Siegel, Ron; Strulovici, Bruno

Working Paper

On the design of criminal trials: The benefits of a three-verdict system

Discussion Paper, No. 1581

Provided in Cooperation with:
Kellogg School of Management - Center for Mathematical Studies in Economics and Management Science, Northwestern University

Suggested Citation: Siegel, Ron; Strulovici, Bruno (2015) : On the design of criminal trials: The benefits of a three-verdict system, Discussion Paper, No. 1581, Northwestern University, Kellogg School of Management, Center for Mathematical Studies in Economics and Management Science, Evanston, IL

This Version is available at:
https://hdl.handle.net/10419/119414

CMS-EMS
Center for Mathematical Studies in Economics
And Management Science

Discussion Paper #1581

On the Design of Criminal Trials: The Benefits of a Three-Verdict System

Ron Siegel*
Bruno Strulovici**

March 3, 2015

JEL Classification: D02, D7, D81, D82, D83, K1, K4

Keywords: Trials, Verdict, Reasonable Doubt, Stigma, Evidence, Plea bargaining

* and ** Northwestern University

On the Design of Criminal Trials: The Benefits of Three-Verdict Systems

– Preliminary and Incomplete –

Ron Siegel and Bruno Strulovici*

Northwestern University

March 3, 2015

1 Introduction

Trial decisions involve a trade-off between two types of errors: convicting the innocent and acquitting the guilty. Both types of errors occur in practice, despite the fact that a guilty verdict requires proof beyond a reasonable doubt. For example, a recent study[1] 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. On average, death-row inmates who ended up being exonerated spent 13 years in jail before their release.

The trade-off is particularly acute when there is significant incriminating evidence, which is not fully conclusive. In such cases, two-verdict systems force jurors to choose between a high probability of acquitting a guilty person and a small probability of convicting an innocent person.

This paper considers the possibility of more expressive verdicts, which better reflect the likelihood a defendant's guilt. We focus on the introduction of a third verdict, which may be

*David Rodina provided excellent research assistance. Strulovici gratefully acknowledges financial support from the NSF (Grant No. 1151410) and the Alfred P. Sloan Foundation. Email: r-siegel@northwestern.edu and b-strulovici@northwestern.edu

[1]See Gross et al. (2014).

and has already been discussed in the legal literature (Bray, 2005). To isolate the issue at hand, we ignore factors such as mitigating circumstances - which in reality affect the length of the sentence for a given verdict, but not the verdict itself. We also initially take the evidence formation process as exogenous.[2]

Splitting the 'guilty' verdict

By providing a richer set of instruments, the introduction of a third verdict weakly improves social welfare in a utilitarian sense. However, could this utilitarian improvement come at the expense of other ethical considerations? Most importantly, could a three-verdict system lead to some innocent defendants being punished more often or more severely than in the two-verdict system? This would be the case if even a light sentence is handed down given evidence that would lead to acquittal in a two verdict system. The prospect of such a system may seem objectionable to many.

The first contribution of this paper is to show that the three-verdict system can always be designed so as to strictly improve social welfare, without increasing the probability of punishing innocent defendants and without increasing their sentence should they be found guilty. To achieve this, we create a third verdict by splitting the guilty verdict of the two-verdict system according to the amount of evidence that has been accumulated against the defendant: the sentence is reduced when the incriminating evidence is comparatively weak, and unchanged otherwise.

While unnecessary for this part of our analysis, it is helpful to think of verdicts as being determined by the posterior probability that the defendant is guilty, given the evidence accumulated before and during the trial. Let p denote this probability and, for the sake of concreteness, suppose that the defendant is found guilty in the two-verdict system if p lies above some cutoff, say 80%.[3] Also suppose that a convicted defendant is sentenced to 10 years in jail. Then, our three-verdict improvement is, qualitatively, of the following form: give only 7 years, say, to de-

[2]Section 8 analyzes how verdict systems affect the incentives for gathering evidence. It suggests that evidence formation need not be adversely affected by the three-verdict system that we propose. If anything, it may actually be enhanced.

