Improving Criminal Trials by Reflecting Residual Doubt: Multiple Verdicts and Plea Bargains

Ron Siegel and Bruno Strulovici*

June 18, 2016

Abstract

We propose adding intermediate verdicts to the two-verdict system used in criminal trials to distinguish convicted defendants based on the residual doubt regarding their guilt at the end of trial. Appropriately designed, additional verdicts improve welfare without increasing wrongful convictions or the incentives to commit crime. We consider plea bargains, a form of intermediate verdict, and show that a properly chosen plea in a two-verdict system increases welfare relative to any multi-verdict system, and characterize the optimal mechanism accounting for the incentives to commit crime. Finally, we consider how additional verdicts affect social stigma and the incentives to gather evidence.

---

*We thank Daron Acemoglu, Robert Burns, Andy Daughety, Eddie Dekel, Louis Kaplow, Fuhito Kojima, Adi Leibovitz, Paul Milgrom, Kathy Spier, Jean Tirole, and Leeat Yariv for their comments and Jennifer Reiganum 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, Duke, Penn State, Johns Hopkins, and the Pennsylvania Economic Theory Conference. 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.

1 Introduction

Criminal trials are imperfect: innocent defendants are sometimes convicted and guilty ones are sometimes acquitted.[1] This is unavoidable, because trials cannot always eliminate all doubt regarding defendants' guilt. How this residual doubt translates into a verdict is determined by the standard used for conviction. In the United States, the standard is "beyond a reasonable doubt," which reflects the view that it is more important not to punish the innocent than it is to mistakenly acquit the guilty.

One way to improve trial outcomes is to reduce the residual doubt regarding defendants' guilt. Technological advances such as DNA profiling sometimes achieve this, but attaining absolute certainty regarding a defendant's guilt in every case is not realistic. This paper proposes a different improvement, which builds on the observation that residual doubt varies across trials. Consider, for example, a trial in which a defendant is found guilty based on a confession and an eye witness report. These pieces of evidence may establish the defendant's guilt "beyond a reasonable doubt," but because confessions and eye-witness reports are known to be unreliable to some extent, some residual doubt remains. A similar trial in which additional evidence is available, such as clear footage of the defendant committing the crime, would result in less residual doubt regarding the defendant's guilt. This variance in residual doubt across trials cannot be reflected in a two-verdict system, in which the defendant is found either guilty or not guilty.

We propose introducing intermediate verdicts as possible outcomes in criminal trials. In particular, introducing a third verdict when the residual doubt is close to "reasonable" improves welfare when the judge or jury is torn between convicting and acquitting a defendant. In this case, an intermediate punishment reduces the welfare loss of convicting an innocent defendant or acquitting a guilty one. The possibility of an additional verdict has been proposed in the legal literature by Bray (2005), but has received little formal analysis.[2]

[1]For example, a recent study by Gross et al. (2014) of 7,482 death row convictions from 1973 to 2004 in the United States estimates that at least 4.1% of death-row defendants have been wrongfully convicted. Given the high burden of proof required for convictions, acquittals of guilty defendants are likely even more frequent.

[2]Bray's proposal concerns the addition of a "not-proven" verdict to the U.S. criminal system, which does not carry any jail time and distinguishes acquitted defendants, unlike the intermediate verdicts which we introduce in Section 2 and distinguish convicted defendants . Daughety and Reinganum (2015a) consider the effect of informal sanctions on defendants and prosecutors. In an extension discussed later in that paper, they consider the effect

Punishments in criminal trials that can be viewed as "intermediate" currently arise for other reasons. The punishment for homicide, for example, may depend on whether the defendant is charged with murder or manslaughter;[3] a single crime may lead to multiple charges, and a defendant may be convicted of only a subset of them; extenuating circumstances may substantially affect the sentence associated with a conviction. Notice, however, that the variability in punishment in these cases arises because of the variability in the nature and circumstances of the crime, not because of the degree of certainty that the defendant committed the crime. To the extent that these instruments are used to reflect residual doubt, they are not designed for this and can lead to arbitrary, unfair, and suboptimal outcomes, as explained in Section 6.

A natural question is why criminal trials today do not commonly use additional verdicts. One possibility is that such verdicts would be an open admission of the system's imperfection. After all, in an ideal world guilty defendants would be convicted and innocent ones would be acquitted, so additional verdicts would be of no value. But criminal trials are in fact imperfect, and, as we show, welfare can be increased by recognizing this fact and introducing an intermediate verdict. Another possibility is that additional verdicts give rise to several concerns. One concern is that more innocent defendants would be convicted. Another is that the incentives to commit crimes would increase. A third is that the incentives to gather evidence may be diminished. A fourth is that implementing the addition would require infeasible changes to the system, or would only be beneficial if the current punishments and conviction standard are close to optimal, which may not be the case.

We show that adding a verdict with an appropriate sentence increases welfare and addresses all these concerns. The intermediate verdict will be used to distinguish among defendants who would be convicted in the current system. Among those defendants, the ones for whom more doubt remains will be punished less severely than those whose guilt is more certain. We show that for any punishment in the current system and any doubt threshold exceeding the one currently used for conviction, there is a way to set the punishments above and below the threshold that increases welfare relative to the current system and does not increase the incentives to commit crimes. This guarantees that every defendant who would be acquitted in the current system would also be acquitted in the new system. In particular, no additional innocent defendants

of introducing a not-proven verdict. Daughety and Reinganum (2015b) consider several implementations of the not proven verdict through defendant choice and compensation.

[3]Homicide is an exceptional crime in that it is associated with several different criminal counts.

would be convicted. If the punishment in the current system is not too inefficiently lenient, we obtain the stronger result that welfare can be improved without increasing the punishment relative to the current system. This guarantees not only that no additional innocent defendants are punished in the new system, but also that those who are punished are never punished more severely than in the current system.[4]

The additional verdict can be introduced into criminal trials in several ways. One possibility is to have the jury first determine whether the defendant is guilty according to the standard used in the current system. If the jury find the defendant guilty, then in a second stage the jury would further indicate whether they find the defendant guilty "beyond a reasonable doubt" or "beyond all doubt," with a lower sentence for the former. This distinction has recently been advocated in the context of capital trials (see Section 6). A second possibility is not to change the jury's current role and instead to relegate the distinction between the two degrees of guilt uncertainty to the sentencing stage. This two-step implementation is explored in Section 7. If the jury find the defendant guilty, then the judge would determine the sentencing category based on the residual doubt regarding the defendant's guilt. A third possibility is not to change the jury's or the judge's current role and instead introduce rules or guidelines (via legislation or other means) that determine the degree of residual doubt following a conviction based on the strength of evidence produced during the trial. It may also be possible to combine some of these methods or introduce additional ones. Notice that in all the methods jurors would still be given, and should follow, the current guidelines for conviction, so the set of convicted defendants would not change.[5]

A potential concern is that jurors and other agents of the criminal justice system may reduce their effort to acquire and seriously consider the evidence if an intermediate verdict is introduced. To gain a better understanding of this issue, we consider how the introduction of a third verdict affects the value of evidence in a trial. Since gathering evidence is costly, the socially optimal

[4]Our result about the welfare-improving addition of a third verdict holds more generally: for any multi-verdict system one can add another verdict and lower the punishments in a way that increases social welfare.

[5]Jurors are currently instructed to focus only on determining the defendant's guilt and ignore the punishment carried by a conviction (Sauer, 1995). To the extent that jurors deviate from these guidelines more in the new system than in the current system, social welfare would be further improved, as long as jurors have society's best interests in mind. Section 7 discusses how jurors' incentives may be affected by the introduction of an intermediate verdict.

amount of evidence to be gathered (and jurors' incentives to fully process this evidence) depends on the verdict structure. We show that adding a verdict can increase the value of evidence and therefore the optimal amount of evidence that should be gathered. We obtain this result both in a two-period discrete model and in a continuous-time model in which the defendant's likelihood of guilt is updated stochastically as long as evidence is gathered.

Another approach to reducing residual doubt is to induce defendants to reveal whether they are guilty. Defendants for which this is done successfully would not go through a trial, so any residual doubt regarding their guilt would be avoided. Of course, if guilty defendants are to be punished, then simply asking defendants whether they are guilty would not work. One way to induce defendants to reveal their guilt is to offer them a plea bargain, which is an admission of guilt along with a lower sentence than the one associated with a conviction.

Plea bargains are an important instrument in the United States criminal justice system.[6] Because defendants choose whether to accept the plea, and guilty defendants are (presumably) more likely than innocent ones to be found guilty during a trial, the plea can serve as a screening device. Building on the framework of Grossman and Katz (1983), in which guilty defendants are indeed more willing to take a plea, we analyze the value of plea bargains relative to other verdict systems. We show that an appropriate two-verdict system with a plea dominates *any* multi-verdict system without pleas, regardless of the number of verdicts in the system, provided that the defendant's utility function is independent of his guilt and the punishments in the original system are not too inefficiently harsh from an interim perspective. In fact, under the same conditions we show that there is a two-verdict system with a plea that maximizes welfare among all incentive compatible mechanisms, and does not increase the incentives to commit crimes.[7] If some punishments in the original system are inefficiently harsh from an interim perspective, which may be ex-ante optimal to generate deterrence, then a two-verdict system with a plea may not be optimal. In such cases, however, a random scheme in which guilty defendants face one of two high sentences, which can be determined independently of any information a trial would generate, is optimal. This "random plea" is consistent with plea bargains in which the

[6]More than 90% of criminal cases in the United States are settled by plea bargains (Burns (2009)). The corresponding percentage in many European countries is much lower, especially for serious crimes.

[7]The characterization of the optimal mechanism does not follow from standard results, because the mechanism design environment does not include transfers.

judge has discretion over the sentence after the defendant agreed to the plea bargain.[8]

Despite its generality, the result on the superiority of two-verdict systems with plea bargains omits several issues. When some innocent defendants are more risk averse than guilty ones, for instance, these innocent defendants may prefer to plead guilty rather than face the lottery of the trial, particularly if the sentence set for a guilty verdict is set at level meant to be optimal conditional on a convicted defendant being surely guilty. Since some innocent defendants are also convicted, that maximal sentence may be too harsh, leading some innocent defendants to accept the plea bargain. We demonstrate (see Appendix C) that when the guilty sentence is suboptimally harsh, the two-verdict system with a plea may be inferior to a three-verdict system.[9] The result is, however, robust in other dimensions. For example, Silva (2015) studies a general mechanism with multiple defendants whose types (guilty or innocent) may be correlated and whose sentences may depend on one another's reports, and finds that there exists an optimal confession-inducing scheme in which confessions are met with a flat sentence similar to a plea bargain.

We also consider using the additional verdict to distinguish among defendants who would be acquitted in the two-verdict system. Since these defendants are not punished in the two-verdict system, they would not be punished in the three-verdict system. But acquitted defendants may suffer from the stigma of having been tried.[10] Because this stigma is likely related to the perceived likelihood that they are in fact guilty, distinguishing among these defendants based on the residual doubt at the end of the trial may affect the stigma they face. We treat the stigma mechanism as exogenous, since it is determined by society and cannot be legislated in the same way that sentences are. Consequently, this additional verdict does not always increase welfare, in contrast to our first result, since its socially detrimental effect on acquitted defendants who are in fact guilty may outweigh the socially beneficial effect on innocent defendants. We provide conditions under which welfare does increase, as well as comparative statics.

Several countries, including Israel, Italy, and Scotland, do in fact distinguish among acquit-

[8]A recent example is the case of 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.

[9]One may also construct examples in which an innocent defendant who overestimates the probability of being found guilty in a trial, perhaps through persuasion or intimidation, may take a plea. In this case, a three-verdict system can again dominate the two-verdict system with a plea.

[10]Economic analyses of the stigma faced by convicts are provided by Lott (1990) and Grogger (1992, 1995)

ted defendants based on the residual doubt regarding their guilt. In Scotland, for example, a conviction in a criminal trial leads to a "guilty" verdict, but an acquittal leads to either a verdict of "not guilty" or "not proven." Neither of the two acquittal verdicts carries any jail time, but the latter indicates a higher likelihood that the defendant is in fact guilty.[11] The likelihood is, however, insufficiently high for conviction.[12]

The appendix provides a micro-foundation for the Bayesian formulation used in later parts of the paper. It establishes that trial technology conceptualized as a mapping from accumulated evidence to a verdict can always be reformulated in Bayesian fashion: accumulated evidence is a signal that turns the prior probability that the defendant is guilty into a posterior probability, on which the verdict is based. Moreover, this transformation establishes a relationship between two notions of 'incriminating' and 'exculpatory' evidence. One notion is based on decisions and the other on beliefs. What makes a piece of evidence 'incriminating' is the fact that it increases the likelihood of guilt of a defendant and, hence, results in a longer expected sentence. In particular, there is no loss of generality when one says that a guilty defendant is more likely than innocent defendant to generate incriminating evidence.

