}\left(\overline{U}\right)$, where $\underline{U} = \max\left\{v < U^g : \hat{W}\left(v\right) = \bar{W}\left(v\right)\right\}$ and $\overline{U} = \min\left\{v > U^g : \hat{W}\left(v\right) = \bar{W}\left(v\right)\right\}$, with probabilities α and $1 - \alpha$ such that $\alpha\underline{U} + (1 - \alpha)\,\overline{U} = U^g$. In this case, again by (12), for any optimal $\hat{u}$ and any signal t, the support of $\hat{u}\left(t\right)$ is a subset of $\left\{\underline{U}, \overline{U}\right\}$. The induced distribution over $\left\{\underline{U}, \overline{U}\right\}$ assigns probability α to $\underline{U}$ and probability $1 - \alpha$ to $\overline{U}$. A sentence function $\tilde{S}(\cdot, \hat{g})$ generates this distribution over $\left\{\underline{U}, \overline{U}\right\}$ if and only if it assigns probability α to sentence $u^{-1}\left(\underline{U}\right)$ and probability $1 - \alpha$ to sentence $u^{-1}\left(\overline{U}\right)$. In particular, the constant stochastic sentence function (which is independent of the signal) that for every signal t assigns sentence $u^{-1}\left(\underline{U}\right)$ with probability α and sentence $u^{-1}\left(\overline{U}\right)$ with probability $1 - \alpha$ is optimal.

If W is single peaked at $\hat{s}$, then the fact that u is decreasing implies that $\hat{W}$ is single peaked at $u(\hat{s})$, which proves that the two-point support lies in $[0, \hat{s}]$ or in $[\hat{s}, \bar{s}]$, and in $[0, \hat{s}]$ if the mechanism is interim optimal.[45] If, in addition, u and $W\left(\cdot, g\right)$ are concave on $[0, \hat{s}]$, then $\hat{W}$ is concave on the utility interval $[u(\hat{s}), u(0)]$. In this case, $\hat{W}$ coincides with $\bar{W}$ for $U \geq u(\hat{s})$, so $U^g \geq u(\hat{s})$ is optimally achieved by a single sentence.[46] 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.

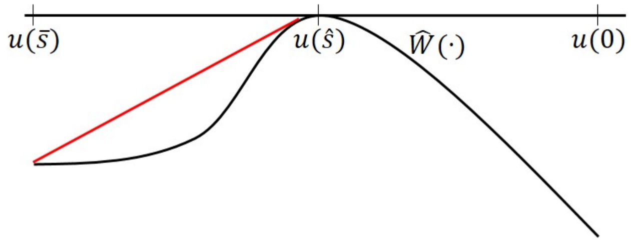

Figure 1: Lottery over sentences.

Since guilty defendants are indifferent between this sentence function and $\tilde{S}(\cdot, \hat{\imath})$, innocent defendants prefer $\tilde{S}(\cdot, \hat{\imath})$.

Appendix C proves the genericity claim in part (ii). ■

[44] 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).

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

[46] This observation provides proves the last claim in part (iv) and also provides an alternative proof for Theorem 1.

Theorem 2 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, it is useful to consider the case of concave utility u and concave and single-peaked welfare $W(\cdot, g)$. 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 $\hat{W}$ is not concave below $u(\hat{s})$, and in addition $\hat{W}(U^g) < \bar{W}(U^g)$.[47]

4 From Judicial Processes to Direct Revelation Mechanisms

This section considers the reduction from a multi-agent judicial process to a single-agent mechanism focused on the defendant, and presents the idea underlying a micro-foundation for Assumption 1, which is formalized in Appendix A. Recall that Assumption 1 states that for any feasible signal structure the designer can optimize over all truthful sentencing schemes that use this signal structure. Versions of this assumption (both implicit and explicit) are pervasive in the law and economics literature. 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.

Such assumptions are not needed in standard mechanism design settings, in which players observe their type and are then asked to report it to the mechanism before the mechanism's outcome is realized. But in the present setting, as in much of the existing law and economics literature, the signals regarding the defendant's guilt may realistically depend on the actions of various actors in the judicial system. These actors may respond to different incentives, which means that the signals available to a mechanism designer could a priori depend on his chosen sentencing scheme. This, in turn, implies that the set of mechanisms over which the designer can optimize is potentially very complex to describe. Assumption 1, and similar assumptions made in other papers, put enough structure on the set of feasible mechanisms to make the analysis tractable. Existing papers, however, do not provide a micro-foundation for these assumptions. Indeed, it is not immediately clear what properties of the multi-agent environment justify such assumptions.