[3]In an ideal world where evidence can be mapped unambiguously into a posterior probability of guilt, utilizing such a threshold is the optimal way of resolving the trade-off between the two types of errors in a two-verdict system.

fendants whose posterior probability lies between 80% and 90% and 10 years when the posterior is above 90%. This modification keeps the probability of punishing the defendant unchanged, compared to the two-verdict system, and never increases his sentence.

By isolating "marginal" convictions of the two-verdict system and assigning them a lower sentence, the three-verdict system clearly benefits the innocent. But it also reduces the sentence of guilty defendants, so how can this always lead to a utilitarian improvement? Intuitively, a guilty defendant is more likely than an innocent one to produce incriminating evidence. Thus, conditional on being found guilty in a two-verdict system, an innocent defendant is more likely to be a marginal case (between 80% and 90%, in the above example) than a more definitive one (above 90%), relative to a guilty defendant. The sentence reduction introduced by the third verdict is thus more likely to benefit an innocent defendant than it is to mistakenly benefit a guilty one, which results in a welfare improvement.

From evidence to posterior probability

The posterior-probability formalism used above is not needed for the result. However, we show 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. This property is just a consequence of what it means, precisely, for some piece of evidence to be incriminating.

Splitting the "not guilty" verdict in the presence of social stigma

In addition to their sentences, defendants are also subject to social stigma.[4] This creates another channel through which three-verdict systems can be beneficial. Because a high evidence requirement is set for convictions, defendants may be acquitted for lack of sufficient evidence. Such acquittals do not always imply that the defendants really are innocent, or that they are (or should necessarily be) perceived as such in the eye of the public. For example, a defendant in a sexual offense trial may be found not guilty for lack of proof, but knowledge of the case may be a valid cause of concern for those with whom the defendant interacts.

The higher the burden of proof required for conviction, the more likely a defendant is to be guilty conditional on being acquitted. In the above example, a defendant receives a "not guilty" verdict as long as the posterior probability of being guilty is less than 80%, but a defendant who is even 50% likely to have committed a crime is at least of some concern to society.

The two-verdict system cannot discriminate between defendants who are cleared through the evidence provided during the trial and defendants who are only acquitted for lack of evidence. Just knowing that the defendant has been acquitted in a rape case, for instance, says little about his precise likelihood of guilt.

A third verdict may be introduced, by splitting the "not guilty" verdict according to the evidence provided. This is actually done in a few legal systems,[5] which distinguish between "innocent" and "not proven" verdicts. In these systems, a defendant is found "innocent" if the evidence produced in the trial has cleared him from wrongdoing. A "not proven" verdict signals that, while the evidence is insufficient to convict the defendant, it casts serious doubt on his innocence. Both verdicts carry no jail time: the defendant is free to go.

These systems exploit social stigma to improve welfare: by signaling a lower likelihood of guilt, the "innocent" verdict reduces the stigma faced by the defendant, which is beneficial since the defendant is indeed more likely to be innocent. Likewise, a "not proven" verdict increases the stigma of the defendant. Social stigma may be beneficial in at least two ways: first, it protects society from likely offenders. Second, it indirectly punishes defendants who are actually guilty but have received no jail time.

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

[5]Such systems are in place in Scotland, Israel, and Italy.

Because splitting the innocent verdict has no effect on the sentence imposed on the defendant, it is immune to the ethical concerns raised earlier: no innocent defendant ever goes to jail more often or for longer than in the two-verdict system.

In contrast to splitting the guilty verdict, however, splitting the innocent verdict may decrease social welfare, even though an innocent defendant is more likely than a guilty one to generate an 'innocent' verdict relative to a 'not proven' verdict. This is because, intuitively, if the stigma function is convex in the posterior p, then the stochasticity added by a third verdict can harm risk averse defendants. We provide general conditions under which splitting the innocent verdict is socially beneficial, and conditions under which it is detrimental.

Plea bargains

More than 90% of criminal cases in the United States are settled by plea bargains. In a plea bargain, the defendant does not go to trial and accepts a lower sentence than the one associated with a guilty verdict. While plea bargains are expedient, they have been severely criticized for a number of reasons, including the extreme power they give prosecutors, the secrecy of the agreements, the lack of judicial oversight, and the fact that innocent defendants are sometimes scared into taking pleas (studies estimate 2-8 percent of the plea bargains are taken by innocent defendants).