2 Reflecting residual doubt in trial outcomes

We consider a trial whose objective is to determine whether a defendant is guilty of committing a certain crime and to deliver the corresponding sentence. In our baseline model the trial is summarized by two numbers: the probability π_g that the defendant is found guilty if he is actually guilty, and the probability π_i that the defendant is found guilty if he is actually innocent.[13] Corresponding to a guilty verdict is a sentence $s > 0$, interpreted as jail time (so a

[11] The introduction of a not-proven verdict is considered by Daughety and Reinganum (2015a), who study how the effect of informal sanctions on defendants and prosecutors affect the plea bargaining process and its acceptance rate, and consider the effect of a not-proven verdict in this context. Daughety and Reinganum (2015b) consider two implementations of a not-proven verdict. In the first one, the defendant can choose between the standard binary verdict system and the system with a not-proven verdict. In equilibrium, all defendants choose the latter system. The authors also analyze an alternative implementation in which some defendants who are found not guilty are compensated.

[12] This may happen, for example, if an eye-witness testimony exists, but the testimony cannot be corroborated.

[13] It is natural to assume that $\pi_g > \pi_i$, i.e., a defendant is more likely to be found guilty if he is actually guilty than if he is innocent. This assumption is, however, not required for this section.

higher value of s corresponds to a harsher punishment).[14]

Society wishes to avoid punishing the defendant if he is innocent, and adequately punish him if he is guilty. This dual goal is modeled by an ex-post, differentiable welfare function, denoted W. Jailing an innocent defendant for s years leads to a welfare of $W(s,i)$, with $W(0,i)=0$ and W decreasing in s. Jailing a guilty defendant leads to a welfare of $W(s,g)$, which has a single peak at $\bar{s}>0$. Thus, $\bar{s}$ is the punishment deemed optimal by society if it is certain that the defendant is guilty. The assumption that $W(s,g)$ increases up to $\bar{s}$ and then decreases is in line 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."[15]

The relative importance of punishing the defendant if he is guilty and not punishing him if he is innocent depends on the prior probability $\lambda \in (0,1)$ that the defendant is guilty. 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. This is captured by the interim social welfare from the defendant going to trial when the punishment of being found guilty is s:

$$\mathcal{W}_2(s) = \lambda\left[\pi_g W(s,g) + (1-\pi_g)W(0,g)\right] + (1-\lambda)\left[\pi_i W(s,i) + (1-\pi_i)W(0,i)\right]. \qquad (1)$$

Since $W(\cdot,i)$ is decreasing and $W(\cdot,g)$ peaks at $\bar{s}$, it is never interim optimal to choose $s>\bar{s}$.

Society's ex-ante welfare also depends on whether the crime is committed in the first place. The incentives to commit the crime play a key role in seminal economic analyses of criminal justice systems (Becker (1966), Stigler (1970)), and received renewed emphasis from Kaplow (2011). To model this, we consider an individual's decision whether to commit the crime, and assume that at most one individual is prosecuted for the crime if it is committed.[16] In a large society, the probability that any particular innocent individual is prosecuted for the crime is infinitesimal, so for expositional convenience we assume that an innocent individual treats this

[14]We leave aside such issues as mitigating circumstances, which are tangential to the focus of the paper.

[15]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.

[16]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 such issues.

probability as 0.[17] If the individual commits the crime, he obtains a benefit b (in utility terms), but faces a probability η_g of being arrested and prosecuted.[18] Thus, the individual commits the crime if

$$b + \eta_g \left(\pi_g u(s) + (1 - \pi_g)u(0)\right) > 0, \tag{2}$$

where $u(s) \leq 0$ is the defendant's differentiable utility from a sentence s, and the utility from not being prosecuted is normalized to 0. Denote by $H(s)$ the fraction of individuals who commit the crime, i.e., for whom (2) holds. The benefit b is distributed in the population according to an absolutely continuous cdf B, so by (2) we have

$$H(s) = 1 - B\left(-\eta_g \left(\pi_g u(s) + (1 - \pi_g)u(0)\right)\right). \tag{3}$$

By normalizing the welfare from no crime to 0, we obtain that the ex-ante social welfare is

$$H(s)\left(\eta_g \left(\pi_g W(s, g) + (1 - \pi_g)W(0, g)\right) + \eta_i \left(\pi_i W(s, i) + (1 - \pi_i)W(0, i)\right) - h\right), \tag{4}$$

where η_i is the probability that an innocent defendant is prosecuted and h is the social harm from the crime.[19] In particular, since that sentence s determines which individuals commit the crime and which are deterred,[20] it affects social welfare in addition to the direct affect of the sentence on the ex-post welfare W. Finally, when an individual is prosecuted the crime has already been committed so the social harm h from the crime is "sunk," and the prior that the defendant is guilty is $\lambda = \eta_g/(\eta_g + \eta_i)$, so we recover (1).

Because we will later consider multiple verdicts, we rewrite the interim social welfare (1) more generally as

$$\lambda E_g\left(W\left(\tilde{s}, g\right)\right) + (1 - \lambda)E_i\left(W\left(\tilde{s}, i\right)\right), \tag{5}$$

where the sentence $\tilde{s}$ is a random variable whose distribution depends on whether the defendant committed the crime. Similarly, we rewrite (2) as

$$b + \eta_g E_g\left(u\left(\tilde{s}\right)\right) > 0, \tag{6}$$

[17]The probability $1 - \lambda > 0$ that a prosecuted individual is innocent is, however, not infinitesimal.

[18]This probability can be endogenized by including the amount of costly law enforcement as a decision variable without changing any of the results.

[19]The benefit from committing the crime can be considered explicitly as well without affecting any of the results.

[20]Guidelines 18 U.S.C § 3553 state that another goal of punishment is "to afford adequate deterrence to criminal conduct."

and rewrite (3) as

$$H(\tilde{s}) = 1 - B(-\eta_g E_g(u(\tilde{s}))), \tag{7}$$

where $H(\tilde{s})$ is the fraction of individuals who commit the crime, i.e., for whom (6) holds. We rewrite (4) as

$$H(\tilde{s})(\eta_g E_g(W(\tilde{s}, g)) + \eta_i E_i(W(\tilde{s}, i)) - h). \tag{8}$$

Throughout the analysis we assume that all sentences are interior, in the sense that they can be made more severe.[21] We also assume that the harm caused by the crime exceeds the social welfare from punishing the perpetrator, i.e.,

$$W(\bar{s}, g) - h < 0. \tag{9}$$

2.1 Intermediate 'guilty' verdict

We consider adding a verdict that refines the 'guilty' verdict from the two-verdict system.[22] Those defendants who would be convicted in the two-verdict system now receive one of two "guilty verdicts," which we denote 1 and 2. Defendants who would be acquitted in the two-verdict system are still acquitted and are released.[23] The distinction between the two 'guilty' verdicts may be based on the evidence available before and during the trial, so that among the collections of evidence that would lead to a conviction in the two-verdict system some lead to verdict 1 and the remaining to verdict 2.[24] Denote by π_i^1 the probability that the defendant

[21] This can be done by imposing a longer or harsher imprisonment term. Even an execution can be made more severe by making it less humane. While an extreme sentence would maximize crime deterrence, it would also deter (or "chill") desirable behavior (Kaplow (2011)) and excessively punish those individuals who were not deterred and committed the crime, either because they were ignorant of the possible punishment or did not rationally assess the consequences of their crime before its commission. Formally, the optimal sentence will be interior, even taking deterrence into account, if i) the maximal benefit from the crime exceeds the maximal disutility from the harshest possible sentence (e.g., benefits have an unbounded support and the defendant's utility is bounded below), and ii) social welfare becomes sufficiently negative as defendants' punishment becomes sufficiently harsh.

[22] Further intermediate verdicts may similarly be added, as discussed below.

[23] Section 7 discusses how to implement the additional verdict in a way that is likely not to affect jurors' decision whether to acquit the defendant. It also discusses how the analysis might change if their decision is affected.

[24] Evidence leading to a homicide conviction in the two-verdict system may include, for example, the discovery, in the defendant's house, of the gun from which the bullet was fired, a confession by the defendant, a death threat made by the defendant to the victim shortly before the murder, or any subset of these.

receives verdict 1 if he is innocent, and define π_i^2, π_g^1, and π_g^2 similarly.[25] Because the same set of defendants is acquitted as in the two-verdict case, we have

$$\pi_i = \pi_i^1 + \pi_i^2 \quad \text{and} \quad \pi_g = \pi_g^1 + \pi_g^2.$$

Without loss of generality[26]

$$\frac{\pi_g^2}{\pi_i^2} > \frac{\pi_g}{\pi_i} > \frac{\pi_g^1}{\pi_i^1},$$

so verdict 1 is an "intermediate verdict:" a guilty defendant is more likely to receive verdict 2, relative to an innocent defendant, than verdict 1.

Let s_j denote the sentence associated with verdict j. Given s_1 and s_2, the interim social welfare is given by

$$\begin{aligned} \mathcal{W}_3(s_1, s_2) = \lambda \left[\pi_g^1 W(s_1, g) + \pi_g^2 W(s_2, g) + (1 - \pi_g) W(0, g)\right] + \\ (1 - \lambda) \left[\pi_i^1 W(s_1, i) + \pi_i^2 W(s_2, i) + (1 - \pi_i) W(0, i)\right]. \end{aligned} \tag{10}$$

Our first result shows that s_1 and s_2 can be chosen so that this welfare is higher than the interim social welfare in the two-verdict system.

Proposition 1 *For any sentence $s > 0$ in the two-verdict system and any verdict technologies π_i, π_g, π_i^j, etc., there exists a three-verdict system with sentences s_1 and s_2 in which the interim welfare is higher than in the two-verdict system, i.e., $\mathcal{W}_3(s_1, s_2) > \mathcal{W}_2(s)$. Moreover, the welfare is higher conditional on the defendant being innocent and conditional on the defendant being guilty. If $s \leq \bar{s}$, then $s_1 < s < s_2$.*

One key aspect of Proposition 1 is that it applies to all two-verdict systems, even those with a suboptimal sentence $s > 0$, and all technologies for splitting of the conviction probabilities π_i and π_g. In particular, it applies whether s was chosen with an ex ante or an interim perspective in mind. Another key aspect of Proposition 1 is that the three-verdict system does not increase the probability of punishing the innocent relative to the two-verdict system. Instead, it modifies the sentence to reflect the richer information that verdicts 1 and 2 convey regarding the relative likelihood of the defendant being guilty or innocent.

[25] In keeping with most of the literature on trial design, we take a reduced-form approach to modeling these probabilities. We provide a micro-foundation for these probabilities in Appendix B.

[26] For any a, b, c, d of $\mathbb{R}_{++}$ we have $\min\{a/b, c/d\} \leq (a + c)/(b + d) \leq \max\{a/b, c/d\}$, with strict inequalities if $a/b \neq c/d$, a generic condition which we will assume throughout (it is easy to impose conditions to guarantee it: for example, one can rank bodies of evidence in terms of the posterior that they generate, as in Appendix B).

Proof. If $s > \bar{s}$, then setting $s_1 = s_2 = \bar{s}$ suffices, since the ex-post welfare $W(s,i)$ and $W(s,g)$ decreases in $s > \bar{s}$. Suppose that $s \leq \bar{s}$. First, observe that $\mathcal{W}_3(s,s) = \mathcal{W}_2(s)$: if we give verdicts 1 and 2 the sentence associated with the guilty verdict of the two-verdict case, then we clearly obtain the same welfare as in the two-verdict case. We are going to create a strict welfare improvement by slightly perturbing the sentences s_1 and s_2. Consider any small $\varepsilon > 0$ and let $s_1 = s - \varepsilon$ and $s_2 = s + \varepsilon\gamma$. The welfare impact of this perturbation is

$$\mathcal{W}_3(s_1, s_2) = \mathcal{W}_2(s) + \lambda\varepsilon W'(s,g)\left(-\pi_g^1 + \gamma\pi_g^2\right) + (1-\lambda)\varepsilon W'(s,i)\left(-\pi_i^1 + \gamma\pi_i^2\right) + o(\varepsilon), \quad (11)$$

where W' denotes the derivative of W with respect to its first argument. Since $W(\cdot, i)$ is decreasing, $W'(s,i)$ is negative. Similarly, because $s \leq \bar{s}$ and $W(\cdot, g)$ is increasing on that domain, $W'(s,g)$ is positive. Since $\pi_g^1/\pi_g^2 < \pi_i^1/\pi_i^2$, we can choose γ between these two ratios. Doing so guarantees that $-\pi_g^1 + \gamma\pi_g^2$ is positive and $-\pi_i^1 + \gamma\pi_i^2$ is negative, which shows the claim. ■

Proposition 1 considers interim social welfare, after the crime has taken place. The incentives to commit the crime may *a priori* be influenced by the introduction of a third verdict, as (6) indicates. The proof of Proposition 1 shows that the welfare-improving sentences in the three verdict system can in fact be chosen in a way that does not increase the set of individuals who commit the crime. To see this, recall that the range of welfare-improving ratios γ for $s < \bar{s}$ is $\left[\pi_g^1/\pi_g^2, \pi_i^1/\pi_i^2\right]$, which is independent of the function $W(\cdot, g)$. For any $s > 0$, choosing $\gamma = \pi_g^1/\pi_g^2$ would not change, to a first order, the welfare for a guilty defendant, and would increase the welfare for an innocent defendant. Replacing $W(\cdot, g)$ with the individual's utility function $u(\cdot)$ and setting $\gamma = \pi_g^1/\pi_g^2$ would make a guilty defendant indifferent between the two- and three-verdict systems, so the left-hand side of (6) would not change. Therefore, the three-verdict system would deter all the individuals deterred by the two-verdict system.[27]

This observation immediately implies the following corollary of Proposition 1.

