Siegel, Ron; Strulovici, Bruno

Working Paper

On the design of criminal trials: The benefit of three-verdict systems

CSIO Working Paper, No. 0132

Provided in Cooperation with:
Department of Economics - Center for the Study of Industrial Organization (CSIO), Northwestern University

Suggested Citation: Siegel, Ron; Strulovici, Bruno (2015) : On the design of criminal trials: The benefit of three-verdict systems, CSIO Working Paper, No. 0132, Northwestern University, Center for the Study of Industrial Organization (CSIO), Evanston, IL

This Version is available at:
https://hdl.handle.net/10419/142006

On the Design of Criminal Trials: the Benefit of Three-Verdict Systems

Ron Siegel and Bruno Strulovici*

August 2015

Abstract

We propose adding a third, intermediate, verdict to the two-verdict system used in criminal trials. We show that the additional verdict can be used to distinguish between convicted defendants, based on the residual doubt regarding their guilt at the end of the trial, in a way that improves welfare and does not increase the set of innocent defendants who are wrongly convicted. It can also be guaranteed that wrongfully convicted defendants do not serve longer sentences, provided that the sentence in the two-verdict system was not too inefficiently low.

Since even acquitted defendants may face a social stigma, we also consider using the additional verdict to distinguish between acquitted defendants, and provide conditions under which this improves welfare. Generalizations to multi-verdict systems with a larger number of verdicts are also explored.

We also consider plea bargains, and show that a properly chosen plea in a two-verdict system leads to higher welfare than any multi-verdict system, and is in fact the optimal mechanism.

Finally, we consider the impact of multiple verdicts on the incentives to gather evidence, and show that the effect is generally positive.

*We thank Robert Burns, Eddie Dekel, Fuhito Kojima, Adi Leibovitz, Paul Milgrom, and Jean Tirole and seminar participants at UC Berkeley, Seoul National University, and the NBER Summer Institute Law and Economics Workshop for helpful comments. David Rodina provided outstanding 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

A key component of criminal trials is the standard for determining guilt (liability). In the United States, the prevailing standard is "reasonable doubt," which reflects the view that it is more important not to punish the innocent than it is to mistakenly acquit the guilty. But even this demanding standard does not eliminate wrongful convictions of innocent defendants.[1] Indeed, as a practical matter, every justice system accepts some wrongful convictions in order to punish the guilty.

This indicates that criminal trials do not fully eliminate the uncertainty regarding the defendant's guilt. But this uncertainty can only be expressed in a limited way in the current system, in which the defendant is found either guilty or not guilty.[2] A more expressive system would be able to better reflect the uncertainty remaining at the end of the trial, and this can potentially increase social welfare.

We propose introducing a third, intermediate, verdict as a possible outcome in criminal trials. The possibility of an additional verdict has been proposed in the legal literature by Bray (2005), but has not been analyzed formally.[3] The intermediate verdict will be used when the doubt regarding the defendant's guilt is close to "reasonable," so the expected social cost because of a wrong decision in the two-verdict system is highest. Because this three-verdict system better reflects the uncertainty that remains regarding the defendant's guilt at the end of the trial, together with an appropriately chosen sentence it can improve upon the two-verdict system.

A substantial concern is that the introduction of a third verdict may lead to more innocent defendants being punished. To address this concern, we conduct much of our analysis under the restriction that an innocent defendant is only punished in the three-verdict system if he would be punished in the two-verdict system. We consider two ways in which the intermediate verdict can be incorporated. First, the intermediate verdict may be used to distinguish among defendants who would be convicted in the two-verdict system. Among those defendants, the ones for whom more doubt remains will be punished less severely than those whose guilt is

[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.

[2]Civil suits are less restrictive, since penalties are more continuous.

[3]Bray's proposal concerns the addition of a "not proven" verdict to the U.S. criminal system. Daughety and Reinganum (2015a) consider the effect of informal sanctions on defendants and prosecutors. In an extension discussed later in this paper, they consider the effect of introducing a not-proven verdict. Daughety and Reinganum (2015b) consider several implementations of the "not proven" verdict through defendant choice and compensation.

more certain. We show that for any punishment in the two-verdict system and any doubt threshold exceeding the one in the two-verdict system, there is a way to set the punishments above and below the threshold that increases welfare relative to the two-verdict system. If the punishment in the two-verdict system is not too inefficiently low, we obtain the stronger result that welfare can be improved without increasing the punishment. This guarantees not only that no additional innocent defendants are punished in the three-verdict system, but also that those who are punished are not punished more severely than in the two-verdict system. We generalize this result and show that it holds for any multi-verdict system. That is, for any multi-verdict system one can add another verdict and lower the punishments in a way that increases social welfare.

The additional verdict can be introduced into criminal trials in the United States 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. 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. If the jury find the defendant guilty, then the judge would determine the sentencing category based on remaining guilt uncertainty. 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 guilt certainty 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. Determining the best method involves many considerations that are beyond the scope of this paper.

The additional verdict can also be used 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.[4] Because this stigma is likely related to the perceived likelihood that they are in fact guilty, distinguishing among these defendants based on the residual guilt uncertainty 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

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

in the same way that sentences are. Consequently, this introduction of a third verdict does not always increase welfare, 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 this third verdict increases welfare, as well as comparative statics.

Several countries, including Israel, Italy, and Scotland, do in fact distinguish among acquitted defendants. 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. The likelihood is, however, insufficiently high for conviction.[5]

We also consider how to optimally incorporate a third verdict without the restriction on the probability that innocent defendants are punished. We show that an optimal three-verdict system will generally punish defendants more frequently than the two-verdict system, since the intermediate verdict will carry a positive sentence, but the additional defendants who are punished, as well as some defendants who would be punished in the two-verdict system, optimally receive a lower punishment than convicted defendants in the two-verdict system. However, those defendants who are punished in the two-verdict system and regarding whose guilt little uncertainty remains at the end of the trial are optimally punished more severely in the three-verdict system.

We then turn to investigating plea bargains, which are an important instrument in the United States criminal justice system, and of growing importance in many other countries.[6] In a plea bargain, the defendant does not go to trial and instead pleads guilty and accepts a lower sentence than the one he would likely get if convicted. Pleas may therefore be seen as a third verdict, which is proposed to the defendant before the trial. Because the defendant chooses whether to accept the plea, and guilty defendants are (presumably) more likely to be found guilty during a trial, the plea can serve as a screening device. Building on the framework of Grossman and Katz (1983), who show that guilty defendants are more willing to take the plea, we analyze the value of plea bargains relative to other verdict systems. We show that an appropriate two-verdict system with pleas 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

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

[6]In the United State, more than 90% of criminal cases are settled by plea bargains (Burns (2009)).

guilt. In fact, we show that there is a two-verdict system with a plea that maximizes welfare among all incentive compatible mechanisms. In this optimal mechanism, the guilty sentence coincides with the sentence that is optimal when one is certain that the defendant is guilty.