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

Appendix A provides an explicit micro-foundation for Assumption 1, which we explain informally here. Intuitively, Assumption 1 entails two distinct commitment features:

1. Whatever mechanisms and strategy profiles of players other than the defendant are available to the designer, these profiles are still available if the sentences are modified, provided that the strategy of the defendant is unchanged;
2. Given an available mechanism and a strategy profile, for any modification of the sentences there is another mechanism in which the set of strategies of the defendant consists of exactly the two strategies (one for each type) used by the defendant in equilibrium in the original mechanism.

To implement the first 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. In the appendix we give a micro-foundation that is more general and 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 second feature, we postulate the existence of a mediator.[48]

Our mechanism-design approach aims to capture the most minimal incentive compatibility (IC) constraints: weak IC for defendant (weak in the sense that we are only concerned with two strategies, one for each defendant type, captured by the mediated system, and not all the strategies available to the defendant in the initial judicial system) and no IC for other actors. We also want to capture the limited evidence technology, without imposing much structure on it. These two features give us the single-agent formulation in the paper: the lack of IC for the other actors and the existence of some evidence technology give us Assumption 1.

In reality, there may well be other IC constraints, both for the defendant and for the other actors. But in our analysis we do not impose them. Thus, we optimize over what is likely a super set of the judicial processes available in reality. In the next section we compare features of the welfare-maximizing mechanisms under full commitment to existing judicial systems. We also compare Assumption 1 to institutional features of the American criminal justice system, and note that some of these features, such as the narrow focus of jurors on finding facts regardless of the sentence resulting from this finding, resemble the commitment assumption we rely on to derive the optimal mechanisms. There are also differences between the optimal mechanisms and existing judicial systems. We discuss these similarities and differences in the next section.

Another, more technical question is whether the signal can without loss of generality be taken to be one dimensional. In reality, the judicial process generates a disparate body of evidence, combined

[48] Of course, the micro foundation we provide is only one particular way to justify Assumption 1.

with arguments, cues, and other soft information, which are ultimately mapped into a sentence. But the law and economics literature has overwhelmingly focused on settings in which the signal about the defendant is one dimensional. This appendix also provides a foundation for this assumption.

5 Discussion

Many features of the optimal judicial mechanisms identified by Theorems 1 and 2 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. In our model, this fixed sentence arises for two reasons. First, in an optimal mechanism, an innocent defendant has no incentive to mimic a guilty one, which means that the guilty defendant's sentencing scheme is not distorted: it maximizes welfare subject to providing the defendant a given expected utility. Second, defendants are risk averse, which allows the social planner to extract a higher sentence from the guilty defendant by insuring him against an uncertain sentence. When the social welfare function is concave, a fixed sentence is also socially beneficial and these considerations all go in favor of a fixed plea. Notice that this is true even when evidence is costless and informative, so the fact that evidence is optimally not used to determine a guilty defendant's sentence is not due to cost saving.

Alternatively, the optimal scheme for a guilty defendant may involve a lottery over two sentences. This randomization arises when effective deterrence requires sentences that are higher than is ex post optimal. In this case, replacing a sentence lottery by its certainty equivalent leads to a higher sentence than is socially optimal. By providing a random sentence that is sometimes closer to the ex-post optimum, a lottery may be preferable. Even then, the concavification argument used to prove Theorem 2 shows that it is always optimal to use two sentences in this lottery. All that matters for this lottery is the probability of each sentence, so a mapping that correctly randomizes between the two sentences and completely disregards the signal is optimal. This is in contrast to the sentencing scheme for a defendant who claims to be innocent, which optimally depends on the signal. 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.[49] Since the punishment in such pleas is determined without a trial, it does not depend on the evidence that a trial would have generated.

Another feature of the optimal mechanisms that is familiar from the American legal system is that

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

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, resembling "beyond a reasonable doubt," if it it much more important to acquit innocent defendants than it is to punish guilty ones. This is 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.