Corollary 1 *For any sentence $s > 0$ of the two-verdict system and any verdict technologies π_i, π_g, π_i^j, etc., there exists a three-verdict system with sentences s_1 and s_2 in which the set of individuals who commit the crime is no larger, and the interim and ex-ante welfare is strictly higher, than in the two-verdict system.*

[27]The utility of an innocent defendant would increase, so even more individuals would be deterred if the individual took into account the negligible probability he would be charged with the crime if he didn't commit it.

While the improvement in Proposition 1 does not increase the probability of punishing an innocent defendant (or a guilty one), an erroneously convicted defendant may face a harsher sentence ex-post when $s_2 > s$. The next next result shows that if the sentence associated with a conviction in the two-verdict system is optimal, then there is an improvement that does not increases the sentence.

Proposition 2 *1. Suppose that s^* is the optimal interim sentence in the two-verdict system, i.e., the one that maximizes $\mathcal{W}_2(s)$. Then, there exists an $s_1 < s^*$ such that the interim welfare in the three-verdict system with sentences s_1 and s^* is higher than in the two-verdict system, i.e., $\mathcal{W}_3(s_1, s^*) > \mathcal{W}_2(s^*)$. 2. Suppose that s^{**} is the optimal ex-ante sentence in the two-verdict system. Then, there exists an $s_1 < s^{**}$ such that the ex-ante welfare in the three-verdict system with sentences s_1 and s^{**} is higher than in the two-verdict system.*

The proof of Proposition 2, in Appendix A, shows that the results hold even when the original sentence is not optimal, as long as it is not too suboptimally lenient. Thus, it may be generally possible to improve upon the two-verdict system even under the strong restriction of not harming any innocent defendant more than in the two-verdict system.

2.2 The Bayesian conviction model

The analysis thus far did not impose any structure on how verdicts were determined. Because some later parts of the paper will require it, we now show how to specialize the setting to a class of verdicts based on the posterior probability that the defendant is guilty. Starting with the prior probability $\lambda = \eta_g/(\eta_g + \eta_i)$, the trial generates evidence that is used to form the posterior. This is summarized by distributions $F(\cdot|g)$ and $F(\cdot|i)$, which describe the posterior based on whether the defendant is actually guilty or innocent.[28] For expositional convenience, we assume that $F(\cdot|g)$ and $F(\cdot|i)$ have positive densities $f(\cdot|g)$ and $f(\cdot|i)$.

In a two-verdict system based on the defendant's posterior, it is natural to follow a cut-off rule. Appendix B shows that any "reasonable" verdict rule based on evidence in the two-verdict

[28] In order to match the prior λ, the distributions must satisfy the conservation equation

$$\lambda = E[p] = \lambda \int_0^1 p dF(p|g) + (1-\lambda) \int_0^1 p dF(p|i).$$

system can be formalized as a Bayesian model with posterior cut-off rule. If the posterior p is below a threshold p^*, then the defendant is acquitted, receiving a sentence of $s = 0$. If p exceeds p^*, then the defendant receives a sentence $s^* > 0$. The cutoff rule is a particular case of the previous section, with $\pi_g = Pr[p > p^*|g] = 1 - F(p^*|g)$ and $\pi_i = 1 - F(p^*|i)$.

The interim social welfare is given by

$$\begin{aligned}\mathcal{W}_2(p^*, s^*) = \lambda \left[(1 - F(p^*|g))W(s^*, g) + F(p^*|g))W(0, g)\right] + \\ (1 - \lambda) \left[(1 - F(p^*|i))W(s^*, i) + F(p^*|i)W(0, i)\right].\end{aligned} \tag{12}$$

Similarly, the ex-ante social welfare is given by

$$H(s^*) \left(\eta_g \left((1 - F(p^*|g))W(s^*, g) + F(p^*|g)W(0, g)\right) + \eta_i \left((1 - F(p^*|i))W(s, i) + F(p^*|i)W(0, i)\right) - h\right). \tag{13}$$

In what follows, we will denote by (p^*, s^*) the cutoff and sentence used in the two-verdict system. These variables may be chosen to maximize (12) or (13). In that case, they correspond to the interim or ex-ante utilitarian optimum for the two-verdict case.

2.3 Multi-verdict systems

Our analysis can be extended to more than three verdicts, and doing so prepares the ground for the general optimality result, in Section 3, concerning plea bargains. Granted an arbitrary number of verdicts, from an interim perspective one would wish to associate with each posterior belief p the sentence $s(p)$ that maximizes the welfare objective

$$pW(s, g) + (1 - p)W(s, i) \tag{14}$$

with respect to s. Rewriting the objective function as

$$\mathcal{W}(p, s) = p[W(s, g) - W(s, i)] + W(s, i),$$

we notice that it is supermodular in (p, s).[29] This implies that the selection of maximizers of (14) is isotone. In particular, there exists a nondecreasing selection $s(p)$ of optimal sentences. The same is true when choosing sentences to maximize the ex-ante welfare.

[29] $W(s, g)$ increases in s over the relevant range $[0, \bar{s}]$ while $W(s, i)$ is decreasing in s. This implies that $\partial\mathcal{W}/\partial p = W(s, g) - W(s, i)$ increases in s and, hence, supermodularity of $\mathcal{W}(p, s)$. See Milgrom and Shannon (1994).

The arguments used for Propositions 1, Corollary 1, and 2 easily generalize to yield the following results. For $k \geq 2$, we define a k-verdict system by a vector $(p_0, s_0, p_1, s_1, \ldots, p_{k-1}, s_{k-1})$ of strictly increasing cutoffs and sentences, with $p_0 = 0$, $p_{k-1} < 1$, $s_0 = 0$ and $s_{k-1} \leq \bar{s}$. In this system, a defendant receives sentence $s_{k'}$ whenever his posterior p lies in $(p_{k'}, p_{k'+1})$.

Proposition 3 *Suppose that the posterior distributions are continuous for both the guilty and innocent defendants. Then, for any k-verdict system there is a $k+1$ verdict system that strictly increases ex-ante and interim welfare. Moreover, if a k-verdict system is optimal among all k-verdict systems and $k \geq 2$, then there is a $k+1$-verdict system that strictly increases ex-ante and interim welfare and in which any defendant receives a weakly lower sentence.*

3 Plea bargaining

More than 90% of criminal cases in the United States conclude in a plea bargain instead of a trial. Plea bargains can be viewed a kind of intermediate verdict, which corresponds to an intermediate sentence that is lower than the one associated with a trial conviction. This verdict differs from what has been discussed so far, because it involves a strategic decision by the defendant of whether to accept the plea, in contrast to his passive role in a multi-verdict trial. As we shall see, this strategic aspect has a substantial impact on welfare.

We model pleas similarly to Grossman and Katz (1983)—hereafter "GK." In the first stage, the defendant is offered a plea sentence, denoted s^b. If the defendant accepts the plea, he gets this sentence and the case is concluded. If he rejects the plea, he goes to trial, where he is found either guilty or not guilty based on a signal structure like the one in the previous sections. The welfare functions $W(\cdot, i)$ and $W(\cdot, g)$ are concave and twice differentiable, and the defendant's utility coincides with society's welfare for an innocent individual, i.e., $W(\cdot, i) = u(\cdot)$.

GK, who consider two-verdict systems and interim welfare and do not consider multi-verdict systems or the incentives to commit the crime, show that the optimal system with a plea bargain is separating: the plea s^b is chosen to make a guilty defendant indifferent between taking the plea and going to trial, a guilty defendant takes the plea, and an innocent defendant goes to trial.

We now show that *any* multi-verdict system without pleas, no matter how many verdicts it has, can be improved upon in terms of interim welfare by a separating plea bargain system with

only two verdicts. In fact, we show that such a plea system with two verdicts is optimal within a much broader class of mechanisms. We then show that plea systems are often optimal in terms of ex-ante welfare, when incentives to commit the crime are taken into account. When they are not optimal, a plea bargain with two possible sentences is optimal, which may correspond to the uncertainty in punishment associated with some real-world plea bargains.[30]

3.1 The welfare value of plea bargaining

We denote by $t \in T$ the signal (evidence) generated during the trial. We assume that t is real-valued and denote by $F_g(t)$ and $F_i(t)$ respectively the signal distributions conditional on the defendant being guilty or innocent.[31] We assume that these distributions are absolutely continuous with positive densities $f_g(t)$ and $f_i(t)$. We also assume, without loss of generality, that the signal space is $T = [0, 1]$ and that the signals are ordered according to the monotone likelihood ratio property (MLRP): the density ratio $f_g(t)/f_i(t)$ is increasing in t (see Appendix B.1). Finally, a (measurable) multi-verdict system is a map $s : t \to s(t)$ from signals into sentences.

Given a multi-verdict system, the intuition for why there exists a separating plea bargain system with two verdicts that improves interim welfare is as follows. First, replace the multi-verdict system with a two-verdict system in which the defendant is either acquitted or punished severely, and set the signal conviction threshold such that a guilty defendant is indifferent between the two systems. Because a guilty defendant is more likely than an innocent defendant to generate a higher signal, the innocent defendant strictly prefers the two-verdict system. Choose a plea sentence that the guilty defendant is just willing to accept in lieu of going to trial. This sentence is higher than the expected trial sentence, because the defendant is risk averse. Welfare increases, because the sentence is higher and certain and society is risk averse.

Proposition 4 *For any multi-verdict system $s(\cdot)$, there exists a two-verdict system with a plea*

[30] According to the Federal Rules of Criminal Procedure, in an 11(c)(1)(B) plea agreement the court may impose a sentence other than the one stipulated in the agreement, and the defendant cannot withdraw his plea in this case.

[31] General evidence structures are discussed in Appendix B.1. If signals were multidimensional, we could order them according to their likelihood ratios and treat the resulting ratio as the signal, so that the real-valued assumption is without loss as long as the likelihood ratio of each signal is well-defined. For example, if T is a Borel subset of $\mathbb{R}^K$ for some dimension K, the ratios will be well defined as long as the signal distributions are absolutely continuous with respect to the Lebesgue measure induced over T and have positive densities.

that generates higher interim and ex-post welfare.

Proof. We begin by constructing a two-verdict system $\hat{s}$ that give the guilty defendant the same expected utility as $s(\cdot)$. In this system, there is a cutoff $\hat{t}$ below which the sentence is zero and above which the sentence is $s^M = \max_{t\in[0,1]} s(t)$. Moreover, the cutoff is chosen so that

$$U^g = \int_0^1 u(s(t))f_g(t)dt = \int_0^1 u(\hat{s}(t))f_g(t)dt = u(0)F_g([0,\hat{t}]) + u(s^M)F_g([\hat{t},1]) = \hat{U}^g, \qquad (15)$$

recalling that $u(s)$ denotes the defendant's utility from getting sentence s, and u is decreasing and concave. That such a $\hat{t}$ exists follows because the right-hand side of (15) is continuous in the cutoff t, ranging all values from $u(0)$ to $u(s^M)$, and because U^g clearly lies between $u(0)$ and $u(s^M)$ as a convex combination of utilities that lie in this interval. Moreover, the new verdict system increases the expected utility of an innocent defendant. To show this, notice that by construction we have

$$\int_0^{\tilde{t}} [u(\hat{s}(t)) - u(s(t))]f_g(t)dt \geq 0$$

for all $\tilde{t} \in [0,1]$. Since $f_i(t)/f_g(t)$ is positive and decreasing in t, this implies that[32]

$$\int_0^1 [u(\hat{s}(t)) - u(s(t))]f_i(t)dt \geq 0,$$

or

$$\hat{U}^i \geq U^i.$$

We now consider the guilty defendant's certainty equivalent s_g^{ce}, such that the guilty defendant is indifferent between getting s_g^{ce} for sure and going to trial in the two-verdict system. That is, s_g^{ce} satisfies

$$u(s_g^{ce}) = U^g = \hat{U}^g.$$

Since the guilty is indifferent, the innocent strictly prefers going to trial because i) guilty and innocent share the same utility function, but ii) an innocent defendant is less likely to be found guilty than a guilty one, so the trial is more appealing (see GK for a formal argument). We set

[32]The argument proceeds by a simple integration by parts. See Quah and Strulovici (2012, Lemma 4) for a similar proof in a more general environment. The claim may also be shown by showing that the defendant's expected utility has the single-crossing property in the defendant's type: the integrand has the single-crossing property in t and the type of the agent is affiliated with the posterior, which implies that the expected utility has the single-crossing property (see, e.g., Athey, 2002).

the plea sentence s^b to be the lower of s_g^{ce} and where $\bar{s}$ is the ideal sentence for a surely guilty defendant, that is,

$$s^b = \min\left\{s_g^{ce}, \bar{s}\right\}.$$

If $s_g^{ce} > \bar{s}$, then decrease the conviction threshold $\hat{t}$ so that the guilty is indifferent between getting $s^b = \bar{s}$ and going to trial in the two-verdict system with threshold $\hat{t}$. This further increases the innocent defendant's utility from going to trial.