Despite its generality, the result on the superiority of two-verdict systems with plea bargains omits at least two issues. First, the sentence faced by defendants if they go to trial may not have been set optimally. For example, the sentence corresponding to a guilty verdict is often chosen to be the sentence that would be optimal *if* one were certain the defendant is guilty. But since some innocent defendants are also convicted, that maximal sentence may be too harsh, especially since a long sentence may lead innocent defendants who are concerned about the risk of being convicted to accept the plea bargain. To demonstrate this, we show by example that when the guilty sentence is set at a suboptimally high level, the two-verdict system with a plea may be dominated by a three-verdict system of the form described above, which distinguishes among convicted defendants according to the remaining uncertainty regarding their guilt. Second, one may 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.

Setting pleas aside, we conclude the paper by investigating how the introduction of a third verdict affects the value of evidence in a trial. Since gathering evidence is costly, the socially optimal amount of evidence to be gathered depends on the verdict structure. Focusing on a three-verdict system that distinguishes among convicted defendants according to the remaining uncertainty regarding their guilt, we show that the introduction of the third verdict generally increases the value of evidence and therefore the optimal amount of evidence that should be gathered.

Daughety and Reinganum (2015a) consider the effect of informal sanctions on defendants and prosecutors. The effect on the former depend on the public's belief regarding the defendant's guilt, and the effect on the latter depend on public's belief that the verdict was mistaken. They show how the informal sanctions interact and affect the plea bargain and its acceptance rate, and also consider the effect of introducing a not-proven verdict, i.e., splitting the innocent verdict (but not splitting the guilty verdict). In their setting this is always welfare improving in that it makes guilty defendants worse off, innocent defendants better off, and leads to lower expected loss due to misclassification by the observers. Daughety and Reinganum (2015b) consider two implementations of the 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 verdict. The authors also analyze an alternative implementation in which some defendants who are found not guilty are compensated.

Appendix A provides a micro foundation for the Bayesian formulation we use throughout 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 to generate incriminating evidence than an innocent defendant.

2 Baseline model: two verdicts, no pleas

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.[7] Corresponding to a guilty verdict is a sentence $s > 0$, interpreted as jail time (so a higher value of s corresponds to a higher punishment).[8]

Society's goal is to avoid punishing innocent defendants and adequately punish guilty ones. This dual goal is modeled by a 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 relative importance of these objectives depends on the prior probability λ that the defendant is in fact guilty. The more likely the defendant is ex-ante to be guilty, the more

[7] 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 restriction is, however, not required for this section.

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

important it is to adequately punish the defendant if he is in fact guilty; the less likely the defendant is ex-ante to be guilty, the more important it is to avoid punishing the defendant if he is in fact innocent. This is captured by the ex-ante social welfare function

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

Since $W(\cdot, i)$ is decreasing and $W(\cdot, g)$ peaks at $\bar{s}$, it is never optimal to choose $s > \bar{s}$. We therefore restrict attention to sentences s in $[0, \bar{s}]$.

Assumption: Any sentence s satisfies $s \in [0, \bar{s}]$.

3 Adding an intermediate verdict

3.1 Intermediate "guilty" verdict

We introduce a third verdict in such a way that 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. 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.[9] Denote by π_i^1 the probability that the defendant receives verdict 1 if he is innocent, and define π_i^2, π_g^1, and π_g^2 similarly. Because the probability of not acquitting the defendant does not change, 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

$$\frac{\pi_g^1}{\pi_i^1} < \frac{\pi_g}{\pi_i} < \frac{\pi_g^2}{\pi_i^2},$$[10]

so verdict 1 is an "intermediate verdict:" an innocent defendant is more likely to receive verdict

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

[10] It is straightforward to check that for any a, b, c, d of $\mathbb{R}_{++}$, $\min\{a/b, c/d\} \leq (a+c)/(b+d) \leq \max\{a/b, c/d\}$, with strict inequalities generically. The inequalities will be strict if, for example, the verdict is decided according to the posterior probability that the agent is guilty.

1, relative to a guilty defendant, than verdict 2.

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

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

Our first result shows that the expected welfare can be improved provided that the sentence s associated with a conviction in the two-verdict system is interior, i.e., $s < \bar{s}$.

Proposition 1 *For any interior sentence s of the two-verdict system and verdict technologies π_i, π_g, π_i^j, etc., there are sentences s_1 and s_2 such that $s_1 < s < s_2$ and $\mathcal{W}_3(s_1,s_2) > \mathcal{W}_2(s)$.*

The key aspect of Proposition 1 is that it does not increase the probability of punishing the innocent, compared 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.[11]

Proof. First, observe that $\mathcal{W}_3(s,s) = \mathcal{W}_2(s)$: if we give the same sentence s for both verdicts 1 and 2, equal to the sentence for the guilty verdict of the 2-verdict case, then we are back to the two-verdict case and achieve the same welfare. 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

$$\mathcal{W}_3(s_1,s_2) = \mathcal{W}_2(s) + \lambda(W_g'^1 + \gamma\pi_g^2) + (1-\lambda)(W_i'^1 + \gamma\pi_i^2) + o(\varepsilon), \tag{3}$$

where W' denotes the derivative of W with respect to its first argument. Since $W(\cdot,i)$ is decreasing, $W'(v,i)$ is negative. Similarly, because $s \leq \bar{s}$ and $W(\cdot,g)$ is increasing on that domain, we have $W'(s,g) > 0$. Since also $\pi_g^1/\pi_g^2 < \pi_i^1/\pi_i^2$, we can choose γ between these two ratios. Doing so guarantees that $W_g'^1 + \gamma\pi_g^2)$ and $W_i'^1 + \gamma\pi_i^2)$ are both positive, which shows the claim. ■

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 worse sentence ex-post, because $s_2 > s$. The next next result shows that if the sentence associated

[11] While our model abstracts from the incentives to commit crimes, our design can easily accommodate an increase in s_2 to maintain deterrence.

with a conviction in the two-verdict system was interior and optimal to begin with, then there is improvement that does not increases the sentence.

Proposition 2 *Suppose that s^* maximizes $\mathcal{W}_2(s)$ and is interior. Then, there exists $s_1 < s$ such that $\mathcal{W}_3(s_1, s^*) > \mathcal{W}_2(s^*)$.*

The proof of Proposition 2 shows that the result holds even when the original sentence was not optimal, as long as it was not too suboptimally low. 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.