The result that an optimal guilty verdict carries a maximal sentence is not new. It is, in fact, arguably the main finding of Becker's (1968) seminal analysis. But the mechanisms leading to this result are completely different in Becker's paper and in ours. In Becker's paper the maximal sentence is used to maximize deterrence while economizing on the cost of law enforcement. If a guilty defendant is apprehended with probability $\gamma(e)$, where e is the amount spent on law enforcement and γ 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. Here, by contrast, the maximal sentence is used to achieve the maximal possible separation between guilty and innocent defendants: using a high sentence permits to concentrate the punishment for the most severe evidence, i.e., for the evidence with the highest likelihood ratio.

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.[50]

There are also important differences between features of the optimal mechanisms and criminal trials in the United States. Our analysis shows that in an optimal mechanism the role of evidence

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

is different from the role 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 the importance of commitment in reality and for our analysis. We also discuss how evidence regains the role of determining guilt when the optimal mechanism is modified slightly.

The importance of commitment and 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.[51] This feature relies on Assumption 1, because the proofs of Theorems 1 and 2 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's 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)). 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. Instructions to the jury entirely focus on describing the procedure for finding facts. As noted by Lee (2014), jurors

[51] 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.

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.[52] 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. 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.[53] 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.[54]

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 α of guilty defendants reject the plea and go to trial. The jury's belief, upon seeing a defendant going to trial and observing signal t regarding the defendant's guilt, is a combination of both pieces of information (rejecting the plea and generating signal t). With Bayesian updating, the posterior probability of guilt corresponding to some signal t can be computed in two steps. First, given prior λ and the fact that the defendant rejected the plea and

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

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

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

went to trial, the probability at the outset of the trial that the defendant is guilty is

$$\hat{\lambda} = \frac{\lambda\alpha}{\lambda\alpha + (1-\lambda)}. \tag{13}$$

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 (13), we have

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

Thus, for any fraction $\alpha > 0$ and conviction threshold $\hat{t}$ there corresponds a posterior belief $\hat{p}(\hat{t})$ of guilt. To get a rough sense of this threshold, suppose that the likelihood ratio at the optimal threshold $\bar{t}$ use 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.[55] Also suppose that, consistent with criminal data in the United States, 90% of defendants are in fact guilty.[56] These assumptions correspond to $\lambda = 0.9$ and $r(\bar{t}) = 10$. The associated posterior probability that the defendant is guilty is

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

For $\alpha = 0.1$, for instance, this implies that the posterior probability of guilt of a defendant who is barely convicted under the optimal scheme is 0.9, or 90%. Thus, even if the BARD doctrine is applied to posterior beliefs that take into account the decision of the defendant to reject the plea, instead of being based purely on the evidence presented at trial, the mechanism proposed here leads to a certainty threshold of 90% regarding the guilt of convicted defendants when 10% of guilty defendants reject the plea.

Thus, under realistic assumptions with regard to the evidence conviction threshold $\bar{t}$ and the prior λ of guilt, our modified mechanism remains consistent with BARD and the observation that most defendants are guilty. With a fraction α of guilty defendant 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.

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

[56] 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.

6 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 formalize the notion of commitment needed to 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 threshold similar to "beyond a reasonable doubt," no punishment following an acquittal, and plea bargains with certain and uncertain punishments. One difference between our results and actual trials is that in the optimal mechanisms only innocent defendant 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.

A From Judicial Processes to Direct Revelation Mechanisms: Formalization

We describe a multi-agent environment and a reduction to a single-agent environment focused on the defendant that lead to Assumption 1. To do this, we take a modern mechanism design approach that uses the concept of a mediator (Myerson (1983, 1986)).

We model a judicial system 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 system. 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 game consists of i) some *admissible evidence* E taking value in some probability space $(\mathcal{E}, \Sigma_E, \nu)$, 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 system.[57]

We let $\mathcal{C}$ denote the class of all judicial systems 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 system is* regular *if*

1) *the defendant's initial type θ is binary (guilty or not guilty), and nature determines it at the root of the game tree;*
2) *the sentence s lies in some interval $[0, \bar{s}]$;*
3) *the defendant's realized utility is a function of the sentence;*
4) *the realized welfare is a function of the sentence and the defendant's type;*
5) *the realized social cost is a function of the realized sequence of actions taken by players other than the defendant;*
6) *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;*
7) *there exists an equilibrium σ.*

Part 6) of the definition says 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 system below. A *judicial process* is a pair consisting of a regular judicial system and an equilibrium profile σ. A judicial process is *associated with* $\mathcal{C}$ if its judicial system belongs to $\mathcal{C}$.