Because plea bargains significantly depart from any ideal of justice,[6] and because they are so prevalent, a number of judges and legal scholars have emphasized the importance of experimenting with different methods to improve upon them.

We analyze the value of plea bargains relative to other verdict systems. Pleas may be seen as a third verdict, which is proposed to the defendant before the trial. Because this third verdict is chosen by the defendant, Grossman and Katz (1983) show that it can serve as a screening device: guilty defendants are more likely to be found guilty during a trial, so they are more willing to take the plea.

Perhaps surprisingly, 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,

[6] See Rakoff (2014).

provided that the defendant's utility function is independent of his guilt. In fact, we show that there is a two-verdict system with a plea that maximizes welfare among all incentive compatible mechanisms. In that optimal mechanism, the guilty sentence coincides with the sentence that is optimal when one is certain that the defendant is guilty.

In analyzing the welfare of plea bargaining, one must take a stand on how a defendant's decision to reject a plea is incorporated in the jury's or judge's posterior belief about the defendant's guilt. This, it turns out, does not affect the set of implementable outcomes in two-verdict systems with pleas: whether the posterior belief used to decide on a verdict includes this information does not affect the set of equilibrium outcomes.

Despite its generality, the result on the superiority of plea bargain systems omits at least two issues. Firstly, the sentences 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* the defendant were 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 idea, we show by example that when the guilty sentence is set at an excessively high level (relative to the utilitarian optimum), the two-verdict system with plea may not perform well. In particular, it may be dominated by a three-verdict system of the form described above, in which the guilty verdict of the two-verdict system is split into two, and the intermediate verdict is assigned a lower sentence. Secondly, one may construct examples in which an innocent defendant who overestimates the probability of being found guilty at the time of trial, perhaps through persuasion or intimidation, may be indeed scared into taking a plea. In this case, a three-verdict system may again dominate the two-verdict system with a plea.

Incentives to Search for Evidence

We have ignored until now how three-verdict systems affect incentives for producing evidence. Obviously, a system without pleas brings more evidence to light through trials than a system in which a large proportion of cases are resolved via plea bargains, as is currently the case in the United States. Even at the pre-trial stage, a prosecutor may have less incentives to accumulate information if he knows that the defendant is likely to accept a plea.

Setting pleas aside, we investigate how a three-verdict system affects the value of evidence formation relative to the two-verdict system from which the three-verdict system has been obtained. We show that, in general, the three-verdict increases the value of evidence.

2 Baseline model: two verdicts, no pleas

We introduce a simple model of trial design. The objectives of a trial are to determine whether a defendant is guilty and to deliver an appropriate sentence. A trial may be conceptualized as a number of signals regarding the defendant's guilt: evidence is produced and interpreted, resulting in each juror forming a posterior belief about the defendant's guilt. This process is complex: it involves numerous agents of different types (the defendant himself, lawyers, judges, witnesses, experts, etc.) and a large number of heterogenous pieces of evidence. The verdict itself may be the result of a collective decision based on a long deliberation and strategic behavior.

For the present purpose, however, this process may conveniently be summarized by two numbers: the probability π_g that a defendant who is actually guilty is found guilty, and the probability π_i that an innocent is found guilty.[7]

We focus on how beliefs about a defendant's guilt translate into a sentence and on how this map captures social objectives. With this in mind, we purposefully leave aside such issues as mitigating circumstances. A guilty verdict is thus simply characterized by a sentence, $s > 0$, interpreted as jail time (so a higher value of s corresponds to a higher punishment).

Society has a dual goal: to avoid punishing the innocent and to adequately punish a guilty defendant. This dual goal is modeled by a welfare function, denoted W. Jailing an innocent 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 social welfare of $W(s, g)$, which is assumed to have 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 guilty: if the defendant is *ex ante* very likely to be guilty, one should be mainly concerned about delivering an adequate punishment following a guilty verdict. If, by contrast,

[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.

the defendant is likely to be innocent, the main concern shifts toward the risk of inflicting a severe sentence to the defendant, should he be found guilty by mistake. These concerns are captured by the ex-ante social welfare function, which is given by

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

Since $W(\cdot, i)$ is decreasing and $W(\cdot, g)$ peaks at $\bar{s}$, it is always beneficial to choose $s \leq \bar{s}$. Even without restricting attention to the welfare-maximizing level of s (which is obviously less than $\bar{s}$, it is natural to assume that the relevant range of s is contained in $[0, \bar{s}]$, and we impose this condition.