Proof. By construction s^* maximizes

$$\lambda \left[\pi_g W(s, g) + (1 - \pi_g) W(0, g)\right] + (1 - \lambda) \left[\pi_i W(s, i) + (1 - \pi_i) W(0, i)\right]$$

with respect to s. Since s^* is interior, it must satisfy the first-order condition

$$\lambda \pi_g W'(s^*, g) + (1 - \lambda) \pi_i W'(s^*, i) = 0. \tag{4}$$

Now consider the derivative of $\mathcal{W}_3(s_1, s^*)$ with respect to s_1, evaluated at $s_1 = s^*$. From (3), we have

$$\left.\frac{\partial \mathcal{W}_3(s_1, s^*)}{\partial s_1}\right|_{s_1 = s^*} = \lambda \pi_g^1 W'(s^*, g) + (1 - \lambda) \pi_i^1 W'(s^*, i). \tag{5}$$

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 (4) implies that the right-hand side of (5) is strictly negative. This shows that decreasing s_1 below s^* strictly improves welfare, yielding the desired improvement. ■

3.2 The Bayesian conviction model

The previous section did not impose any structure on how verdicts are determined. Since more structure is required for the analysis in the remainder of the paper, this section specializes the setting to a class of verdicts based on the posterior probability that the defendant is guilty. Starting with a prior probability λ, 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.[12] Appendix A shows that any "reasonable" verdict rule based on evidence can be formalized as a Bayesian model in this way. For expositional convenience only, we assume that $F(\cdot|g)$ and $F(\cdot|i)$ are differentiable with respect to their first argument and 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. If the posterior p is below a threshold p^*, then the defendant is acquitted, receiving a sentence of $s = 0$. If instead $p > p^*$, 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 ex-ante social welfare is given by

$$\begin{aligned}\mathcal{W}_2(p^*, s^*) = \lambda \left[(1 - F(p^*|g))W(s^*, g) + F(p^*|g))W(s^*, g)\right] + \\ (1 - \lambda) \left[(1 - F(p^*|i))W(s^*, i) + F(p^*|i)W(s^*, i)\right].\end{aligned} \tag{6}$$

In what follows, we will denote by (p^*, s^*) the cutoff and sentence used in the two-verdict system. These variables may be chosen so as to maximize (1). In that case, they correspond to the utilitarian optimum for the 2-verdict case.

Notice that this optimum is constrained by the restriction that an acquittal leads to the sentence $s = 0$. Within the model, this restriction may be questioned: for example, if $p^* = 0.9$, i.e., the defendant is convicted only if there is a 90 percent chance that he is guilty, then an acquittal is not a strong indication of innocence. For example, if $p = 0.8$, then the defendant is acquitted due to insufficient evidence, even though the probability he is guilty is quite high.

In practice, a defendant may face significant stigma even if he is acquitted. Indeed, if guilt must be established "beyond a reasonable doubt" for a conviction, then an acquittal is consistent with significant doubt regarding the defendant's innocence. Such doubt will harm a defendant who is in fact innocent and has been acquitted. To address this issue the next section introduces an additional acquittal verdict, and a defendant who is not convicted receives an acquittal verdict based on the amount of residual doubt regarding his guilty.

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

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

3.3 Intermediate "not guilty" verdict

We introduce a third verdict in such a way that those defendants who would be acquitted in the 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 $p^{iv} < p^*$, 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 posterior above p^* leads to a conviction and 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 guilty is more demanding than stigmatizing him, so $p^s < p^*$.[13] 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 $d^g > 0$.[14] We are interested in the optimal cutoff p^{iv} and the conditions under which introducing the additional verdict increases 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

[13] This implies that the analysis of Section 3.1 does not change as a result of the stigma, since a defendant who receives verdicts 1 or 2 is stigmatized.

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

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

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

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

$$-\lambda f\left(p|g\right) d^g + \left(1-\lambda\right) f\left(p|i\right) d^i > 0 \iff \frac{f\left(p|g\right)}{f\left(p|i\right)} < \frac{\left(1-\lambda\right) d^i}{\lambda d^g}. \tag{7}$$

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

Proposition 3 *Suppose that being acquitted in the two-verdict system carries a stigma. Then, optimally splitting the acquittal into "not guilty" and "not proven" increases welfare if and only if*

$$\frac{f\left(\underline{q}|g\right)}{f\left(\underline{q}|i\right)} < \frac{\left(1-\lambda\right) d^i}{\lambda d^g}.$$

If the condition in Proposition 3 holds, then the optimal cutoff p^{iv} is the minimum between the highest posterior for which (7) holds and the highest posterior such that $p_1 \leq p^s$. Notice that the condition in Proposition 3 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

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

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

$$\lambda f(p|g)\, d^g - (1-\lambda) f(p|i)\, d^i > 0 \iff \frac{f(p|g)}{f(p|i)} > \frac{(1-\lambda)\, d^i}{\lambda d^g}. \tag{8}$$

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

Proposition 4 *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 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 4 holds, then the optimal p^{iv} is the maximum between the lowest posterior for which (8) holds and the lowest posterior such that $p_2 \geq p^s$. Notice that the condition in Proposition 4 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.

3.4 Welfare maximization with three verdicts

Although normatively appealing, the cutoff and sentence restrictions reduce welfare, and it is natural to ask what the optimal three-verdict system looks like. The result is provided by the following proposition.

Suppose that (p^*, s^*) are optimal in the two-verdict system, and let $(p_1^*, p_2^*, s_1^*, s_2^*)$ be optimal in the three-verdict system (if the posterior is below p_1^*, then the sentence is 0, if the posterior is between p_1^* and p_2^*, then sentence is s_1^*, etc.).

Assumption: $W(\cdot, i)$ and $W(\cdot, g)$ are concave in s, the posterior distributions $F(\cdot|i)$ and $F(\cdot|g)$ are both continuous, and W and F are regular, so the implicit function theorem applies.

Proposition 5 $p_1^* \leq p^* \leq p_2^*$ *and* $s_1^* \leq s^* \leq s_2^*$.

Intuitively, the optimal sentence reflects the likelihood that the agent is guilty. Thus, 'higher' sets of priors will lead to a longer sentence. The proof of this proposition is in Appendix B.

4 Multi-verdict systems

Before discussing plea bargains in the next section, it is useful to generalize our analysis to any number of verdicts. Consider first the optimal sentence s as a function of the posterior p. Given a posterior p, the optimal sentence $s(p)$ maximizes the welfare objective

$$pW(s,g) + (1-p)W(s,i) \tag{9}$$

with respect to s. Since both $W(\cdot, g)$ and $W(\cdot, i)$ are decreasing beyond the ideal punishment $\bar{s}$ for a guilty defendant, any optimizer of (9) is lower than $\bar{s}$. Moreover, 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) (see Topkis (1978)), because $W(\cdot, g)$ increases in the relevant range $[0, \bar{s}]$ and $W(\cdot, i)$ is decreasing, which implies that $\partial\mathcal{W}/\partial p = W(s,g) - W(s,i)$ increases in s. This implies that the selection of maximizers of (9) is isotone. In particular, there exists a nondecreasing selection $s(p)$ of optimal sentences.