Since the innocent benefits from the new verdict system, we will have shown that this system improves on the original one if we prove that the social welfare conditional on facing the guilty defendant is also higher. This welfare is equal to $W(s^b, g)$. If $s^b = \bar{s}$, then $W(\cdot, g)$ is maximized. Otherwise, $s^b = s_g^{ce} < \bar{s}$. Suppose this is the case. Because the defendant is risk averse (u is concave), s^b is greater than the average sentence $\tilde{s} = \int_0^1 s(t) f_g(t) dt$ that the guilty gets if he goes to trial. And because $W(\cdot, g)$ is concave, we have $W(\tilde{s}, g) \geq \int_0^1 W(s(t), g) f_g(t) dt$. Since $s^b \geq \tilde{s}$ and $W(\cdot, g)$ increases up to $\bar{s}$, we conclude that $W(s^b, g)$ dominates the expected social welfare conditional on facing the guilty.

In conclusion, the new two-verdict system with a plea improves social welfare regardless of whether the defendant is innocent or guilty. In particular, it is an improvement regardless of the prior distribution. The improvement is strict if either u or $W(\cdot, g)$ is strictly concave. ■

Proposition 4 shows that any multi-verdict system can be improved upon by some two-verdict system with a plea. This raises the question of whether other schemes, for example three-verdict systems with pleas, can do even better. By modifying the proof using a similar intuition, it is possible to prove that the answer is "no." To show this, we note that all the verdict systems, with and without pleas, may be seen as particular mechanisms. In a mechanism the defendant, who privately knows whether he is guilty or innocent, takes one of the actions available in the mechanism, and this action together with any additional information generated about the defendant's guilt is mapped into a sentence. It is well known from the mechanism design literature that in the present setting it is enough to consider direct revelation mechanisms in which it is optimal for the defendant to report his type truthfully: the defendant makes a report $\hat{\theta} \in \{g, i\}$ of his type ($g$uilty or $i$nnocent) and is then assigned a sentence $s(t, \hat{\theta})$ that depends on his report and on the signal t generated during trial. A mechanism is interim optimal if it maximizes interim welfare given the prior probability λ that the defendant is guilty.

Proposition 5 *For any sentence s^M, there is a unique interim optimal mechanism among those*

that assign sentences of at most s^M. This mechanism takes the form of a two-verdict system with a plea: $s(\cdot, g)$ is constant (i.e., like a plea) and no greater than $\bar{s}$, and $s(\cdot, i)$ is a two-step function, which jumps from 0 to s^M. The incentive compatibility constraint of the guilty defendant binds. The signal cutoff at which $s(\cdot, i)$ jumps from 0 to s^M decreases in the prior.

Proposition 5, whose logic is similar to Proposition 4, is proved in Appendix A. Propositions 4 and 5, which concern interim welfare, also have implications for ex-ante welfare. Considering the proofs of these propositions, if the guilty defendant's certainty equivalence s_g^{ce} in the proofs does not exceed $\bar{s}$, then each step of the proofs alters the mechanism in a way that increases welfare but leaves the expected utility of a guilty defendant unchanged. Thus, the two-verdict system with a plea that improves upon the original mechanism increases welfare and generates the same expected utility for a guilty defendant that the original mechanism did, so the set of individuals who commit the crime does not change. In this case, therefore, the ex-ante welfare is also improved. In particular, Proposition 5 characterizes the mechanism that maximizes ex-ante welfare among all mechanisms in which the certainty equivalence of the guilty does not exceed $\bar{s}$.

It may be, however, that deterrence optimally leads to sentences that are so high that s_g^{ce} exceeds $\bar{s}$. In this case, the improving mechanisms of Propositions 4 and 5, which increase interim welfare, lead to higher utility for guilty defendants. This increases the set of individuals who commit the crime and may therefore lower ex-ante welfare. The next result characterizes the ex-ante optimal mechanisms, and shows that this problem is optimally overcome by having the guilty defendant face at most two sentences, which are different from the ones faced by the innocent defendant.

Proposition 6 *For any sentence s^M, consider the ex-ante optimal mechanisms among those that assign sentences of at most s^M. There are two possibilities.*

1. There is a unique optimal mechanism, which takes the form of two verdicts with a plea: $s(\cdot, g)$ is constant, but may be greater than $\bar{s}$, and $s(\cdot, i)$ is a two-step function, which jumps from 0 to s^M. The signal cutoff at which $s(\cdot, i)$ jumps from 0 to s^M decreases in the prior.

2. There are infinitely many, essentially identical, optimal mechanisms. The guilty defendant faces a two-point lottery: function $s(\cdot, i)$ is as in 1 and function $s(\cdot, g)$ takes the same two values, which weakly exceed $\bar{s}$, and with the same probabilities, across all optimal mechanisms.

Moreover, every function that takes these two values with these probabilities is part of an optimal mechanism.

Proof. Consider a direct mechanism $s(\cdot,\cdot)$ in which it is optimal for the defendant to report his type truthfully. As in the proof of Proposition 5, we replace $s(\cdot,i)$ with a two-verdict system $\hat{s}(\cdot,i)$ with a cutoff $\hat{t}$, below which the sentence is zero and above which the sentence is s^M, so that the innocent defendant is indifferent between $s(\cdot,i)$ and $\hat{s}(\cdot,i)$ (that is, (24) holds). The guilty defendant prefers $s(\cdot,i)$ to $\hat{s}(\cdot,i)$; we increase $\hat{t}$ until he becomes indifferent between the two. This increases the utility of an innocent defendant, and therefore social welfare.

We now modify $s(\cdot,g)$. For this, denote by U^g the utility of a guilty defendant in the original mechanism, i.e.,

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

We would like to find the functions $\hat{s}(\cdot,g)$ that maximizes ex-ante social welfare subject to giving the defendant utility U^g. The ex-ante social welfare is given by a modification of (8) that accounts for the different sentencing schemes faced by guilty and innocent defendants:

$$H\left(\hat{s}(\cdot,g)\right)\left(\eta_g E_g\left(W\left(\hat{s}(\cdot,g),g\right)\right)+\eta_i E_i\left(W\left(\hat{s}(\cdot,i),i\right)\right)-h\right).$$

By (7), $H\left(\hat{s}(\cdot,g)\right)$ depends only on the expected utility U^g from $\hat{s}(\cdot,g)$. Thus, the optimal $\hat{s}(\cdot,g)$ are those that solve

$$\arg\max_{s(\cdot)} \int_0^1 W(s(t),g)f_g(t)dt \text{ s.t. } \int_0^1 u(s(t))f_g(t)dt = U^g,$$

where $s(\cdot)$ takes values in $\left[0,s^M\right]$.

To solve this problem, it is convenient to reformulate it in terms of the defendant's utility, that is,

$$\arg\max_{\hat{u}(\cdot)} \int_0^1 \hat{W}(\hat{u}(t))f_g(t)dt \text{ s.t. } \int_0^1 \hat{u}(t)f_g(t)dt = U^g, \tag{16}$$

where $\hat{u}(\cdot)$ takes values in $\left[u(0),u\left(s^M\right)\right]$ and $\hat{W}(U) = W\left(u^{-1}(U),g\right)$ for U in $\left[u(0),u\left(s^M\right)\right]$. The two formulations are equivalent, since $u(\cdot)$ is strictly decreasing.

The advantage of (16) is that the maximal value of the target integral is given by $\bar{W}\left(U^g\right)$, where $\bar{W}$ is the concavification of $\hat{W}$, from the definition of $\bar{W}(U)$ as $\sup\left\{x:(U,x)\in co\left(\hat{W}\right)\right\}$, where $co\left(\hat{W}\right)$ is the convex hull of the graph of $\hat{W}$.[33] Moreover, by this definition, if $\hat{W}\left(U^g\right) =$

[33] A more detailed discussion of this use of concavification appears, for example, in Kamenica and Gentzkow (2011), who apply it in a sender-receiver setting.

$\bar{W}(U^g)$, then this maximal value is achieved by the constant sentence $u^{-1}(U^g)$.[34] And if $\hat{W}(U^g) < \bar{W}(U^g)$, then the maximal value is achieved by randomizing between $u^{-1}(\underline{U})$ and $u^{-1}\left(\overline{U}\right)$, where $\underline{U} = \max\left\{v < U^g : \hat{W}(v) = \bar{W}(v)\right\}$ and $\overline{U} = \min\left\{v > U^g : \hat{W}(v) = \bar{W}(v)\right\}$, with probabilities α and $1-\alpha$ such that $\alpha\underline{U} + (1-\alpha)\overline{U}$.

In addition, because u monotonically decreases and W increases up to $\bar{s}$ and then decreases, $\hat{W}$ is single peaked at $u(\bar{s})$. It is also concave above $u(\bar{s})$, because of the concavity and monotonicity of W and u on $[0,\bar{s}]$. These two observations imply that $\hat{W}$ coincides with $\bar{W}$ above $u(\bar{s})$, so $U^g \geq u(\bar{s})$ is optimally achieved by a single sentence, and that if $U^g < u(\bar{s})$ is achieved by randomizing between two sentences, they both exceed $\bar{s}$.

Since guilty defendants are indifferent between this sentencing scheme and $\hat{s}(\cdot, i)$ (because they are indifferent between this sentencing scheme and $s(\cdot, g)$, and are indifferent between $s(\cdot, g)$ and $\hat{s}(\cdot, i)$), innocent defendants prefer $\hat{s}(\cdot, i)$. ■

Proposition 6 shows that a two-verdict system with a plea may be optimal in terms of ex-ante welfare even when the plea sentence exceed $\bar{s}$. This is not the case when two things happen. First, the optimal level $\hat{U}^g$ of utility for the guilty must be lower than $u(\bar{s})$, which happens 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 is sufficiently less risk averse than the individuals contemplating committing the crime, so $\hat{W}$ is not concave below $u(\bar{s})$, and in addition $\hat{W}\left(\hat{U}^g\right) < \bar{W}\left(\hat{U}^g\right)$.[35]

Under these conditions, the optimal sentence mapping for the guilty involves two sentences, both no lower than $\bar{s}$. All that matters for this mapping is the probability of each sentence, and not which signals are mapped to each sentence.[36] In particular, a mapping that correctly randomizes between the two sentences and completely disregards the signal is optimal. This is consistent with 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

[34] This is always the case for $U^g \geq u(\bar{s})$, because concavity and monotonicity of W and u on $[0,\bar{s}]$ imply that $\hat{W}$ is concave on $[0, u(\bar{s})]$, and therefore coincides with $\bar{W}$ on this range. This observation can also be used to provide an alternative proof for Proposition 5.

[35] For example, for $s^M = 4$, $u^{-1}(U) = \sqrt{-U}$, and $W(s) = -2 + s$ for $s \leq 2$ and $2 - s$ for $s > 2$, we have that for $U^g < -4$ the optimal scheme randomizes between $s = 2$ and $s^M = 4$.

[36] In contrast, the optimal mapping for an innocent defendant, which always involves two sentences, assigns the high sentence when the signal is sufficiently high.

the one specified without allowing the defendant to withdraw his plea.[37] Since the punishment in such pleas is determined without a trial, it does not depend on the signal that a trial would have generated.

Despite these results, plea bargains have been severely criticized for leading innocent defendants to accept jail time rather than go to trial. This may result from the fact that sentences given at trial are excessively harsh, which is a problem that has been pointed out repeatedly.[38] Section C provides an example that illustrates this idea. It should be noted, however, that many of the criticisms leveled at plea bargaining can, at least in principle, be addressed. In the United States, a defendant is entitled to competent counsel at the plea bargaining stage in all federal trials as well as in some state-level trials.

4 Value of evidence with a third verdict

The previous sections have taken as given the technology that generates evidence in favor of or against the defendant. Gathering evidence is costly, however, and the amount of evidence generated in a case depends on the incentives of the agents involved in the evidence-gathering process: law enforcement officers, prosecutors, experts, etc.