3 Splitting the guilty verdict

The evidence collected before and during the trial determines the verdict. Consider the set $\mathcal{S}$ of evidence collections leading to a guilty verdict. Because many possible collections are consistent with a guilty verdict, this set is not a singleton. For example, evidence leading to a guilty verdict may include the discovery of a gun 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 any union of these. To construct a new trial desing, we split $\mathcal{S}$ into two proper subsets $\mathcal{S}_1$ and $\mathcal{S}_2$, arbitrarily, and introduce two verdicts, denoted for now 1 and 2 according to whether the collected evidence belongs to $\mathcal{S}_1$ or $\mathcal{S}_2$. Let π_i^1 denote the probability that the innocent receives verdict 1. π_i^2, π_g^1, and π_g^2 are defined similarly.

Because the "guilty" verdict is now split in these two verdicts and the overall probability hasn't changed, we have

$$\pi_i = \pi_i^1 + \pi_i^2 \qquad \pi_g = \pi_g^1 + \pi_g^2.$$

Without loss of generality,[8] we can label the verdicts 1 and 2 so that

$$\frac{\pi_g^1}{\pi_i^1} < \frac{\pi_g}{\pi_i} < \frac{\pi_g^2}{\pi_i^2}$$

Intuitively, verdict 1 is an intermediary verdict: the likelihood that the innocent receives it, relative to the guilty defendant, is higher than with the verdict 2. Verdict 2 is the most serious verdict: a guilty defendant is more likely to get it.

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

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

The improvement from splitting the guilty verdict is captured by the following result. There are sometimes limits on the maximal sentence that can be used for a given crime. In particular, it makes sense to assume that $s \leq \bar{s}$, where $\bar{s}$ is the optimal sentence when one knows for certain that the defendant is guilty. Say that a sentence s is *interior* if $s < \bar{s}$.

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 modulates the sentence so as to reflect the richer information contained from the verdicts 1 and 2, which are used here as signals.

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

[8] We use the fact, 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 the verdict is split according to the posterior probability that the agent is guilty.

$s_1 = s - \varepsilon$ and $s_2 = s + \varepsilon\gamma$. The welfare impact of this perturbation is

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

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 $W'(s, g)(-\pi_g^1 + \gamma\pi_g^2)$ and $W'(s, i)(-\pi_i^1 + \gamma\pi_i^2)$ are both positive, which shows the claim. ■

Proposition 1 shows that three verdicts can always improve utilitarian welfare, compared to any two-verdict design with an interior sentence. Crucially, the improvement does not increase the probability of punishing an innocent defendant (or a guilty one). However, the improvement requires an *increase* in maximal sentence, from s to $s_2 > s$. Therefore, an innocent is erroneously convicted may face, ex post, a worse jail sentence.

The next proposition sharpens this result by showing that if the sentence s of two-verdict system was optimal to begin with, and if it was interior, then one can construct an improvement that only reduces sentences.

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

This proposition thus says that if the two-verdict system was optimal to begin with, then it is possible to improve it with a three-verdict system by introducing an intermediary verdict that weakens the sentence.

Proof. By construction s^* maximizes

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

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

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

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. ■

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.

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: $W(\cdot, i)$ and $W(\cdot, g)$ are concave in s. The posterior distributions $F(\cdot|i)$ and $F(\cdot|g)$ are both continuous. Finally, W F are regular so that one can apply the implicit function theorem.

Proposition 3 *$p_1^* \leq p^* \leq p_2^*$ and $s_1^* \leq s^* \leq s_2^*$.*

Intuitively, the optimal sentence reflects how likely the agent is guilty. So 'higher' sets of priors will lead to a longer sentence.