Mediated system. Given any judicial process, we consider a modification of its judicial system that reduces the defendant's decision problem to a direct revelation mechanism. This reduction is formalized as follows: starting from the initial judicial system, a new node is inserted immediately after the root of the game tree, at which the defendant privately reports his type to a mediator. The rest of the game tree is unchanged except that the 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

[57] 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.

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 6) of the definition of a regular judicial system 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.[58] The realized welfare and cost are also 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 2 (Feasible Mediation) *Given any judicial system in $\mathcal{C}$, the mediated system associated to it also belongs to $\mathcal{C}$.*

Lemma 1 *Given a judicial process, the associated mediated system has the following properties:*

1) It is regular;

2) Letting σ denote the equilibrium defining the judicial process, the augmented strategy profile in which the defendant reports his type truthfully and other players (including the mediator) follow σ is an equilibrium of the mediated system;

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 system, in which σ 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 σ, 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 system, contradicting the premise that σ was an equilibrium in that system. ■

A mediated system together with the equilibrium mentioned in the previous lemma is called a *mediated process.* In the equilibrium associated with a mediated process, the defendant reports his type truthfully. Lemma 1 shows that any judicial process has an associated mediated process, which has the same welfare and cost distributions.

We now turn to our main invariance assumption. Given any mediated system, a *sentencing scheme* is a map $\tilde{s} : (E, \hat{\theta}) \to \Delta([0, \bar{s}])$, where E is any admissible evidence in the system. A judicial system is an $\tilde{s}$ *-modification* of the mediated system if it identical to the mediated system 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 system, 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 $\hat{\theta}$. 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 a defendant's report $\hat{\theta}$, so that the sentence can indeed depend on $\hat{\theta}$. 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. 2) the utility of the defendant, the welfare, and the social cost are given by the same functions of the sentence, defendant type, and realized actions as in the initial mediated system. 3) other players' utility are arbitrary.

[58] In accordance with the mechanism design literature, we assume that the mediator is committed to following this strategy. Alternatively, we could treat the mediator as any other player in the game, with a utility of 1 if he followed his prescribed strategy and 0 otherwise.

An $\tilde{s}$-modification thus preserves the same structure of the game (except possibly for the leaves, which may be replaced by a lottery followed by new leaves), changes the sentences of the mediated system in a particular way, preserves how the defendant's utility, social welfare, and costs depend on sentences, and allows any change in the realized utilities of the other players.

Given a mediated system, an equilibrium strategy profile σ, and a sentencing scheme $\tilde{s}$, let $u(\hat{\theta};\theta,\tilde{s})$ denote the expected utility of the defendant of type θ who reports $\hat{\theta}$ when other players—with the mediator replacing the defendant as explained above—follow the equilibrium strategy profile σ in any $\tilde{s}$-modification of the mediated system.[59] A sentencing scheme $\tilde{s}$ is *truthful* if $u(\theta;\theta,\tilde{s}) \geq u(\hat{\theta};\theta,\tilde{s})$ for all pairs $(\theta,\hat{\theta})$.

Assumption 3 *Given any mediated process whose system belongs to $\mathcal{C}$ and any truthful sentencing scheme $\tilde{s}$, there exists an $\tilde{s}$-modification in $\mathcal{C}$ such that if the defendant reveals his type truthfully, the strategies prescribed to the other players by the equilibrium of the mediated process form a continuation equilibrium in the $\tilde{s}$-modification. Such a modification is called* admissible.

Intuitively, Assumption 3 says that, starting from the 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. For example, an investigator's effort to look for evidence is unaffected by the possible sentences faced by the defendant. Likewise, a defense lawyer or prosecutor's effort to argue their case is not 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 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 straightforward and omitted.

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 truthfully.*

Definition 2 *A mediated process is* feasible *if the following properties hold in equilibrium:*

1) *The distribution of E has full support over $\mathcal{E}$ for $\theta, \hat{\theta} = g, i$, with positive density functions $f_\theta^{\hat{\theta}}$;*[60]
2) *The likelihood ratio $\ell(E,\hat{\theta}) = f_g^{\hat{\theta}}(E)/f_i^{\hat{\theta}}(E)$ for $\hat{\theta} = g, i$ has a non-atomic distribution with support $\mathbb{R}_+$;*
3) *Conditional on E and $\hat{\theta}$, s is statistically independent of θ.*

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

[59] Note that the defendant's expected utility does not depend on which $\tilde{s}$-modification is considered, conditional on σ.