Setting aside the possible biases in these agents' behavior, the socially optimal amount of evidence to be gathered in a case clearly depends on the verdict structure. For example, a trial system in which a single verdict is given regardless of the evidence produced clearly eliminates any value of gathering evidence. This dependency underlies one criticism of plea bargaining, namely that it reduces the incentives for evidence gathering.

This section investigates the impact on evidence gathering of introducing a third verdict. For simplicity, we focus on the setting of Section 2.1 with the Bayesian conviction model. We consider interim welfare, since for any particular crime investigated it is plausible that the agents involved in the discovery stage of the trial are primarily concerned with the facts pertaining to the specific case.

A (possibly multi-) verdict system leads to welfare

$$w(p) = pW(s(p), g) + (1 - p)W(s(p), i), \tag{17}$$

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

[38] See for example Judge Rakoff's "Why Innocents Plead Guilty," in the *The New York Review*, (November 20, 2014) and Justice Kagan's opinion in Supreme Court Ruling No. 13-7451 on Yates vs. U.S.

where $p \mapsto s(p)$ is a step function that starts at zero, has two levels in a two-verdict system, and three levels in a three-verdict system. The welfare function $w(p)$ is piecewise linear. It start at 0, and decreases until a kink at which the sentence jumps from 0 to a positive level. Figure 1 represents the welfare function for the optimal two-verdict system when $W(\cdot, g)$ and $W(\cdot, i)$ are quadratic, for parameters given in the appendix.

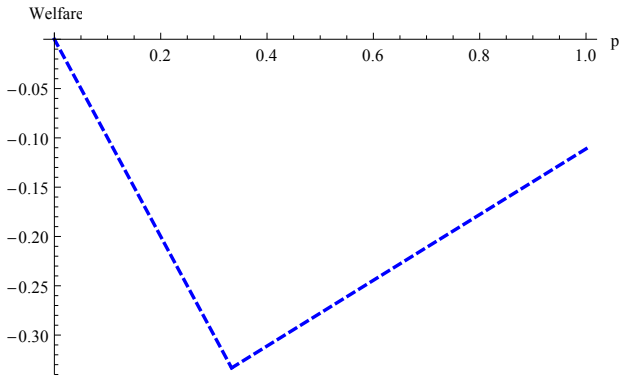

Figure 1: Welfare function, 2 verdicts.

The kink occurs at the cutoff $p^* = 1/3$, at which the sentence jumps from 0 to 2/3. Figure 2 represents the welfare function for the optimal three-verdict system obtained by adding an intermediate verdict and keeping the highest sentence at 2/3. The first cut-off is $p_1 = p^* = 1/3$, and the second cut-off is $p_2 = 1/2$. The welfare function is discontinuous at p_1: this reflects the fact that p_1 is not chosen optimally, but is rather "inherited" from the two-verdict system. In contrast, because p_2 is chosen optimally, the welfare function is kinked but continuous at p_2.

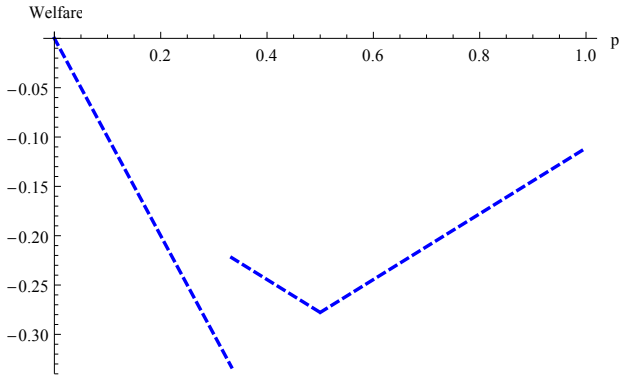

Figure 2: Welfare function, 3 verdicts.

Actual evidence formation processes are complex, involving various actors of different types – forensic experts, lawyers, witnesses—and different forms of evidence. To model evidence

formation, we must abstract from much of this complexity. Instead, we take the viewpoint of a social planner who may gather information until a verdict is reached.

The tradeoff at the heart of this task is clear: effort spent gathering evidence is costly, but provides information about the defendant's guilt. We discuss two ways to model this tradeoff (there are, of course, many others). The first is a one-shot binary evidence-gathering decision, which already captures the rough intuition for why two-verdict and three-verdict systems differ in their effects on evidence gathering. The second is a continuous evidence-gathering process, which provides a more visually appealing representation of the impact of a third verdict on evidence gathering.

4.1 One-shot evidence gathering

Suppose the planner decides whether to gather evidence, at a cost of $c > 0$. Starting with a prior p_0, the evidence returns a higher probability of guilt, say $p_0 + \Delta$ with probability $1/2$, and a lower probability $p_0 - \Delta$ also with probability $1/2$. The belief process is a martingale: the mean of the posterior p' is equal to the prior: $\frac{1}{2}(p + \Delta) + \frac{1}{2}(p - \Delta) = p$.

When is evidence gathering socially desirable? Suppose first that the prior is close to 0, so that the posterior p' surely lies below the cutoff p_1. Then, the additional evidence has no value as the defendant will be acquitted in all cases. Similarly, if p_0 is high enough for p' to lie above the cutoff p_1 no matter what, the additional evidence has no value as the defendant will be convicted regardless of p'. For p_0 slightly below p_1 and Δ such that $p_0 + \Delta$ lies above p_1, the value of evidence is positive, since it can lead to a conviction and increase welfare (relative to an acquittal) when it does. Similarly, evidence is valuable for p_0 slightly above p_1. Thus, evidence is valuable around the kink, where the welfare function is convex.

Consider now the case of three verdicts. For p_0 slightly below p_1, the value of evidence is higher than in the two-verdict case because a positive belief update triggers a large improvement in welfare (see Figure 2). For p in a neighborhood of p_2, the value of evidence is also positive due to the convex kink there, whereas it is 0 (for Δ small enough) in the two-verdict case.

For p_0 slightly above p_1 however, additional evidence may be more valuable in the two-verdict case, which creates a "doughnut hole:" additional evidence is more valuable in the three-verdict case than in the two-verdict case for more extreme beliefs, and less valuable in some intermediate region. This result is easier to visualize in the next model, where evidence gathering is more

gradual.

4.2 Continuous evidence gathering

Now suppose that evidence is gathered continuously. As long as evidence is gathered, a flow cost of c is incurred. During this time the belief p_t that the defendant is guilty evolves continuously in a way that is consistent with Bayesian updating, so that the closer the belief is to 0 or 1 the more slowly it evolves. Let the value function $v(p)$ denote the social welfare that arises from stopping optimally when the current belief is p. If it is optimal to stop immediately, then $v\,(p)$ coincides with $w\,(p)$ given by (17). Otherwise, it is optimal to continue collecting evidence, which leads to p changing continuously. In this case, $v\,(p)$ is the expectation of the value of stopping optimally in the future.

As in the one-shot setting, the value of gathering evidence in the two-verdict setting (formally analyzed in Appendix D) is high when the belief is very close to the convex kink p_1 in Figure 1. For beliefs slightly farther from the kink the value is still high because collecting evidence there leads with high probability to beliefs that are closer to the kink, for which the value of gathering evidence is high. The continuous evidence collection structure smooths the value of evidence collection as a function of the belief. This value decreases continuously as the belief moves away from the kink. For beliefs that are sufficiently far from the kink, it is optimal to stop immediately. Less informative evidence and higher costs of evidence collection lead to lower value of evidence collection, which correspond to functions v with to lower values. This is depicted in Figure 3 for parameters given in the appendix.

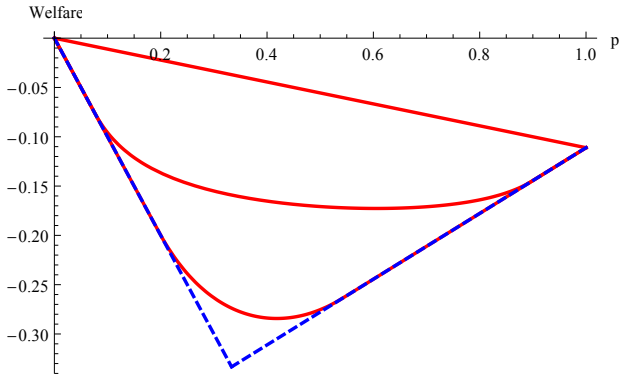

Figure 3: Value function, 2 verdicts, for varying cost levels.

Consider now the case of three verdicts in Figure 2. Around the kink p_2, the value function v

behaves similarly to the two-verdict case. Around the discontinuity p_1 the behavior is different, just like in the one-shot setting. Immediately to the left of the discontinuity the value of collecting evidence is high, so v substantially exceeds w, and this value decreases continuously for lower beliefs. Immediately to the right of the discontinuity there is no value in collecting evidence, so v coincides with w. Higher beliefs have a positive value of evidence collection, because of the kink. This value continuously decreases to 0 as the beliefs increase above the kink. This is depicted in Figure 4.

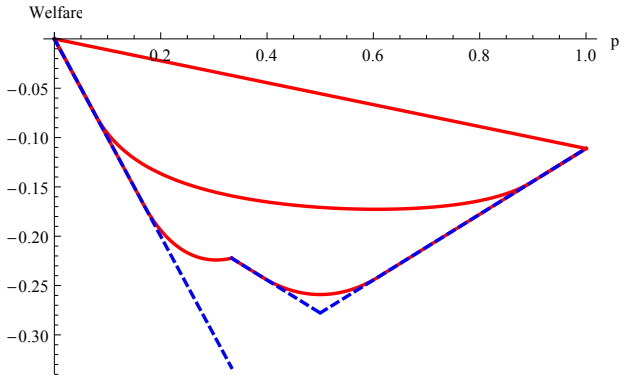

Figure 4: Value function, 3 verdicts, for varying cost levels.

In conclusion, the impact of switching to a three-verdict system by splitting the guilty verdict depends on how informative the evidence is and how costly it is to collect. When evidence is very informative, the posterior is unlikely to end up in the middle region, so the intermediate verdict has little impact. When finding new evidence is very costly, however, the posterior may end up in the middle region. The third-verdict system then increases the value of gathering evidence in two regions, below p_1 and around p_2, and decreases the value immediately above p_1. Overall, because $\tilde{p}_0 < \hat{p}_1$ and $\tilde{p}_2 > \hat{p}_2$, the three-verdict system results in evidence gathering at more extreme beliefs, where in the two-verdict evidence gathering has already stopped.

5 Intermediate "not guilty" verdict

Suppose now that those defendants who would be acquitted in the current two-verdict system now receive one of two verdicts, which we denote 1 and 2. Both verdicts are associated with no jail time, i.e., with $s = 0$. Verdict 1, which we refer to as "not guilty," obtains if the posterior is less than some cutoff $p^{iv} < p^*$, where p^* is the threshold for conviction, and verdict 2, which we refer to as "not proven," obtains if the posterior is between p^{iv} and p^*. We denote by p_i the

probability that a defendant is guilty conditional on verdict $i = 1, 2$. A conviction leads to the same sentence s^* as in the two-verdict system.

We assume that society observes the verdict at the end of the trial, but not the posterior regarding the defendant's guilt. The stigmatization associated with being charged and tried is modeled by a cutoff p^s, such that the defendant is stigmatized if the probability he is guilty conditional on the verdict exceeds p^s. We take p^s as exogenous, and assume that convicting a defendant is more demanding than stigmatizing him, so $p^s < p^*$.[39] We also assume that if the defendant is completely cleared in the trial and the public were fully aware of this, then he would not be stigmatized. That is, $\underline{p} < p^s$, where $\underline{p}$ is the lowest possible posterior. An innocent defendant who is stigmatized lowers welfare by $d^i > 0$, and a guilty defendant who is stigmatized increases welfare by $d^g > 0$.[40] We are interested in the optimal cutoff p^{iv} and the conditions under which introducing the additional verdict increases welfare. For expositional simplicity we will consider interim welfare; the same qualitative results hold for ex-ante welfare.

The relevant part of the welfare function in the two-verdict system is

$$\lambda \left[W(0, g) + 1_{p^{ng} > p^s} d^g \right] + (1 - \lambda) \left[W(0, i) - 1_{p^{ng} > p^s} d^i \right],$$

where p^{ng} is the probability that a defendant is guilty conditional on being acquitted, since whether an acquitted defendant is stigmatized depends on whether p^s is lower or higher than p^{ng}. We consider these two possibilities below.

Suppose first that $p^{ng} \geq p^s$, so an acquitted defendant in the two-verdict system is stigmatized. For any p^{iv}, it must be that $p_2 \geq p^{ng} \geq p^s$, so the defendant is stigmatized if he is found "not proven" in the three-verdict system. The split can have an effect on social welfare only if $p_1 \leq p^s$, in which case the defendant is not stigmatized if he is found "not guilty" in the three-verdict system. Therefore, consider p^{iv} such that $p_1 < p^s$. Eliminating the stigma when the defendant is found "not guilty" increases the relevant part of the welfare function by

$$-\lambda \sum_{p \leq p^{iv}} f(p|g)\, d^g + (1 - \lambda) \sum_{p \leq p^{iv}} f(p|i)\, d^i.$$

For a given posterior $p \leq p^{iv}$ the increase is

$$-\lambda f(p|g)\, d^g + (1 - \lambda) f(p|i)\, d^i > 0 \iff \frac{f(p|g)}{f(p|i)} < \frac{(1 - \lambda)\, d^i}{\lambda d^g}. \tag{18}$$

[39] This implies that the analysis of Section 2.1 does not change as a result of the stigma, since a defendant who receives verdicts 1 or 2 is stigmatized.