The proof of this proposition is in Appendix A

4 From an Evidence-Based Verdict to a Posterior-Based Verdict

We begin this section by the presentation of a Bayesian conviction model in which verdicts are determined by the posterior probability that the defendant is guilty, given all the evidence accumulated during the trial, as well as the prior probability of guilty. We then provide a

foundation for this model, showing that any "reasonable" verdict rule based on evidence may be formalized as a Bayesian model.

4.1 Presentation of the Bayesian Conviction Model

The previous section established that splitting the guilty verdict is always welfare improving, without imposing any structure on how verdicts are determined. 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 which is used to form this posterior belief, p. For now, it suffices to summarize it by distributions $F(\cdot|g)$ and $F(\cdot|i)$ describing the posterior depending on the defendant's actual guilt level.[9]

When a verdict system is based on the defendant's posterior, it is natural to follow a cutoff rule. In a two-verdict system, in particular, the defendant is relaxed if the trial has not gathered enough evidence against him. Formally, this means that the sentence is $s = 0$ if the posterior probability of guilt lies below a threshold p^*. If instead $p > p^*$, 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

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

In what follows, we will denote (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.

This optimum is actually constrained by the restriction that an "innocent" verdict necessarily leads to relaxation ($s = 0$). Within the model, this restriction may be questioned: for example, if $p^* = 90\%$ (i.e., the defendant gets the guilty verdict only if there is a 90 percent chance that

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

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

he is guilty), then an "innocent" verdict captures many likelihoods of guilt. For example, if $p = 80\%$, the defendant gets the "innocent" verdict for lack of sufficient evidence, even though the probability that he is guilty is quite high.

In practice, however, a defendant may face significant stigma even if he is found innocent through his trial. If guilt must be established "beyond any reasonable doubt" to deliver a guilty verdict, then it follows that an "innocent" verdict may well be compatible with significant lingering doubt, and such doubt may harm even a defendant who is innocent and has been found innocent by the jury.

To address this issue, it is helpful to send a public signal about the amount of doubt in the case, even if the defendant is relaxed. The next section explores whether this possibility

[Rm: if the public has access to the evidence, then the stigma depends on the evidence produced, not just on the verdict. The distinction]

4.2 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? Probably yes.

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 4 *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

$$\max\{r(E) : D(E) = i\} < \max\{r(E) : D(E) = g\}. \tag{7}$$

For 2, the answer is "yes:" choose P_g and P_i so that (7) 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 5 *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.

4.3 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.[10] 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 6 *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.

5 Splitting the 'not guilty' verdict

As mentioned in the introduction, the only three-verdict system which has received some attention in the legal literature and has found some implementation in actual jurisdictions is based on the introduction of a 'not proven' verdict. Like the not guilty verdict, the not proven verdict entails no jail time. However, it signals a much weaker belief that the defendant is actually not guilty. The importance of the distinction, as has been argued elsewhere, lies in the stigma faced by defendants after the trial. However, we are unaware of any mathematical model behind the argument. This section formalizes and investigates the claim that such a split increases social welfare.

Let q denote the posterior probability, in the eye of the public, that the agent is guilty given the verdict. Crucially, and as argued in Bray (2005)), the public mainly sees the verdict, not the

[10]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.

detailed evidence brought during the trial. Therefore, the public's information is coarser and this posterior belief q is conditioned only on the verdict.

There are two intuitive ways of modeling stigma. It can enter welfare additively, so instead of $W(s,i)$, we have $W(s,i) - d(q)$ for some positive and increasing function d, or it can act as increase of the sentence, so instead of $W(s,i)$ we have $W(s+d(q),i)$. We only consider splitting the 'not guilty' verdict into two verdicts that still carry a sentence of zero jail time. The threshold for acquitting the agent is the same so there will be the same number of acquittals.

5.1 Stigma enters welfare additively

Generally there are two effects from splitting the innocent verdict. The first effect is that more information is provided. By helping the public discriminate between defendant's type, this effect makes an innocent defendant better off and a guilty defendant worse off. The second effect comes from the curvature of the disutility from stigma. If the cost from the stigma is convex, an agent prefers not to split the verdict, and vice versa if the cost is concave.