[60] The density is taken with respect to measure ν over $\mathcal{E}$.

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.[61]

The proof of the following is straightforward and omitted.

Lemma 3 *Given a feasible mediated process and truthful sentencing scheme $\tilde{s}$, any admissible $\tilde{s}$-modification is also feasible.*

Two judicial processes are said to be welfare equivalent if they induce the same distribution for i) 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 could 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 sentence 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 4 *Given any feasible mediated process, there exists a truthful sentencing scheme $\tilde{s}$ that depends on $(E, \hat{\theta})$ only through $(\ell(E, \hat{\theta}), \hat{\theta})$, 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 θ conditional on $\ell(E, \hat{\theta})$ and $\hat{\theta}$.*

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

For any $\hat{\theta}$ and l, 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 θ, and we have just showed that the distribution of E conditional on l and $\hat{\theta}$ is also independent of θ. Combining these observations, it follows that the distribution of $\tilde{s}$ is well defined independently of θ.[62]

Now suppose that a defendant of type θ reports $\hat{\theta}$ and that other players follow the strategies prescribed by the equilibrium σ of the initial mediated process. By construction, the defendant will face the same sentence lottery

[61] 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 4 below, and not on any other private information held by any of the other actors of the judicial system.

[62] More precisely, we need to show that s is independent of θ conditional on $\hat{\theta}$ and l. Consider any bounded functions g, h of s and θ, respectively, and let $F_l^{\hat{\theta}}$ denote the distribution of E conditional on $\mathcal{E}(l, \hat{\theta})$, which we have shown to be independent of θ. We have $E[g(s)h(\theta)|l, \hat{\theta}] = \sum_{E \in \mathcal{E}(l, \hat{\theta})} E[g(s)h(\theta)|E, \hat{\theta}] dF_l^{\hat{\theta}}(E) = \sum_{E \in \mathcal{E}(l, \hat{\theta})} E[g(s)|E, \hat{\theta}] E[h(\theta)|E, \hat{\theta}] dF_l^{\hat{\theta}}(E) = E[g(s)|l, \hat{\theta}] E[h(\theta)|l, \hat{\theta}]$, where the first equality comes from the fact that any evidence $E \in \mathcal{E}(l, \hat{\theta})$ implies likelihood ratio l, so l can be removed from the conditioning variables; the second equality comes from independence of s and θ 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 in this sequence of equalities then shows that s and θ are independent conditional on l and $\hat{\theta}$.

under the scheme $\tilde{s}$ as he was facing under scheme s by reporting $\hat{\theta}$ and, hence, the same utility distribution. Therefore, the sentencing scheme $\tilde{s}$ 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 σ, the distribution of welfare is the same as under the initial scheme, because by regularity, realized welfare function depends only on the realized sentence and the defendant's type, and we already argued 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 $\tilde{s}$-modification is feasible in the sense of Definition 2 is trivial since the distribution of E has not been modified.

Finally, there remains to check that $\tilde{s}$ is statistically independent of θ conditional on $\hat{\theta}$ and $\ell(E, \hat{\theta})$, but this is a direct consequence of fact that s is independent of θ conditional on $\hat{\theta}$ and E (by feasibility) and the earlier observation that the distribution of E conditional on $\hat{\theta}$ and the event $\ell(E, \hat{\theta}) = l$ is independent of θ for any fixed l. ■

We can finally state the main definition and result of this section.

Definition 3 *A class $\mathcal{C}$ of judicial systems is* feasible *if the following holds:*

1) *All judicial systems in $\mathcal{C}$ are regular;*
2) *All mediated processes associated with $\mathcal{C}$ are feasible;*
3) *All judicial systems have realized sentences taking values in the same fixed interval $[0, \bar{s}]$;*
4) *The defendant's utility is the same function $s \mapsto u(s)$ across all judicial systems;*
5) *The realized welfare is the same function $(s, \theta) \mapsto w(s, \theta)$ across all judicial systems.*

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

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 sentence scheme $s(\hat{\theta}, t) \in \Delta([0, \bar{s}])$. The signal t has a distribution that depends on θ 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 θ, and iii) a realized social cost, which depends on θ, $\hat{\theta}$, and t.