[40] A similar analysis can be conducted for $d^i \leq 0$ and/or $d^g \leq 0$.

Since $f(p|g)/f(p|i)$ increases in the posterior p, a fact we show in Appendix B.1, we obtain the following result.

Proposition 7 *Suppose that being acquitted in the two-verdict system carries a stigma. Then, optimally splitting the acquittal into "not guilty" and "not proven" increases interim welfare if and only if* $\frac{f(\underline{q}|g)}{f(\underline{q}|i)} < \frac{(1-\lambda)d^i}{\lambda d^g}$.

If the condition in Proposition 7 holds, then the optimal cutoff p^{iv} is the minimum between the highest posterior for which (18) holds and the highest posterior such that $p_1 \leq p^s$. Notice that the condition in Proposition 7 is satisfied more easily if the defendant is more likely to be innocent (λ decreases), the stigma for the innocent increases, or the stigma for the guilty decreases.

Now suppose that $p^{ng} < p^s$, so an acquitted defendant in the two-verdict system is not stigmatized. The split can have an effect on social welfare only if $p_2 > p^s$, in which case the defendant is stigmatized if he is found "not proven" in the three-verdict system. Therefore, consider p^{iv} such that $p_2 > p^s$. Stigmatizing the defendant when he is found "not proven" increases the relevant part of the welfare function by

$$\lambda \sum_{p>p^{iv}} f(p|g)\, d^g - (1-\lambda) \sum_{p>p^{iv}} f(p|i)\, d^i.$$

For a given posterior $p > p^{iv}$ the increase is

$$\lambda f(p|g)\, d^g - (1-\lambda) f(p|i)\, d^i > 0 \iff \frac{f(p|g)}{f(p|i)} > \frac{(1-\lambda)\, d^i}{\lambda d^g}. \tag{19}$$

Since $f(p|g)/f(p|i)$ increases in the posterior p, we obtain the following result.

Proposition 8 *Suppose that being acquitted in the two-verdict system does not carry a stigma. Then, optimally splitting the acquittal into "not guilty" and "not proven" increases interim welfare if and only if* $\frac{f(p^*|g)}{f(p^*|i)} > \frac{(1-\lambda)d^i}{\lambda d^g}$.

If the condition in Proposition 8 holds, then the optimal p^{iv} is the maximum between the lowest posterior for which (19) holds and the lowest posterior such that $p_2 \geq p^s$. Notice that the condition in Proposition 8 is satisfied more easily if the defendant is more likely to be guilty (λ increases), the stigma for the innocent decreases, or the stigma for the guilty increases.

6 Reflecting residual doubt in the current justice system

The most explicit inclusion of residual doubt in the U.S. criminal justice system concerns the determination of death sentences. In capital cases, juries must decide, after returning a guilty verdict, whether the defendant should get the death penalty. In this penalty phase, residual or "lingering" doubt may be used as a mitigating circumstance to reject the death penalty.[41] The Capital Jury Project—an academic survey of past jurors in capital cases—has found that lingering doubt was the most important mitigating factor identified by jurors.

There is, however, wide variation in how residual doubt is applied. First, the U.S. penal code (Title 18, §3592) does not explicitly mention residual doubt in its list of mitigating factors, although it does state that mitigating circumstances are not limited to this list. In some cases, jurors are not informed that lingering doubt is a valid mitigating circumstance.[42] In *Franklin v. Lynaugh* (1988), the U.S. Supreme Court rejected a defendant's right to invoke residual doubt at the penalty stage, while in *People v. McDonald* (Supreme Court of Illinois, 1995) a trial judge refused to answer jurors' question on the issue, a decision which was later affirmed by the Supreme Court of Illinois.

Compounding this inconsistency, there is empirical evidence that many jurors get confused with the voting rules used to establish aggravating and mitigating circumstances at the penalty stage. While the unanimity rule is required to find a circumstance aggravating, no such standard exists for mitigating circumstances. The Capital Jury Project found, however, that 45% of jurors failed to understand that they were allowed to consider any mitigating evidence during the sentencing phase of the trial, not just the factors listed in the instructions.[43]

When sentencing is performed by a trial judge, the invocation of residual doubt can be highly controversial. In *State v. Krone* (Arizona Supreme Court, 1995) a trial judge sentenced to life in prison a defendant found guilty of murder, citing doubt about whether he was the true killer.

[41]The more demanding requirement of proving guilty beyond "all doubt" has been discussed in some states, such as the bill proposed in 2003 by then Illinois House Republican leader Tom Cross. Some death penalty advocates have countered that it was impossible to prove anything beyond *all* doubt, and that the bill would in effect rule out the death penalty. Various degrees of lingering doubt have been discussed (e.g., Sand and Rose, 2003) without any mathematical formalism.

[42]See, e.g., *People v. Gonzales and Soliz*, California Supreme Court, 2009.

[43]The CJP's findings concerning jurors' understanding of instructions are summarized at http://www.capitalpunishmentincontext.org/issues/juryinstruct.

In their legal textbook, Dressler and Thomas (2010, pp. 57–61) comment that this decision "borders on the unbelievable." They do not, however, suggest an alternative solution.

In non-capital cases, only five states permit juries to make the sentencing decision. Outside of these states, residual doubt can thus only be expressed by the sentencing judge, whose opinion does not necessarily reflect the views of the jury. Again, residual doubt is not listed as a mitigating factor in sentencing guidelines.

The fact that residual doubt should only be considered in capital cases seems largely arbitrary. Even comparably less serious cases can carry large sentences, resulting in extreme punishments for defendants who are found guilty but for whom residual doubt remains. 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. The explicit inclusion of residual doubt in sentencing would have likely avoided such an extreme outcome.[44]

Some felonies provide an indirect way of expressing doubt by using the lesser-included-offense rule: juries can return a manslaughter verdict, rather than a first- or second-degree murder verdict, or a larceny verdict instead of a robbery verdict. However, each of these verdicts corresponds to a precise charge (e.g., whether premeditation and malice aforethought were involved) and doubt about a particular charge can only be imperfectly expressed by returning a guilty verdict on a lower count. These instruments only offer a limited and, in fact, improper, way of reflecting residual doubt. Furthermore, the less-included-offense rule is not a constitutional right of the defendant; its application is therefore to some extent arbitrary and depends on the inclination of the jury (see Mascolo (1986)).

Even when the lesser-included-offense rule does not apply, residual doubt may be reflected by returning a guilty verdict only on a subset of the charges brought against the defendant. There is anecdotal evidence that such compromise is sometimes used by the jurors to reflect doubt. In

[44]Capital sentences are unique in their irreversibility, which creates an additional reason for avoid this sentence in case of lingering doubt: exonerating evidence may appear after the execution of the defendant, preventing any release and compensation. In practice, this fundamental difference is attenuated by the fact that death-row defendants spend many years in jail before their execution until all recourses have been exhausted, while non-capital defendants serving long sentences may die in jail, which also prevents any release or compensation.

the aforementioned State v. May, for instance, Nelson (2013) notes that "it seems likely that the defendant molested either all of the children or none of them. So why did the jury ultimately reach a verdict of guilty on five counts and not guilty on two? The answer is that the jurors compromised." Dropping some charges is, however, a very coarse instrument to incorporate residual doubt: for example, this approach cannot be used to reduce the sentence of a defendant facing a single but severe count, while it may be used for a defendant facing several counts, the sum of which adds to the same aggregate maximal sentence as in the single-count case.[45] Even when it is feasible, the approach exposes the defendant to another idiosyncratic component of the jury—whether it is sophisticated or willing enough to use this compromise strategy—introducing a source of jury heterogeneity in trial outcomes even for otherwise identical cases.[46]

The U.S. justice system incorporates residual doubt about a defendant's guilt in two other ways. First, a defendant found not guilty in a criminal trial may still be found guilty in a civil suit, which uses the less demanding preponderance-of-evidence standard of proof. However, civil suit sentences carry no jail time and thus may be more limited in preventing recidivism. Furthermore, the connection between criminal and civil trials is generally limited, preventing any coordination and coherent decision across these trials. Second, residual doubt variations also imply different likelihoods of post-trial events such as successful appeals and exonerations, which affect the defendant's ultimate punishment. These events are largely beyond the control of the first court and are not a close substitute for the additional verdicts introduced here.

In summary, the current criminal justice system includes various ways of reflecting residual doubt in outcomes and it appears that these ways are used purposefully by some actors of the system. However, these ways are largely arbitrary, inconvenient, and uncoordinated. This paper proposes a structured, systematic approach for the consideration of residual doubt in criminal justice decisions and explicit designs which are shown to improve welfare in many settings.

[45]The set of charges leveled at the defendant may also be affected by the strategic decisions of the prosecutor, which increases the prosecutor's power and adds to the complexity of this problem.

[46]It should also be noted that under the current law, such compromise is actually illegal if it results from a bargaining between pro-acquittal and pro-conviction jurors. Such an arrangement currently violates the rights of the defendant if the pro-acquittal jurors still believe that the defendant should be found not guilty (Mascolo (1986)).

7 Implementation and jurors' reactions to additional verdicts

Implementation: verdicts vs. sentences

Formalizing the intermediate sentence introduced in this paper as an intermediate *verdict* is consistent with the not-proven verdict, discussed in Section 5, used by some criminal justice systems. In this formulation, the jury must decide, according to some collective rule, among the three verdicts.

An alternative "two-step" implementation maintains the current separation between the fact-finding and sentencing stages. The verdict outcome is still binary ("guilty" or "not guilty"), and residual doubt is expressed in the form of intermediate sentences decided in the sentencing stage.

The second implementation presents a significant advantage: in principle, the jury can be given exactly the same instructions as in the current system, which allows to cleanly split the set of cases which would receive a "guilty" verdict under the current system into multiple sentence levels reflecting the strength of evidence, and thus leaves unchanged the probability of acquitting the defendant.

Intermediate sentences can be decided in a variety of ways, which may involve a sentencing judge, sentencing guidelines (e.g., automatically rule out the death penalty if the evidence is solely based on a confession), or a jury.

Regardless of the implementation, a potential concern is how the jury may react to additional verdicts. The remainder of our discussion focuses on this issue.

Jurors' reaction to additional verdicts

Jury decisions involve collective and psychological considerations: jurors may have limited and uneven ability to understand jury instructions or interpret the evidence, have varied tolerance for erroneous convictions and acquittals, and are subject to individual biases and to persuasion and group-think dynamics, to cite only a few issues. Even abstracting from these issues, jury decisions are difficult to analyze.[47]

[47]Austen-Smith and Banks (1996), Feddersen and Pesendorfer (1996, 1997), and Gerardi and Yariv (2007) identify important informational effects, which may arise even when all jurors have identical preferences. A central mechanism in this literature is that, conditional on being pivotal in a vote, a rational juror may put so

The literature on criminal trial design varies from fully rational to completely reduced-form models of jury behavior. At the most "rational" extreme, Lee (2015) considers jurors who perfectly take into account how prosecutors select the pool of defendants who go to trial. Prosecutors can influence this pool by choosing the plea sentence that they propose to defendants before the trial.[48] Other papers on trial design (Kaplow (2011), Daughety and Reinganum (2015a,b), Da Silveira (2015), Silva (2015)) abstract from any jury decision, focusing on reduced-form thresholds or on a mechanism design approach without jurors.

A key observation is that our Propositions 1 and 2 continue to hold under the two-step implementation mentioned above, provided that jurors are given the same instructions as in the current system to decide between the guilty and not-guilty verdicts, and react to these instructions in the same way, no matter how imperfect, as they currently do. No matter how "tough on crime" or otherwise biased each juror is, and what voting, persuasion or other collective processes are at play, all these components would play out in exactly the same way at the fact-finding stage, under a standard binary verdict, as in the first step of the two-step approach, guaranteeing that no more defendants are found guilty in the three-verdict system than in the current one.

The main question, therefore, is to what extent jurors would know and incorporate in the fact-finding stage the fact that residual doubt may play a significant role in the sentencing stage.

In practice, there is little evidence that jurors incorporate sentencing considerations into their verdict decisions. On the contrary, in recent history judicial practice has been to keep the jury uninformed about the punishment faced by the defendant (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 entirely focus on describing the procedure for finding facts. In many cases—such as *People v. May* above—jurors are unaware of the minimum-punishment guidelines relevant for the case.

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

much weight on other jurors' signals that he significantly discounts, and potentially discards, his own information.