Consider the following parametrization. Let I be the event of acquittal under two-verdict system, and I_1 and I_2 be the two innocent verdicts after we split the innocent verdict into two verdicts, where I_1 occurs for priors below a cutoff, and I_2 for priors above a certain cutoff. Also, let q, q_1, q_2 denote the probability that the agent is guilty conditional events I, I_1, I_2, respectively.

Under a single innocent verdict the relevant part of the welfare function is

$$\lambda[W(0,g) + \gamma d(q)] + (1-\lambda)[W(0,i) - \delta d(q)]$$

Under two innocent verdicts the relevant part of the welfare function is

$$\lambda[W(0,g) + \gamma E[d(q_j)|g]] + (1-\lambda)[W(0,i) - \delta E[d(q_j)|i]]$$

Stigma and welfare Presumably $\delta \geq 0$ and $\gamma \geq -\delta$, so that one wants to avoid stigma of the innocent, and minds less about the stigma of the guilty. Moreover, because neither the not

guilty nor the not proven verdicts entail any jail time, a moderate amount of stigma is arguable beneficial on the guilty: so it is natural to assume that $\gamma > 0$.

The proposition below follows from the following observation:

Observation 1
- $q = Pr(I_1)q_1 + Pr(I_2)q_2$.
- $q \geq Pr(I_1|i)q_1 + Pr(I_2|i)q_2$.
- $q \leq Pr(I_1|g)q_1 + Pr(I_2|g)q_2$.

Proposition 7 *Suppose that the stigma enters welfare additively. Then, the following holds.*

- *If d is linear, then welfare*
  - *doesn't change if $\gamma = -\delta$.*
  - *increases if $\gamma > -\delta$.*
- *If d is concave, then welfare*
  - *increases if $\delta > 0, \gamma \in [-\delta, 0]$.*
- *If d is convex, then welfare*
  - *increases if $\delta = 0, \gamma > 0$.*
  - *decreases if $\gamma = -\delta$.*

5.2 Stigma increases sentence

Consider now that we can split the innocent verdict with an arbitrary garbling, so that q_1 and q_2 are arbitrarily close to q. Since the stigma enters welfare in terms of the sentence like $W(s + d(q), .)$, W and d can be treated as linear if they are smooth. I thought that one can apply a similar argument that was used to show that one can always improve by splitting the guilty verdict into two different sentences. However it is not clear that one can always find a strict improvement, because for a very uninformative garbling such that q_1 and q_2 are arbitrarily close to q, the likelihood ratios for these events between innocent and guilty will be close to 1.

If we consider more general garblings, like having a cutoff as in the previous subsection no general statement seems to be possible.

6 Multi-verdict systems

Before discussing pleas in the next section, it is useful to generalize the analysis to an arbitrary number of verdicts. Suppose, ideally, that the court was free to set arbitrarily the sentence s as a function of the posterior p. Given a posterior p, the optimal sentence $s(p)$ is chosen so as to maximize the welfare objective

$$pW(s,g) + (1-p)W(s,i) \tag{8}$$

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 (8) must be less 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 $W(\cdot, g)$ is increasing in the relevant range $[0, \bar{s}]$ and $W(\cdot, i)$ is decreasing, implying $\partial\mathcal{W}/\partial p = W(s,g) - W(s,i)$ is increasing in s. This implies that the selection of maximizers (8) is isotone. In particular, there exists a nondecreasing selection $s(p)$ of optimal sentences.

The argument used for Propositions 1 and 2 can be easily generalized 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 $p_1 > 0 = p_0$, $p_k < 1 = p_{k+1}$, $s_0 = 0$ and $s_k \leq \bar{s}$. In this system, a defendant gets sentence $s_{k'}$ whenever his posterior p lies in (p_k, p_{k+1}).

Proposition 8 *Suppose that the signal distributions are continuous for both the guilty and innocent defendants. Then, for any k-verdict, there is a $k+1$ verdict that strictly increases welfare. Moreover, if a k-verdict is optimal among all k-verdicts and either $k > 2$ or $k = 2$ and $s_1 < \bar{s}$, there is a $k+1$-verdict that strictly improves upon it and has lower sentences.*

Thus, Proposition 8 implies that

7 Pleas

More than 90% of criminal cases in the United States are concluded by a plea agreement and do not go to trial. It is therefore important to evaluate the social value of this institution. From our viewpoint, pleas are particularly interesting, as they constitute a kind of third verdict: in terms of outcomes, defendants can get sentences corresponding to being found guilty or not guilty during the trial, or they can receive an intermediate sentence at the plea bargain stage. Of course, this third verdict is different from the third verdict discussed so far, because it involves a strategic decision by the defendant of whether to take the plea, and hence a more active role than in a multi-verdict trial where the defendant passively receives one of the verdicts. As we shall see, this strategic aspect plays an interesting role on welfare.