A direct judicial mechanism is *truthful* if truth telling is optimal for the defendant.

Theorem 3 *Suppose that a class $\mathcal{C}$ of judicial systems is feasible and satisfies Assumptions 2 and 3. Then, the following holds:*

1) *Any judicial process is associated with a direct judicial mechanism that is truthful and such that 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 and any truthful sentence scheme, there is another direct judicial mechanism with this sentence scheme and the same tuple F, and social cost function as in the original direct judicial mechanism;*
4) *All realized sentences take values in $[0, \bar{s}]$;*
5) *The defendant's realized utility is $u(s)$;*
6) *Realized welfare is $w(s, \theta)$.*

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

B Proof of uniqueness in Theorem 1

Suppose first that S violates i) over a positive measure of signals. In this case, the step function $\tilde{S}(t,\hat{\imath})$ constructed in the first part of the proof is such that the difference $S(t,\hat{\imath})-\tilde{S}(t,\hat{\imath})$ is strictly positive over a subset T_1 of $[0,\hat{t})$ that has positive Lebesgue measure and strictly negative over a subset T_2 of $(\hat{t},1)$ that has positive Lebesgue measure.[63] Since u is strictly decreasing, this implies that the single-crossing function $\delta : t \mapsto \delta(t) = u(S(t,\hat{\imath})) - u(\tilde{S}(t,\hat{\imath}))$ is strictly negative over T_1 and strictly positive over T_2. We now show that this implies that (7) is strict. Let $H(t) = \int_t^1 \delta(\tau) f_i^{\hat{\imath}}(\tau) d\tau$. By construction, we have $H(0) = H(1) = 0$, $H(t) \geq 0$ for all t, and $H(t) > 0$ for all t in the interior of the convex hull of $T_1 \cup T_2$.[64] Let $\gamma(t) = f_g^{\hat{\imath}}(t)/f_i^{\hat{\imath}}(t)$. By strict MLRP, γ is a strictly increasing function and is thus almost everywhere differentiable. Therefore,

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

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

This shows that (7) holds as a strict inequality, and thus that one may strictly increase the conviction threshold $\bar{t}$ without violating incentive compatibility. More precisely, the argument given in the last paragraph of the proof in the main text applies.

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^a: 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 incentive compatibility.

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

[64] The fact that $H(0) = 0$ is simply a restatement of(6). Nonnegativity of H comes from the fact that the integrand of H, $\delta(t)f_i^i(t)$, is first negative and then positive and integrates up to 0, and the strict inequalities come from the fact that the integrand is strictly negative over T_1 and strictly positive over T_2.

C Proof of generic uniqueness in Theorem 2

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

The notion of "almost all" that we choose is the standard mathematical notion of "prevalence," which is used to conceptualized genericity for infinite-dimensional sets, like the set of functions that we consider here. [65]

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

In our case, the functions of interest are of the form $U \mapsto \hat{W}(U) = W(u^{-1}(U), g)$. Since u^{-1} is continuous[66] 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 taken to be $[0, 1]$.

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

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

To prove this result, the finite-dimensional subset $\mathcal{V}$ that we choose[67] is the set $\{af : a \in \mathbb{R}\}$, where $f(x) = x^2$. $\mathcal{V}$ is thus one dimensional.

Given a function $w \in \mathcal{W}$, let $w_a = w + af$, and let $A(w) = \{a \in \mathbb{R} : w_a \text{ violates the two-contact property}\}$. We wish to show that $A(w)$ has zero Lebesgue measure. For any fixed a, let $\{I_k^a\}_k$ denote the collection of maximal intervals of $[0, 1]$ over which the concavification $\bar{w}_a$ of w_a is linear and coincides with w_a at three or more points points of these intervals. Since these intervals are maximal, they are closed. Moreover, if a is increased slightly, it is straightforward to see,[68] by strict convexity of f, that there are at most two points of contact over I_k^a for all $a' > a$: all interior points x of I_k^a are such that $w_{a'}(x) < \bar{w}_{a'}(x)$.

If w_a violates the two-contact property for some a, this implies that for any $a' > a$ the set of maximal intervals over which $w_{a'}$ violates the two-contact property consists of intervals $I_{k'}^{a'}$ that are either in the closure of

[65] 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.