The arguments used for Propositions 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 gets sentence $s_{k'}$ whenever his posterior p lies in $(p_{k'}, p_{k'+1})$.

Proposition 6 *Suppose that the signal 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 welfare. Moreover, if a k-verdict system is optimal among all k-verdict systems and either $k > 2$ or $k = 2$ and $s_1 < \bar{s}$, then there is a $k+1$-verdict system that strictly improves upon it and has lower sentences.*

5 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 third verdict, which corresponds to an intermediate sentence that is lower than the one associated with a trial conviction. This third verdict is different from what has been discussed so far, because it involves a strategic decision by the

defendant of whether to take the plea, in contrast to his passive role in 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 s^b. If the defendant accepts the plea, he gets this sentence and the case is concluded. If the defendant rejects the plea, he goes to trial and faces the same signal structure as in the previous sections. The welfare functions $W(\cdot, i)$ and $W(\cdot, g)$ are also as in the previous sections. Following GK, we assume that the information revealed by the choice of the defendant to reject the plea is not taken into account during the trial. This is consistent with legal requirements to focus only on the evidence presented during trial when assessing the defendant's guilt. But this assumption may seem troubling given the separating equilibrium described below, in which only an innocent defendant goes to trial. It turns out, however, that the assumption is irrelevant, since the same outcomes can be achieved with and without incorporating the information resulting from the defendant's decision into the trial. This is shown in Appendix C.

A two-verdict system with pleas is thus characterized by four parameters: the plea s^b, the guilty sentence s, and the probabilities π_g and π_i that the defendant is found guilty during the trial, depending on whether he is actually guilty or innocent. We assume that the defendant's utility function, u, does not depend on whether he is actually guilty. But because a guilty defendant is more likely to be found guilty if he goes to trial ($\pi_g > \pi_i$), his incentive to go to trial is strictly weaker than an innocent defendant's.

Therefore, depending on the parameters, three types of equilibrium behavior can arise. Either the defendant takes the plea regardless of his guilt, or he rejects the plea and goes to trial regardless of his guilt, or only the guilty defendant takes the plea.[15] GK show the optimal system with a plea bargain is separating. The plea s^g is chosen so 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.

In the next subsection we show that this equilibrium outperforms *any* multi-verdict system without pleas, including ones with two and three verdicts, and is in fact optimal within a much broader class of mechanisms.

[15]Mixing can be showed to be suboptimal.

5.1 The welfare value of plea bargaining

As discussed in previous sections, we assume that the posterior distribution for the defendant is continuous.

Proposition 7 *Consider any multi-verdict system $s : p \to s(p)$ that is nondecreasing and taking values in $[0, \bar{s}]$. There exist a two-verdict system with a plea that increases welfare.*

Proof. We begin by constructing a two-verdict system $\hat{s}$ that give the guilty defendant the same expected utility as $\mathbf{s}$. In this system, there is a cutoff $\hat{p}$ below which the sentence is zero and above which the sentence is $s(1) = \max_p s(p)$. Moreover, the cutoff is chosen so that

$$U^g \int_0^1 u(s(p)) f_g(p) dp = \int_0^1 u(\hat{s}(p)) f_g(p) dp = u(0) F_g([0, \hat{p}]) + u(s(1)) F_g([\hat{p}, 1]) = \hat{U}^g, \qquad (10)$$

recalling that $u(s)$ denotes the defendant's utility from getting sentence s, and u is decreasing and concave. Because the right-hand side of (10) is continuous in the cutoff p, ranging all values from $u(0)$ to $u(s(1))$, and because U^g clearly lies between $u(0)$ and $u(s(1))$ as a convex combination of utilities that lie in this interval, the existence of $\hat{p}$ is clear. Moreover, the new verdict system increases the expected utility of an innocent defendant. To show this claim, notice that by construction we have

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

for all $\tilde{p} \in [0, 1]$. Since $f_i(p)/f_g(p)$ is positive and decreasing in p, this implies that[16]

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

or

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

[16]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, as a special case of the previous argument. This is done by observing that the integrand has the single crossing property in p and that the type of the agent is affiliated with the posterior. Single crossing of the expected utility follows (see, e.g., Athey, 2002).

We now introduce the plea s^b, setting it so as to make the guilty defendant indifferent between taking the plea and going to trial: that is, we choose s^b so that

$$u(s^b) = U^g = \hat{U}^g.$$

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).

Since the innocent benefits from the new verdict system, we will have shown that it improves of the old if we prove that the social welfare conditional on facing the guilty defendant is also higher. This welfare is equal to $W(s^b, g)$. Notice that s^b is the certainty equivalent sentence for the guilty which makes him indifferent with going to trial. Because the defendant is risk averse (u is concave), s^b is greater than the average sentence $\tilde{s} = \int_0^1 s(p) f_g(p) dp$ that the guilty gets if he goes to trial. Moreover, because $W(\cdot, g)$ is also concave, we have $W(\tilde{s}, g) \geq \int_0^1 W(s(p), g) f_g(p) dp$. Finally, since $s^b \geq \tilde{s}$ and $W(\cdot, g)$ is increasing, we conclude that $W(s^b, g)$ dominates the expected social welfare conditional on facing the guilty.

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

By modifying the proof slightly, it is possible to prove that the following, stronger result. All the verdict systems, with and without pleas, may be seen as particular mechanisms. It is well-known from the mechanism design literature that in the present setting there exists a direct revelation mechanism: the defendant makes a reports $\hat{\theta}$ of his type (guilty or innocent) and is then assigned a sentence $s(p, \hat{\theta})$ that depends on his report and on the posterior (signal) p generated during trial, which is assumed to be continuous on $[\underline{p}, \overline{p}]$ and satisfy the monotone-likelihood ratio property. A mechanism is feasible if $s(p, \hat{\theta}) \leq \bar{s}$ for all p and θ, i.e., it does not punish the defendant more than would be optimal if the defendant were known to be guilty. A feasible mechanism is optimal if it maximizes welfare given the prior probability λ that the defendant is guilty.

Proposition 8 *There is a unique optimal mechanism. This mechanism takes the form of a two-verdict system with a plea: $s(\cdot, g)$ is constant (i.e., like a plea), and $s(\cdot, i)$ is a two-step*

function, which jumps from 0 *to* $\bar{s}$*. The incentive compatibility constraint of the guilty defendant is binding. The cutoff at which* $s(\cdot, i)$ *jumps from* 0 *to* $\bar{s}$ *decreases in the prior from* $\overline{p}$ *to* $\underline{p}$*.*

5.2 The failure of plea bargaining with excessive sentencing

Despite the result of the previous section, pleas 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.[17] We now provide an example that illustrates this idea.