[48]The approach presumes that jurors are aware of the plea sentence offered to the defendant. In practice, the jury is often instructed to consider only the evidence produced at trial.

increases with the sentence that they face.[49] Such a correlation cannot be easily explained away by prosecutor behavior: if, in particular, 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 trial sentence. This relation seems contradicted by the data.[50]

More generally, there is strong evidence that jurors have a limited understanding of the sentences faced by defendants. For example, the aforementioned Capital Jury Project found that most jurors "grossly underestimated" the amount of time spent in jail entailed by a guilty verdict. It is reasonable to believe that jurors would be as unaware of, say, maximum-sentencing guidelines as they currently are of minimum-sentencing guidelines.

Finally, if contrary to expectations jurors incorporated the intermediate verdict into their decision, they might adopt a different standard of proof to convict defendants, knowing that the corresponding cases would result in a different sentence than in the current system. To the extent that jurors did this with the social welfare objective in mind, such a change would likely be beneficial. Jurors may, however, have their own objective in mind. For example, they may ignore, from the interim perspective in which they are placed, the deterrence value, ex ante, of higher expected punishments—this issue arises even in a two-verdict system, and may explain the fact, mentioned earlier, that jurors are specifically asked to focus on finding facts and left relatively uninformed about the strength of the punishment implied by a guilty verdict. Jurors may also worry about the length of deliberation, and be willing to continue deliberation only if the social value of doing so is high. The analysis of Section 4 suggests that this value is not lowered by the introduction of an intermediate verdict, and may in fact be higher for a wide range of beliefs.

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

[50]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.

A Omitted proofs

A.1 Proof of Proposition 2

Consider the first part of the proposition. By construction s^* maximizes $\mathcal{W}_2(s)$ with respect to s. In particular, $s \leq \bar{s}$. Since all sentences are interior, s^* must satisfy the first-order condition

$$\lambda \pi_g W'(s^*, g) + (1-\lambda)\pi_i W'(s^*, i) = 0. \tag{20}$$

Now consider the derivative of $\mathcal{W}_3(s_1, s^*)$ with respect to s_1, evaluated at $s_1 = s^*$. From (11), we have

$$\left.\frac{\partial \mathcal{W}_3(s_1, s^*)}{\partial s_1}\right|_{s_1 = s^*} = \lambda \pi_g^1 W'(s^*, g) + (1-\lambda)\pi_i^1 W'(s^*, i). \tag{21}$$

Since $\frac{\pi_g^1}{\pi_i^1} < \frac{\pi_g}{\pi_i}$, $W'(s^*, g) > 0$, and $W'(s^*, i) < 0$, the first-order condition (20) implies that the right-hand side of (21) is strictly negative. This shows that decreasing s_1 below s^* strictly improves welfare, yielding the desired improvement.

For the second part of the proposition, differentiating (4) gives the first-order condition satisfied by s^{**} in the two-verdict system:

$$\frac{dH(s^{**})}{ds}\left[\eta_g\left(\pi_g W(s^{**}, g) + (1-\pi_g)W(0, g)\right) + \eta_i\left(\pi_i W(s^{**}, i) + (1-\pi_i)W(0, i)\right) - h\right] + H(s^{**})\left(\eta_g \pi_g W'(s^{**}, g) + \eta_i \pi_i W'(s^{**}, i)\right) = 0. \tag{22}$$

The equivalent derivative for the three-verdict case with respect to s_1 at s^{**} is

$$\frac{dH_1(s^{**})}{ds_1}\left[\eta_g\left(\pi_g W(s^{**}, g) + (1-\pi_g)W(0, g)\right) + \eta_i\left(\pi_i W(s^{**}, i) + (1-\pi_i)W(0, i)\right) - h\right] + H(s^{**})\left(\eta_g \pi_g^1 W'(s^{**}, g) + \eta_i \pi_i^1 W'(s^{**}, i)\right), \tag{23}$$

where $H_1(s_1)$ denotes the fraction of individuals who commit the crime in the three-verdict system as a function of the punishment s_1 when $s_2 = s^{**}$ (in particular, $H_1(s^{**}) = H(s^{**})$).

Dividing (22) by π_g we obtain

$$\frac{1}{\pi_g}\frac{dH(s^{**})}{ds}\left[\eta_g\left(\pi_g W(s^{**}, g) + (1-\pi_g)W(0, g)\right) + \eta_i\left(\pi_i W(s^{**}, i) + (1-\pi_i)W(0, i)\right) - h\right] + H(s^{**})\left(\eta_g W'(s^{**}, g) + \eta_i \frac{\pi_i}{\pi_g} W'(s^{**}, i)\right) = 0.$$

Dividing (23) by π_g^1 we obtain

$$\frac{1}{\pi_g^1}\frac{dH_1(s^{**})}{ds_1}\left[\eta_g\left(\pi_g W(s^{**}, g) + (1-\pi_g)W(0, g)\right) + \eta_i\left(\pi_i W(s^{**}, i) + (1-\pi_i)W(0, i)\right) - h\right] + H(s^{**})\left(\eta_g W'(s^{**}, g) + \eta_i \frac{\pi_i^1}{\pi_g^1} W'(s^{**}, i)\right).$$

From (3) and (7) we have

$$\frac{1}{\pi_g}\frac{dH(s^{**})}{ds} = \frac{1}{\pi_g^1}\frac{dH_1(s^{**})}{ds_1}.$$

Since $W'(s^{**}, i) < 0$, to prove that the derivative for the three-verdict case is negative, which would demonstrate the claim in the statement of the proposition, it suffices that $\pi_i^1/\pi_g^1 > \pi_i/\pi_g$, which holds by definition of verdict 1.

A.2 Proof of Proposition 5

Consider a direct mechanism $s(\cdot,\cdot)$ in which it is optimal for the defendant to report his type truthfully. We begin by replacing $s(\cdot,i)$ with a two-verdict system $\hat{s}(\cdot,i)$ with a cutoff $\hat{t}$ below which the sentence is zero and above which the sentence is s^M. The cutoff is chosen so that the innocent defendant is indifferent between $s(\cdot,i)$ and $\hat{s}(\cdot,i)$, that is,

$$U^i = \int_0^1 u(s(t,i))f_i(t)dt = \int_0^1 u(\hat{s}(t,i))f_i(t)dt = u(0)F_i([0,\hat{t}]) + u(\bar{s})F_i([\hat{t},1]) = \hat{U}^i. \tag{24}$$

The guilty defendant prefers $s(\cdot,i)$ to $\hat{s}(\cdot,i)$, i.e., his incentive compatibility continues to hold, when $s(\cdot,i)$ is replaced with $\hat{s}(\cdot,i)$. This is because by construction we have

$$\int_0^1 [u(s(t,i)) - u(\hat{s}(t,i))]f_i(t)dt = 0,$$

and since $h(\cdot) = u(s(\cdot,i)) - u(\hat{s}(\cdot,i))$ crosses 0 once from below on $[0,1]$ and $f_i(t)/f_g(t)$ is positive and decreasing in t, we obtain (see the previous footnote)

$$\int_0^1 [u(s(t,i)) - u(\hat{s}(t,i))]f_g(t)dt \geq 0.$$

Thus, because the guilty defendant prefers $s(\cdot,g)$ to $s(\cdot,i)$, he also prefers $s(\cdot,g)$ to $\hat{s}(\cdot,i)$. Now replace $s(\cdot,g)$ with the certainty equivalence s_g^{ce} such that the guilty defendant is indifferent between s_g^{ce} and $s(\cdot,g)$, that is,

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

This increases welfare because the guilty defendant and society are risk averse, as in the proof of Proposition 4. Let $s^b = \min\left\{s_g^{ce}, \bar{s}\right\}$, which increases welfare if s_g^{ce}, because $W(\cdot,g)$ decreases above $\bar{s}$.

Because the guilty defendant is indifferent between s_g^{ce} and $s(\cdot,g)$, he prefers s^b to $\hat{s}(\cdot,i)$. If the preference is strict, modify $\hat{s}(\cdot,i)$ by increasing $\hat{t}$ until the guilty defendant is indifferent between s^b and $\hat{s}(\cdot,i)$. This increases welfare since it increases the utility of the innocent defendant, and also guarantees that the innocent defendant prefers $\hat{s}(\cdot,i)$ to s^b (because the guilty defendant is indifferent between the two). This shows that the optimal mechanism is of the form described in the statement of the proposition, and that the incentive constraint of the guilty defendant binds. Thus, each such mechanism is pinned down by the cutoff $\hat{t}$. It is straightforward to see that the welfare-maximizing $\hat{t}$ decreases in the prior λ.

B Foundation of the Bayesian Conviction Model

We now study whether actual court proceedings can be translated into a Bayesian updating process and a threshold. We address this by considering an evidence-based trial technology. There is a set X of evidence elements, and "evidence collection" refers to a subset of X. The court technology is a mapping $D: 2^X \to \{G,N\}$, which for every evidence collection decides whether the defendant is guilty or not guilty.[51] Distributions P_θ on 2^X, for $\theta \in \{g,i\}$, describe the probability that different evidence collections arise conditional on the defendant being actually guilty or innocent. We assume that both distributions have full support. Letting π_θ^k denote the probability that a defendant of type θ receive verdict k, we have $\pi_\theta^k = P_\theta\left(D^{-1}(k)\right)$ for each type θ and verdict k in $\{G,N\}$. Recall that $\pi_i^G < \pi_g^G$, i.e., $P_i\left(D^{-1}(G)\right) < P_g\left(D^{-1}(G)\right)$, and that λ is the prior that the defendant is guilty. We ask several questions.

[51] The analysis can be generalized to stochastic decisions.

1. Given D, P_i, P_g, and λ, can D be rationalized as the result of Bayesian updating with a threshold on the posterior for determining guilt? At a minimum, this would require D to respect "incriminating" and "exculpatory" evidence sets, which are determined by whether they indicate that the defendant is more likely to be guilty than innocent.

2. Given D and λ, can P_i and P_g be chosen to rationalize D as the result of Bayesian updating with a threshold on the posterior for determining guilt?

3. Given λ, can D, P_i, and P_g be chosen to rationalize D as the result of Bayesian updating with a threshold on the posterior for determining guilt?

To answer these questions, we formally order defendant types i and g so that $i < g$, and we order verdicts as $N < G$. Then, we say that D ***can be rationalized*** as the result of Bayesian updating with a threshold on the posterior if for every $E, E' \subseteq X$ we have $D(E) < D(E')$ if and only if the posterior that the defendant is guilty is higher under E' than under E, i.e.,

$$\frac{\lambda P_g(E)}{\lambda P_g(E) + (1-\lambda) P_i(E)} < \frac{\lambda P_g(E')}{\lambda P_g(E') + (1-\lambda) P_i(E')}.$$

This condition is equivalent to $\lambda P_g(E)(\lambda P_g(E') + (1-\lambda) P_i(E')) < \lambda P_g(E')(\lambda P_g(E) + (1-\lambda) P_i(E))$ and, after rearranging, to

$$\frac{P_g(E)}{P_i(E)} < \frac{P_g(E')}{P_i(E')}.$$

The likelihood ratios are thus ordered independently of λ. For every evidence set $E \subseteq X$, denote by $r(E) = P_g(E)/P_i(E)$ its likelihood ratio. This shows the following proposition.

Proposition 9 *D can be rationalized if and only if for every $E, E' \subseteq X$ the following holds:*

$$r(E) \leq r(E') \Rightarrow D(E) \leq D(E').$$

While we started with a Bayesian definition of rationalizability, this concept is in fact non-Bayesian: it is purely based on the likelihood ratio of guilty given the observed evidence and, in particular, is independent of any prior.

Equipped with this result, we can answer the questions above. For 1, the answer is "yes" if and only if

$$\max\{r(E) : D(E) = N\} < \max\{r(E) : D(E) = G\}. \tag{25}$$

For 2, the answer is "yes:" choose P_g and P_i so that (25) holds. Since 2 implies 3, that answer to 3 is also "yes."

Incriminating and exculpatory evidence: definitions and properties

When D can be rationalized, we say that evidence $e \in X$ is **D-incriminating** if for every $E \subseteq X$ with $e \notin E$, $D(E) = g$ implies that $D(E \cup \{e\}) = g$. We say that evidence $e \in X$ is *P-incriminating* if for every $E \subseteq X$ with $e \notin E$ we have that $r(E) \leq r(E \cup \{e\})$. Decision- and belief-based notions of exculpatory evidence are defined similarly. The following result establishes the logical connection between these concepts.

Proposition 10 *If D is rationalized by P, any P-incriminating evidence is also D -incriminating.*

The reverse need not hold: one can easily construct examples in which some evidence collection E suffices to convict the defendant (i.e., $D(E) = g$) and the additional piece of evidence e reduces the 'guilt' ratio ($r(E \cup \{e\}) < r(E)$), but not enough the change the decision ($D(E \cup \{e\}) = g$).

Our definition and characterization of rationalization extend without change to probabilistic functions D, in which the image of D is the probability that the defendant is found guilty.