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

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

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

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

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

D Extension: Discrepancy between Welfare and Utility

Although the social preference may be broadly aligned with the defendant's utility when the defendant is innocent, this does necessarily not imply that the corresponding objective functions should be identical.

In this appendix, we relax the assumption common in the literature (Grossman and Katz (1983)), that $W(\cdot, i) = u(\cdot)$. More precisely, we assume 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. Specifically, our result in this section concerns the case in which ϕ is convex, which means that the social preferences when facing an innocent defendant exhibits less risk aversion than the defendant's preferences. From a positive perspective, this assumption captures the idea that society may not quite internalize the full extent of an innocent defendant's exposure to the judicial process.

Proposition 1 *Suppose that ϕ is increasing and convex and that the other assumptions of Theorem 2 hold. Then, there exists a welfare-maximizing optimal mechanism satisfies all the conclusions of Theorem 2.*

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

For expositional simplicity, let us normalize the utility functions as follows: $u(0) = 0$, $u(\bar{s}) = -1$, $\phi(0) = 0$ and $\phi(-1) = -M$. This normalization is without loss of generality, as is easily checked.[69] We wish to show that

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

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

$$\int_0^1 \phi(u(S(t, \hat{\imath}))) f_i^{\hat{\imath}}(t) dt \leq -M F_i^{\hat{\imath}}([\bar{t}, 1]), \tag{14}$$

It follows from the indifference equation (5) and the above normalization that the cutoff $\bar{t}$ satisfies

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

[69] 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.

Since $u(\cdot)$ takes values in $[-1,0]$ we can view $-u(\tilde{s})$ as a weight in a convex combination. Since also $u(0) = \phi(0) = 0$, $u(\bar{s}) = -1$, and $\phi(-1) = -M$, we have[70]

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

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

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

Combining this with (15) then yields (14). ■

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

References

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

Athey, S. (2002) “Monotone Comparative Statics under Uncertainty,” *Quarterly Journal of Economics*, Vol. 117, pp. 187–223.

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

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

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

Ben-Porath, E., Dekel, E., and Lipman, B. (2014) “Optimal Allocation with Costly Verification,” *American Economic Review,* Vol. 104, pp. 3779-3813.

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

Da Silveira, B. (2015) Bargaining with Asymmetric Information: An Empirical Study of Plea Negotiations,” *Working Paper*, Washington University.

Daughety, A., Reinganum, J. (2015a) “Informal Sanctions on Prosecutors and Defendants and the Disposition of Criminal Cases,” forthcoming in the *Journal of Law, Economics, and Organization.*

Daughety, A., Reinganum, J. (2015b) “Selecting Among Acquitted Defendants: Procedural Choice vs. Selective Compensation,” *Working Paper*, Vanderbilt University.

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

Friedman, J., Holden, R. (2008) “Optimal Gerrymandering: Sometimes Pack, but Never Crack,” *American Economic Review,* Vol. 98, pp. 113–144.

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

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

Jehiel, P., MeyerÂ€terÂ€Vehn, M., Moldovanu, B., and W. Zame (2006) “The Limits of Ex Post Implementation,” *Econometrica*, Vol. 74, pp. 585–610.

Kamenica, E., and Gentzkow, M. (2011) “Bayesian Persuasion,” *American Economic Review*, Vol. 101, pp. 2590–2615.

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

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

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

Karlin, S., and Rubin, H. (1956) "The theory of decision procedures for distributions with monotone likelihood ratio." *The Annals of Mathematical Statistics*, Vol. 27, pp. 272–299.

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

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

Myerson, R. (1983) "Mechanism Design by an Informed Principal," *Econometrica*, Vol. 51, pp. 1767–1797.

Myerson, R. (1986) "Multistage games with communication," *Econometrica*, Vol. 54, pp. 323–358.

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

Reinganum, J. (1988) "Plea Bargaining and Prosecutorial Discretion," *American Economic Review*, Vol. 78, pp. 713-728.

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

Siegel, R., and Strulovici, B. (2017) "Multiverdict Systems," *Working Paper*, Pennsylvania State University and Northwestern University.

Silva, F. (2016) "If We Confess Our Sins," *Working Paper*, University of Pennsylvania.

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

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