The first step is to 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 3.1 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:

- 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 $s_1 = 16$;

[17] See for example Rakoff (2014) and Kagan's opinion in Supreme Court Ruling No. 13-7451 on Yates vs. U.S.

- 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.

6 Incentives for Evidence Formation

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

Leaving aside the possible biases in these agents' behavior, the socially optimal amount of information to be acquired 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. Such criticism has been leveled at plea bargaining: that so many defendants take a plea reduces incentives for information acquisition.

This section compares the impact on evidence formation of introducing a third verdict. For simplicity, we focus on the setting of Section 3.1 with the Bayesian conviction model.

A (possibly multi-) verdict system leads to welfare

$$w(p) = pW(s(p), g) + (1 - p)W(s(p), i), \tag{11}$$

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.

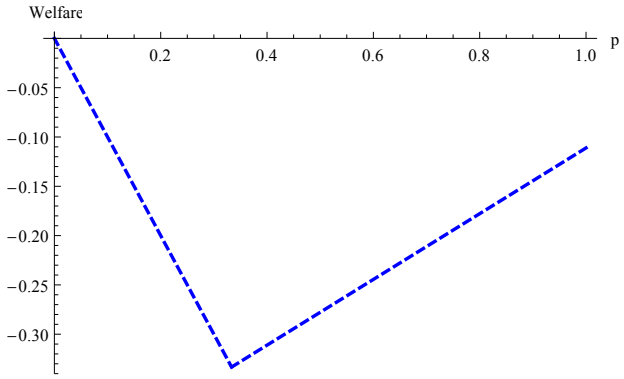

Figure 1: Welfare function, 2 verdicts.

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 the same. 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.

Actual evidence formation processes are complex, involving numerous actors of different types – forensic experts, lawyers, witnesses – and various forms of evidence. To model this information acquisition task, we must abstract from much of this complexity. Instead, we take the viewpoint

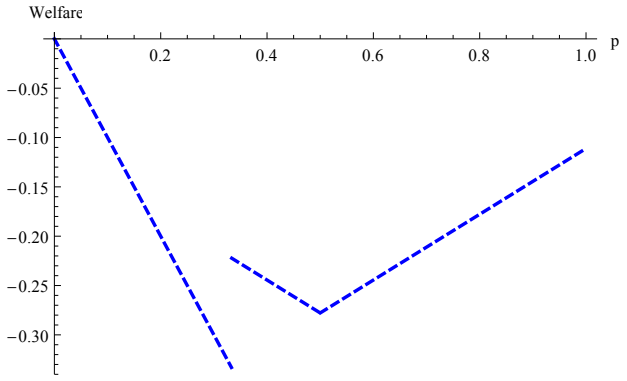

Figure 2: Welfare function, 3 verdicts.

of a social planner who may gather information until a verdict is reached.

The tradeoff at the heart of this task is clear: more effort spent gathering evidence means higher costs for society but more precise information about the defendant's guilt. We discuss two ways to model this tradeoff (there are, of course, many others). This first is a one-shot 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.

6.1 One-shot evidence gathering

Suppose the planner decides whether to gather evidence, which has a cost $c > 0$. 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 $1/2(p + \Delta) + 1/2(p - \Delta) = p$, i.e., the prior.

When is evidence gathering socially desirable? Suppose first that the prior is close to 0, so the posterior p' surely lies on the first branch of the graph in Figure XXX. Then, the value of evidence is zero, due to the linearity of w_2, and further evidence will not be gathered. Similarly, if p_0 is high enough for p' to surely lie on the second branch of w_2, the value of evidence is zero. Intuitively, the evidence is not enough to change the verdict and hence is valueless.

Consider now the case of three verdicts. For p slightly above a $p_1 + \Delta$, evidence is valueless as well for Δ small enough, because p' will lie between p_1 and p_2 regardless of the verdict. Thus, in this region, moving to a third verdict *reduces* the incentive to gather evidence.

For p slightly less than p_1, however, the value of evidence is high, because a positive belief

update triggers a large improvement in welfare. Similarly, for p in a neighborhood of p_2, the value of evidence is positive, whereas it is 0 (for Δ small enough) in the two-verdict case.

6.2 Continuous evidence gathering

Now suppose that evidence is gathered for 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 as a martingale according to a continuous signal, modeled 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 $w(p_T)$.

Let $v(p)$ denote the value function corresponding to stopping optimally Adapting the arguments of Bolton and Harris (1999) to our environment, v must satisfy the Bellman equation

$$0 = \max\{w(p) - v(p); -rv(p) - c + D^2p^2(1-p)^2v''(p)\}, \tag{12}$$

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:

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

All solutions to this equation are in closed form when $D^2/r = 3/4$:

$$v(p) = -\frac{c}{r} + \left(A_1 + A_2\left(p - \frac{1}{2}\right)(1-p)^{-2}\right)p^{-\frac{1}{2}}(1-p)^{\frac{3}{2}}, \tag{13}$$

where A_1 and A_2 are free integration constants. For simplicity, in what follows we set $r = 1$ and

$D^2 = 3/4$ and vary the cost c.[18]

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 (13). The result is depicted in Figure 3.

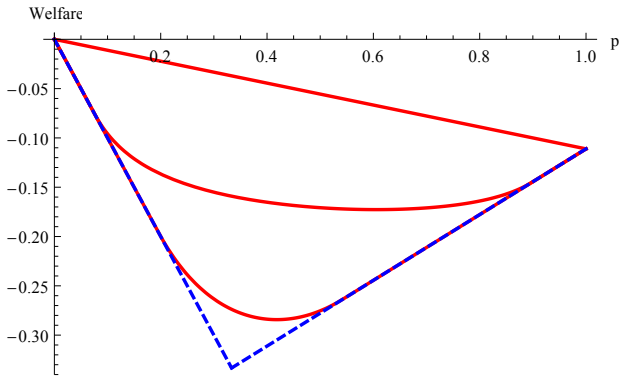

Figure 3: Value function, 2 verdicts, for varying cost levels.

The three-verdict case is more interesting. 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.

[18]Changing r has an equivalent effect if one changes the signal accuracy parameter D to keep D^2/r constant at $3/4$ and the cost parameter c to keep c/r constant.

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.[19]

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 (13) for that region. The result is depicted in Figure 4.

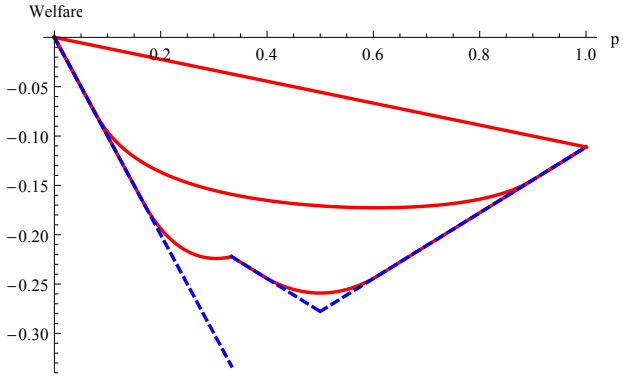

Figure 4: Value function, 3 verdicts, for varying cost levels.