B.1 Ordering posterior distributions with the MLRP

In the Bayesian conviction model, the posterior belief is formed by combining a prior with the signals observed about the defendant. One may view each evidence collection E as a signal, and signals may be ordered according to the likelihood ratio $r(E)$. The distributions P_i and P_g over evidence collections can then be mapped into distributions over likelihood ratios r. In a Bayesian conviction model, only the likelihood ratio matters for the decision, and one can thus without loss identify any signal with r. Thus, without loss, signals may be ranked according to this likelihood ratio. Let R_g and R_i denote the distributions of r, conditional on being guilty and innocent, respectively. When the signal distributions, conditional on being guilty or innocent, are continuous, let ρ_g and ρ_i denote their densities. By construction, we have $\rho_g(r)/\rho_i(r) = r$. In statistical terms, this means that R_g and R_i are ranked according the MLRP: the ratio of their density is increasing in the signal. Moreover, because the posterior $p(r)$, given a signal r, is equal to the conditional probability of $\theta = g$ given r, it inherits the MLRP.[52] Let F_g and F_i denote the distributions of p, conditional on being guilty and innocent, respectively, and let f_g and f_i denote the densities of F_g and F_i (which exist as long as R_g and R_i are continuous), we have $f_g(p)/f_i(p)$ is increasing in p.

Proposition 11 *Suppose that both signal distributions, conditional on being guilty and innocent, are continuous. Then both distributions of the posterior p are continuous, and their density functions satisfy the MLRP.*

This property, which holds without loss (except for the continuity assumption, of a technical nature), plays a key role for subsequent results.

C Plea bargaining with excessive sentences (Online)

We introduce a model in which some innocent defendants indeed take the plea. Following GK, we achieve this by introducing two types of innocent defendants, which vary according to their degree of risk aversion. To simplify the analysis, we assume that there are three types of defendants in equal proportion: risk neutral guilty defendants with utility $u(s) = -s$, risk neutral innocent defendants with the same utility, and risk averse innocent defendants with a piecewise linear utility function given by $u(s) = -\frac{3}{16}s$ for $s \leq 16$ and $u(s) = -3 - 2(s - 16)$ for $s \in [16, 20]$. Again for simplicity, we assume that the social welfare as a function of the guilty defendant's punishment is linear with a peak at 20 years: $W(s, g) = -|s - 20|$. We thus only consider sentences lower than the sentence $\bar{s} = 20$ that is optimal if the defendant is known to be guilty.

Finally we suppose that the trial can generate two types of evidence against the defendant, weak or strong. A guilty defendant generates strong evidence with probability 30% and weak evidence with probability 50%. An innocent defendant generates (regardless of his risk aversion) strong evidence with probability 10% and weak evidence with probability 30%. When no evidence is found against the defendant, he is acquitted.

We now show that plea bargaining with two verdicts when the guilty sentence is excessively high is worse than a three-verdict system as in Section 2.1 that keeps the excessively high sentence for the verdict associated with strong evidence.

Because of the linear structure of payoffs, it is easy to show that the only relevant sentence levels are $s_1 = 16$ and $s_2 = 20$. The following facts are easy to establish in this example:

[52] This fact is well-known and straightforward to establish.: if θ is the state of the world, r is a signal, and the conditional distributions $\rho(r|\theta)$ are ranked according to MLRP, then the posterior distributions $\rho(\theta|r)$ are also ranked according to the MLRP.

- In a two-verdict system without a plea, it is optimal to punish the defendant for either type of evidence (weak or strong), and the optimal sentence is $s_1 = 16$;
- The same is true in an optimal two-verdict system with a plea, and only the guilty defendant takes the plea;
- If, however, the conviction sentence is suboptimally set to $s_2 = 20$ at the trial stage (which is the ex post optimum if the defendant is indeed guilty), then the optimal plea is $s^b = 0.8 * s_2 = 16$, and both guilty and the risk averse innocent defendants take the plea.
- Subject to keeping a high sentence equal to $s_2 = 20$, the three-verdict system that gives a sentence of $s_1 = 16$ if weak evidence is presented, and $s_2 = 20$ if strong evidence is presented is optimal and yields a higher expected welfare than the two-verdict system with a plea that has a trial conviction sentence of $s_2 = 20$.

This result shows that the introduction of an intermediate verdict with a lower sentence may be more efficient than a plea to counteract the effects of a suboptimally high sentence for the guilty. This illustrates how ethical considerations (here, providing the right ex post punishment if the defendant is guilty) shape the optimal verdict system: in a purely utilitarian world, a suboptimally high guilty sentence would be reduced (here, to 16) and plea bargains may be optimal. If, however, it is difficult to reduce the guilty sentence, due to political or other considerations, plea bargaining not be the best solution.

Another reason plea bargains may be suboptimal is that an innocent defendant may think that his likelihood of being convicted is higher than it really is. Revisiting the example, suppose that the risk averse innocent defendant erroneously believes that the probability of weak evidence being found against him is 75%. Then he may prefer to take the plea rather than run the risk of being found guilty in trial. In this case, even if the guilty sentence is set to $s = 16$, welfare is suboptimal compared to a three verdict system.

D Modeling continuous evidence gathering (Online)

As long as evidence is gathered, the belief p_t that the defendant is guilty evolves as a martingale as in Bolton and Harris (1999):

$$dp_t = Dp_t(1 - p_t)dB_t,$$

where B is the standard Brownian motion and D is a measure of the quality of the signal: the higher D is, the faster p evolves toward the true probability that the defendant is guilty (0 or 1). At some time T, the evidence formation process is stopped and the verdict is chosen based on the posterior p_T, which results in social welfare $w(p_T)$.

Adapting the arguments of Bolton and Harris (1999) to our environment, the value function $v(\cdot)$ must satisfy the Bellman equation

$$0 = \max\{w(p) - v(p); -rv(p) - c + \frac{1}{2}D^2p^2(1-p)^2v''(p)\}, \tag{26}$$

where r is a discount rate that captures the idea that longer judicial processes are penalizing for all parties. The first part of the equation implies that $v(p) \geq w(p)$, which means that the value function always exceeds the welfare obtained by stopping immediately. This is natural, since the option of stopping is available at any time. The second part of the equation describes the evolution of the value function while evidence is accumulated:

$$0 = -rv(p) - c + \frac{1}{2}D^2p^2(1-p)^2v''(p).$$

All solutions to this equation are in closed form when $D^2/r = 3/2$:

$$v(p) = -\frac{c}{r} + \left(A_1 + A_2 \left(p - \frac{1}{2} \right) (1-p)^{-2} \right) p^{-\frac{1}{2}} (1-p)^{\frac{3}{2}}, \tag{27}$$

where A_1 and A_2 are free integration constants. For simplicity, in what follows we set $r = 1$ and $D^2 = 3/2$ and vary the cost c.

The region in which evidence is gathered and value functions are determined by the conditions that v is continuous, weakly above w, and when it hits w, it satisfies the smooth pasting property whenever w is continuously differentiable at the hitting point.

Starting with the two-verdict case, one should expect v to coincide with w when p is either close to 0 or close to 1: in this case, there is a high degree of confidence in the defendant's guilt and the value of further evidence gathering is low. Near w's kink (i.e., the threshold p^* at which the sentence switches), however, the value of additional evidence is high, so v should be strictly above w. Thus, it suffices to connect v and w on both sides of p^*. At the connection points, $\hat{p}_1$ and $\hat{p}_2$ such that $\hat{p}_1 < p^* < \hat{p}_2$, v must be equal to w (this is the "value matching" condition) and the derivatives must also coincide (this is the smooth pasting condition).

This imposes four conditions (two value matching and two smooth pasting), and there also four free parameters: the cutoffs $\hat{p}_1$ and $\hat{p}_2$, and the constants A_1 and A_2 arising in equation (27). The result is depicted in Figure 3.

Now consider the three-verdict case. Around the kink p_2, we still have a two-way smooth connection between w and v, as in the two-verdict case. Around $p_1 = p^*$, however, w is discontinuous, jumping upward from $\underline{w} = -1/3$ to $\bar{w} = -2/9$ as p passes p_1. In this case, if $v(p_1) > \bar{w}$ (the cost is low), then the situation is exactly as in the two-verdict case. Intuitively, the cost is low enough that the intermediate verdict doesn't matter: evidence is gathered until either the not guilty or the guilty verdict is reached. This a situation in which the trial technology is quite accurate, so a two-verdict system suffices.

For larger costs, however, v hits w exactly at $p_1 = p^*$, due to the upward jump. The smooth pasting condition is violated, because the left derivative of v is higher than its right derivative at p_1, and v is equal to w on a right neighborhood of p_1. Intuitively, this kink in the value function reflects the fact that $p_1 = p^*$ was not chosen optimally for the three-verdict system, but rather inherited from the two-verdict system.

The evidence-gathering region now has two parts. When p is below p_1, there is a large incentive to gather evidence, because such evidence can change the sentence from 0 to s_1, and s_1 was tailored to provide a fairer sentence around p_1 than both 0 and s_2. This also implies that not gathering evidence in a right-neighborhood of p_1 is optimal. The second evidence-gathering region is around p_2, as before.[53]

Because the first region violates the smooth pasting condition at p_1, its determination is slightly different. We must determine the threshold $\tilde{p}_0$ at which the region begins, and we know that the region ends at the cutoff p_1. At $\tilde{p}_0$, we have two conditions: the value matching and the smooth pasting conditions. At p_1, however, we only have the value matching condition $v(p_1) = \bar{w}$, since the smooth pasting condition is violated. This gives three conditions. There are also three free parameters: the cutoff $\tilde{p}_0$ and the constants $\hat{A}_1$ and $\hat{A}_2$ in (27) for that region. The result is depicted in Figure 4.

Because the welfare w_3 is always higher than the welfare w_2, it is is straightforward to establish that the value function v_3 in the three-verdict case is (weakly) higher than the two-verdict value function v_2. This matters for high enough cost, i.e., when $v(p_1) = \bar{w}$. In that case, v_3 is strictly above v_2 around p_1, and it is also strictly above v_2 in the second evidence-gathering region, closer to p_2. This implies that the cutoff $\tilde{p}_0$ is lower than the cutoff $\hat{p}_1$ of the two-verdict case, and the right cutoff $\tilde{p}_2$ of the second evidence-gathering region in the three-verdict case is greater than $\hat{p}_2$.

[53] As the search cost decreases, the two search regions become connected when $v(p_1) \geq \bar{w}$.

References

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

Austen-Smith, D., Banks, J. (1996) "Information Aggregation, Rationality, and the Condorcet Jury Theorem," *American Political Science Review,* Vol. 90, pp. 34–45.

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

Bolton, P., Harris, C. (1999) "Strategic Experimentation," *Econometrica*, Vol. 67, pp. 349–374.

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

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

Burns, R. (2009) *The Death of the American Trial*, University of Chicago Press.

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.

Dressler, J., Thomas, G. (2010) "Does the Process Protect the Innocent," in *Criminal Procedure: Prosecuting Crime*, Fourth Edition, West Academic Publishing.

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.

Feddersen, T., Pesendorfer, W. (1996) "The Swing Voter's Curse," *American Economic Review,* Vol. 86, pp. 408–424.

Feddersen, T., Pesendorfer, W. (1997) "Voting Behavior and Information Aggregation in Elections with Private Information," *Econometrica,* Vol. 65, pp. 1029–1058.

Gerardi, D., Yariv, L. (2007) "Deliberative Voting," *Journal of Economic Theory*, Vol. 134, pp. 317–338.

Grogger, J. (1992) "Arrests, Persistent Youth Joblessness, and Black-White Employment Differentials," *Review of Economics and Statistics*, Vol. 74, pp. 100–106.

Grogger, J. (1995) "The Effect of Arrest on the Employment and Earnings of Young Men," *Quarterly Journal of Economics*, Vol. 90, pp. 51–72.

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

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

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.

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

Lott, J. (1990) "The Effect of Conviction on the Legitimate Income of Criminals," *Economics Letters*, Vol. 34, pp. 381–385.

Mascolo, E. (1985) "Procedural Due Process and the Lesser-Included Offense Doctrine," *Albany Law Review*, Vol. 50, pp. 263–304.

Milgrom, P., Segal, I. (2002) "Envelope Theorems for Arbitrary Choice Sets," *Econometrica*, Vol. 70, pp. 583–601.

Milgrom, P., Shannon, C. (1994) "Monotone Comparative Statics," *Econometrica*, Vol. 62, pp. 157–180.

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

Quah, J., and Strulovici, B. (2012) "Discounting, Values, and Decisions," *Journal of Political Economy*, Vol. 121, pp. 898–939.

Sand, L., Rose, D. (2003) "Proof Beyond All Possible Doubt: Is there a Need for Higher Burden of Proof When the Sentence May Be Death," *Chicago-Kent Law Review*, Vol. 78, pp. 1359–1376.

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

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

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