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

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

case, and the right cutoff $\tilde{p}_2$ of the second evidence-gathering region in the three-verdict case is greater than $\hat{p}_2$.

In conclusion, the impact of switching to a three-verdict system by splitting the guilty verdict depends on the evidence gathering cost. When the trial technology is very accurate, 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.

7 Conclusion

This paper considered the introduction of additional verdicts into the standard two-verdict system. It showed that even when restricted to not punishing the innocent more severely or with higher probability, additional verdicts that refine the 'guilty' sentence can be used to increase welfare. Refining the 'innocent' verdict can also increase welfare, if being charged with a crime may carry a stigma. The precise conditions for this increase depend on whether an acquittal (and not only a conviction) leads to the defendant being stigmatized. Two-verdict systems with plea bargains, which can be viewed as a three-verdict systems in which the defendant chooses whether to accept the third verdict (the plea) in lieu of a trial, are shown to be superior to a trial system with any number of verdicts. This is because the defendant knows whether he is guilty when he chooses to take the plea, and this can be used to obtain a more accurate outcome, which increases welfare. This benefit of plea bargains may be attenuated if the defendant is intimidated by an overly harsh 'guilty' sentence, is more risk averse than the general defendant population, or assesses incorrectly the probability of being convicted.

We also considered the effect of a third verdict on the incentives to gather evidence. A third verdict introduces additional regions of the defendant's guilt posterior in which new evidence can change the trial's verdict. When the cost of gathering evidence is substantial, so are these regions, which implies that the additional verdict often leads to more evidence being gathered.

The analysis suggests many extensions and directions for future research. One direction is to replace our reduced form trial technology with a more realistic model of judge and jury. The

strategic interaction among the jurors, and between the judge and the jury, may be affected by the added verdicts in interesting ways. More generally, such changes in the judicial system may affect the incentives and actions of the police, prosecutors, and other agents. These agents may have conflicting goals, and may not always have society's best interests in mind.

A 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 $D : 2^X \to \{g, i\}$, which for every evidence collection decides whether the defendant is guilty or innocent (this can be generalized to a stochastic decision). Distributions P_θ on 2^X, for $\theta \in \{g, i\}$, describe the probability that different evidence collections arise conditional on the defendant being 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 $\{g, i\}$. 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.

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 $i < g$. 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 is equivalent to

$$\begin{aligned}
& \lambda P_g(E)\left(\lambda P_g(E') + (1-\lambda) P_i(E')\right) < \lambda P_g(E')\left(\lambda P_g(E) + (1-\lambda) P_i(E)\right) \\
\iff \quad & \lambda P_g(E)(1-\lambda) P_i(E') < \lambda P_g(E')(1-\lambda) P_i(E) \\
\iff \quad & P_g(E) P_i(E') < P_g(E') P_i(E) \\
\iff \quad & \tfrac{P_g(E)}{P_i(E)} < \tfrac{P_g(E')}{P_i(E')},
\end{aligned}$$

that is, the likelihood ratios are ordered. Observe that this ordering is independent 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:*

$$D(E) < D(E') \Leftrightarrow r(E) < r(E').$$

It is worth emphasizing that, while we started with a Bayesian definition of rationalizability, this definition 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

$$\max\{r(E) : D(E) = i\} < \max\{r(E) : D(E) = g\}. \tag{14}$$

For 2, the answer is "yes:" choose P_g and P_i so that (14) holds. Since 2 implies 3, that answer to 3 is "yes."

Definition of incriminating and exculpatory evidence

If D can be rationalized, then 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.

We immediately have the following result:

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

The reverse need not hold: in particular, one can easily construct examples in which some evidence collection E suffices to convict the defendant ($D(E) = g$), the additional evidence e

reduces the ratio ($r(E \cup \{e\}) < r(E)$), not enough the change the decision, i.e., we still have $D(E \cup \{e\}) = g$.

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.

A.1 The posterior distribution obeys the monotone likelihood ratio property

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 Monotone Likelihood Ratio Property (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.[20] 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 an important role in several of the results below.

[20]This fact is well-known: 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. It is straightforward to establish.

B Comparison of cutoffs and sentences under two- and three-verdict systems

Let (p^*, s^*) be optimal under two-verdict system, and $(p_1^*, p_2^*, s_1^*, s_2^*)$ be optimal under three-verdict system (so that if posterior is below p_1^*, the sentence is 0, if posterior is between p_1^* and p_2^*, sentence is s_1^*, etc.).

Assumption: $W(s, .)$ is concave. Distributions of posterior given state are continuous. Distribution and welfare functions $F(p|.), W(s, .)$ are well behaved so that one can apply the implicit function theorem.

B.1 Comparison of p^* and p_1^*

Proposition 12 *Two-verdict system results in more acquittals: $p^* \geq p_1^*$.*

Proof. We first observe that the two-verdict system can be replicated by a three-verdict system where $p_2 = 1$. Consider a constrained maximization problem where the planner cannot choose p_2. Define $p_1^*(p_2)$, $s_1^*(p_2)$, $s_2^*(p_2)$ as the solutions to this maximization problem. The proposition follows if we can show that $p_1^*(p_2)$ is nondecreasing. In the constrained problem, the actual choice of s_2 has no effect on the optimal choice of p_1, so the part corresponding to s_2 will be dropped.

Therefore, the planner maximizes

$$\begin{aligned}\mathcal{W}_{constrained}(p_1, s_1|p_2) = \quad & \lambda\left[(F(p_2|g) - F(p_1|g))W(s_1, g) + F(p_1|g))W(0, g)\right] \\ & +(1-\lambda)\left[(F(p_1|g) - F(p_1|i))W(s_1, i) + F(p_1|i)W(0, i)\right].\end{aligned}$$

Consider $s_1^*(p_1, p_2)$, which is the optimal sentence taking p_1, p_2 as given, and plug it into the objective, which is now only a function of p_1, p_2 (call the objective $\mathcal{W}_{reduced}(p_1, p_2)$.)

To show that $p_1^*(p_2)$ is nondecreasing, it is enough to show that $\mathcal{W}_{reduced}(p_1, p_2)$ is supermodular. The cross partial equals (applying the Envelope theorem)

$$\frac{\partial^2 \mathcal{W}_{reduced}}{\partial p_1 \partial p_2} = \frac{\partial s_1^*(p_1, p_2)}{\partial p_1}[\lambda f(p_2|g)W'(s_1^*(p_1, p_2), g) + (1-\lambda)f(p_2|i)W'(s_1^*(p_1, p_2), i)]$$

Claim: $\frac{\partial s_1^*(p_1,p_2)}{\partial p_1}$ **is positive.** By the implicit function theorem, $\frac{\partial s_1^*(p_1,p_2)}{\partial p_1}$ has the same sign

as

$$-\lambda f(p_1|g)W'(s_1^*(p_1,p_2),g) - (1-\lambda)f(p_1|i)W'(s_1^*(p_1,p_2),i).$$

However we know that $s_1^*(p_1,p_2)$ satisfies the FOC

$$\lambda[F(p_2|g)-F(p_1|g)]W'(s_1^*(p_1,p_2),g) + (1-\lambda)[F(p_2|i)-F(p_1|i)]W'(s_1^*(p_1,p_2),i) = 0.$$

The claim follows from $\frac{f(p_1|g)}{f(p_1|i)} < \frac{F(p_2|g)-F(p_1|g)}{F(p_2|i)-F(p_1|i)}$ (by MLRP), and $W'(s_1^*(p_1,p_2),g) > 0 > W'(s_1^*(p_1,p_2),i)$.

Claim: $[\lambda f(p_2|g)W'(s_1^*(p_1,p_2),g)+(1-\lambda)f(p_2|i)W'(s_1^*(p_1,p_2),i)]$ **is positive.** This is shown using the FOC for $s_1^*(p_1,p_2)$:

$$\lambda[F(p_2|g)-F(p_1|g)]W'(s_1^*(p_1,p_2),g) + (1-\lambda)[F(p_2|i)-F(p_1|i)]W'(s_1^*(p_1,p_2),i) = 0,$$

and observing that by the MLRP, $\frac{f(p_2|g)}{f(p_2|i)} > \frac{F(p_2|g)-F(p_1|g)}{F(p_2|i)-F(p_1|i)}$, and that $W'(s_1^*(p_1,p_2),g) > 0 > W'(s_1^*(p_1,p_2),i)$. ■

B.2 Comparison of p^* and p_2^*

Proposition 13 *The two-verdict system convicts more often than the three-verdict system gives the highest sentence:* $p^* \leq p_2^*$.

Proof. The approach for the proof is similar to that of the previous proposition. Observe that if we treat s_1 as a parameter, we get the two-verdict system if we set $s_1 = 0$. So the proposition follows if one can show that $p_2^*(s_1)$ is increasing. Consider the optimization over p_1, p_2, s_2, for a given s_1. Similarly to the previous proof, plug into the objective $p_1^*(s_1,p_2)$ and $s_2^*(s_1,p_2)$, which are the optimal values taking both s_1 and p_2 as given. Now the planner maximizes $\mathcal{W}(s_1,p_2)$ over p_2.

$$\begin{aligned}\mathcal{W}(s_1,p_2) = \lambda[&F(p_1^*(s_1,p_2)|g)W(0,g) + (F(p_2|g)-F(p_1^*(s_1,p_2)|g))W(s_1,g) + (1-F(p_2|g))W(s_2^*(s_1,p_2),g)] \\ +(1-\lambda)[&F(p_1^*(s_1,p_2)|i)W(0,i) + (F(p_2|i)-F(p_1^*(s_1,p_2)|i))W(s_1,i) + (1-F(p_2|i))W(s_2^*(s_1,p_2),i)].\end{aligned}$$

Using the implicit function theorem, the envelope theorem, and the fact that $s_2^*(s_1,p_2)$ is

independent of s_1, one can show that

$$\frac{dp_2^*(s_1)}{ds_1} = \frac{\frac{\partial^2 \mathcal{W}(s_1, p_2^*(s_1))}{\partial s_1 \partial p_2}}{-\frac{\partial^2 \mathcal{W}(s_1, p_2^*(s_1))}{\partial p_2^2}} = \frac{\lambda f(p_2^*(s_1)|g) W'(s_1, g) + (1-\lambda) f(p_2^*(s_1)|i) W'(s_1, i)}{-\frac{\partial^2 \mathcal{W}(s_1, p_2^*(s_1))}{\partial p_2^2}}$$

Step 1: at $s_1 = s_1^*$, $\frac{dp_2^*(s_1)}{ds_1} > 0$. $\frac{dp_2^*(s_1)}{ds_1}$ has the same sign as $\lambda f(p_2^*(s_1)|g) W'(s_1, g) + (1-\lambda) f(p_2^*(s_1)|i) W'(s_1, i)$. At s_1^*, one can use the monotone MLRP and the FOC for s_1^* in the unconstrained problem to show that this is positive.

Step 2: if $\frac{dp_2^*(s_1)}{ds_1} = 0$, then $\frac{d^2 p_2^*(s_1)}{ds_1^2} < 0$.

$$\frac{d^2 p_2^*(s_1)}{ds_1^2} = \frac{\frac{d\partial^2 \mathcal{W}(s_1, p_2^*(s_1))/\partial s_1 \partial p_2}{ds_1} \left(-\frac{\partial^2 \mathcal{W}(s_1, p_2^*(s_1))}{\partial p_2^2}\right) - \frac{\partial^2 \mathcal{W}(s_1, p_2^*(s_1))}{\partial s_1 \partial p_2} \frac{d(-\partial^2 \mathcal{W}(s_1, p_2^*(s_1))/\partial p_2^2)}{ds_1}}{\left(-\frac{\partial^2 \mathcal{W}(s_1, p_2^*(s_1))}{\partial p_2^2}\right)^2}$$

Now when $\frac{dp_2^*(s_1)}{ds_1} = 0$, this has the same sign as $\frac{d\partial^2 \mathcal{W}(s_1, p_2^*(s_1))/\partial s_1 \partial p_2}{ds_1}$.

$$\frac{d\partial^2 \mathcal{W}(s_1, p_2^*(s_1))/\partial s_1 \partial p_2}{ds_1} = \frac{\partial^3 \mathcal{W}(s_1, p_2^*(s_1))}{\partial s_1 \partial p_2^2} \frac{dp_2^*(s_1)}{ds_1} + \frac{\partial^3 \mathcal{W}(s_1, p_2^*(s_1))}{\partial s_1^2 \partial p_2} = \frac{\partial^3 \mathcal{W}(s_1, p_2^*(s_1))}{\partial s_1^2 \partial p_2}$$

whenever $\frac{dp_2^*(s_1)}{ds_1} = 0$. Also,

$$\frac{\partial^3 \mathcal{W}(s_1, p_2^*(s_1))}{\partial s_1^2 \partial p_2} = \lambda f(p_2^*(s_1)|g) W''(s_1, g) + (1-\lambda) f(p_2^*(s_1)|i) W''(s_1, i) < 0$$

This finishes step 2.

Step 1 and 2 imply that on $s_1 \in [0, s_1^*]$, $\frac{dp_2^*(s_1)}{ds_1} > 0$. ■

B.3 Comparison of s^*, s_1^* and s_2^*

Finally, the ordering of the sentences follows from the previous two propositions. Intuitively, the optimal sentence reflects how likely the agent is guilty. So 'higher' sets of priors will lead to a longer sentence.

C Pleas with Bayesian updating on the defendant's decision

The following notation will be used in this section:

- λ: prior that defendant is guilty
- α_i, α_g: probability that an innocent or guilty defendant goes to trial
- t: signal observed during trial, $\bar{t}$: threshold for acquittal in terms of signal
- s_b, s_t: sentence from bargain and trial
- $F(t|i), F(t|g)$: cdfs of signal that satisfy MLRP

C.1 Optimal plea without Bayesian updating

From Grossman, Katz we know that the optimal plea has

- $\alpha_i = 1, \alpha_g = 0$
- $s_b = s_b^*, s_t = s_t^*$
- $\bar{t} = \bar{t}^*$

for some numbers $s_b^*, s_t^*, \bar{t}^*$. That the policy for acquitting has a threshold property follows from MLRP. Moreover the guilty is indifferent between trial and bargain, while the innocent strictly prefers trial. In terms of the posterior, the optimal cutoff is

$$\bar{p}^* = \frac{\lambda f(\bar{t}^*|g)}{\lambda f(\bar{t}^*|g) + (1-\lambda) f(\bar{t}^*|i)}$$

This computation is only based on the signal, and not on the decision to go to trial or not.

C.2 Almost optimal plea with Bayesian updating

First note that any policy that is implementable with Bayesian updating is also implementable without. The argument is as follows. Suppose that with updating, the defendant is found guilty

after some set of p's. If both α_i, α_g are strictly positive, any p in that set can be translated to some signal $t(p)$, which in turn can be translated into some p' using Bayes' rule that ignores the decision to go to trial or not. If on the other hand one of the $\alpha's$ is 0, then either no one or everyone gets sentenced guilty, but this can be achieved also under 'no updating'.

If sentencing is forced to be based on the posterior and the posterior takes into account whether the defendant goes to trial or not, now the above policy is not feasible because the posterior at the trial stage is zero.

Consider the following mechanism:

- $\alpha_i = 1, \alpha_g = \epsilon$
- $s_b = s_b^*, s_t = s_t^*$
- $\bar{t} = \bar{t}^*$

In words, we keep sentences and threshold in terms of the signal the same, and ask a small fraction of the guilty to go to the trial. This is acceptable for them since they are indifferent. To implement this in a way where sentencing is only based on the posterior, let the new threshold be

$$\bar{p}^* = \frac{\epsilon\lambda f(\bar{t}^*|g)}{\epsilon\lambda f(\bar{t}^*|g) + (1-\lambda) f(\bar{t}^*|i)}$$

Apart from a small fraction of the guilty defendants, the allocation is the same and hence the welfare difference is of order ϵ.

D Parameters for the welfare functions of Section 6

We set the ideal sentence $\bar{s}$ for the guilty and use quadratic loss functions: $W(s|g) = -(1-s)^2$, $W(s|i) = -s^2$. We also assume that the prior is equal to 1/2: the defendant is equally likely to be guilty or innocent ex ante. To obtain simple expressions for the optimal cutoffs and sentences, we reverse-engineer the signal structure. Recall that the optimal cutoff is given by the indifference condition

$$p^* W(s^*, g) + (1-p^*) W(s^*, i) = p^* W(0, g),$$

or

$$p^*(1 - (s^*)^2) + (1-p^*)(-(s^*)^2) = p^*.$$

The optimal sentence is given by the first-order condition deriving from

$$s^* \in \arg\max_s \frac{1}{2} Pr(p \geq p^*|g)W(s|g) + \frac{1}{2} Pr(p \geq p^*|i)W(s|i),$$

i.e.,

$$(1 - F(p^*|g))(1 - s^*) = (1 - F(p^*|i))s^*.$$

By choosing $F(\cdot, g)$ and $F(\cdot, i)$ so that the ratio $q = \frac{1-F(p|i)}{1-F(p|g)}$ is equal to $1/2$ when evaluated at $p = 1/3$, we verify that $p = 1/3$ and $s = 2/3$ solve the problem. Note that q must be less than 1, from MLRP.

With three verdicts, we impose the restrictions $p_1 = 1/3$ and $s_2 = 2/3$ (so that we are indeed splitting the guilty verdict, and not increasing the guilty sentence), and optimize over the remaining two parameters, p_2 and s_1. These parameters are again characterized by the indifference equation for p_2, given the sentences s_1 and s_2 that are given above and below p_2,

$$p_2 W(s_1, g) + (1 - p_2)W(s_1, i) = p_2 W(s_2, g) + (1 - p_2)W(s_2, i),$$

and by the optimality condition for s_1, which is

$$s_1 \in \arg\max_s \frac{1}{2} Pr(p \in [p_1, p_2]|g)W(s|g) + \frac{1}{2} Pr(p \in [p_1, p_2]|i)W(s|i),$$

which yields the first-order condition

$$F([p_1, p_2]|g)(1 - s_1) = F([p_1, p_2]|i)s_1.$$

Again doing reverse engineering, we choose $F(\cdot|g)$ and $F(\cdot, i)$ so that the ratio $q' = \frac{F([p_1,p_2]|i)}{F([p_1,p_2]|g)}$ evaluated at $p_1 = 1/3$ and $p_2 = 1/2$ be equal to 2. With this condition, $s_1 = 1/3$ and $p_2 = 1/2$ satisfy all conditions. Note that the ratio q' must be greater than q, by MLRP.

This yields the welfare functions $w_2(p) = w_3(p) = -p$ for $p < 1/3$, $w_2(p) = -p/9 - (1-p) \times 4/9$ for $p \geq 1/3$, and $w_3(p) = -p/9 - (1 - p) \times 4/9$ for $p \geq 1/2$, and $w_3(p) = -p\frac{4}{9} - (1 - p)\frac{1}{9}$ for $p \in [1/3, 1/2)$.

References

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

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

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

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

[5] Daughety, A., Reinganum, J. (2015a) "Informal Sanctions on Prosecutors and Defendants and the Disposition of Criminal Cases," *Working Paper*, Department of Economics and Law School, Vanderbilt University.

[6] Daughety, A., Reinganum, J. (2015b) "Selecting Among Acquitted Defendants: Procedural Choice vs. Selective Compensation," *Working Paper*, Department of Economics and Law School, Vanderbilt University.

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

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

[9] 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.

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

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

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

[13] Rakoff, J. (2014) "Why Innocents Plead Guilty," *The New York Review*, November 20, 2014 Issue.