Following Grossman and Katz (1983), hereafter "GK", we model pleas as follows: in the first stage, the defendant is offered a plea sentence s^b. If the defendant accepts it, he gets this sentence and the case is closed without trial. 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 like in the previous sections. Following GK, we assume that the jury or judge does not use the information revealed by the choice of the defendant of rejecting the plea. This assumption may seem troubling at first: especially in a separating equilibrium, only an innocent defendant goes to trial. However, it turns out that the distinction is irrelevant: the same outcomes can be achieved with and without updating based on the defendant's decision. This fact is shown in Appendix B. Ignoring the defendant's decision in assessing his guilty is also consistent with legal requirements to focus only on the evidence presented during trial to assess the guilt of the defendant.

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, as a function of his type. These four numbers determine an entry decision for each type of defendant. It is assumed that both types of defendants have the same utility function, u, regardless of their guilt. Because a guilty defendant is more likely to be found guilty if he goes to trial (i.e., $\theta_g > \theta_i$) his incentive to go to trial are strictly lower than an innocent defendant's.

Therefore, depending on the parameters, only three situations can arise: both types of

defendants take the plea, only the guilty defendant takes the plea, or both defendants go to trial (mixing can be showed to be suboptimal).

Grossman and Katz (1983) prove that the optimal system is separating. Precisely: i) the plea s^g is chosen so as to make the guilty defendant indifferent between taking the plea and going to trial, ii) the guilty defendant takes the plea, and iii) the innocent defendant goes to trial.

Because plea bargains have been widely criticized, one might expect them to do poorly in terms of welfare. Surprisingly, however, we show that in the present setting, they outperform *any* multi-verdict system without pleas, including the two and three verdicts.

7.1 The welfare value of plea bargaining with only two types of defendant

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

Proposition 9 *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

$$U^g \int_0^1 u(s(p)) f_g(p) dp = \int_0^1 u(\hat{s}(p)) f_g(p) dp = u(0) F_g([0, \hat{p}]) + u(s(1)) F_g([\hat{p}, 1]) = \hat{U}^g, \qquad (9)$$

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 (9) 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. 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[11]

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

or

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

Thus, the new verdict system increases the expected utility of an innocent defendant. 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

[11] 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.

well-known from the mechanism design literature that in the present setting there exists a direct revelation mechanism: the defendant make a reports $\hat{\theta}$ of his type (guilty or innocent) and is then assigned a sentence $s(t, \hat{\theta})$ that depends on his report and on the signal t generated during trial, which is assume to be continuous on $[\underline{t}, \bar{t}]$ and satisfy the MLRP. A mechanism is feasible if $s(t, \hat{\theta})$ is less than $\bar{s}$ for all t and θ: i.e., it does not punish the defendant by 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 10 *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, jumping from 0 to $\bar{s}$. The IC constraint of the guilty defendant is binding. The cutoff t^* at which $s(\cdot, i)$ jumps from 0 to $\bar{s}$ decreases in the prior from $\bar{t}$ to $\underline{t}$.*

7.2 The failure of plea bargaining with excessive sentencing

Despite the encouraging result of the previous section, pleas have been severely criticized for leading innocent defendants to accept jail time rather than go to trial. A recurring problem, which has been pointed out repeatedly[12] is that the sentences given at trial are excessively harsh. This section provides an example that captures 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 (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 also assume that the social welfare function facing a guilty defendant is linear with a peak at 20 years: $W(s, g) = -|s - 20|$. Given this, we only consider sentences which are lower than the sentence $\bar{s} = 20$ that is socially optimal to give to a guilty defendant.

Finally we suppose that the trial can generate two types of evidence against the defendant,

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

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 relaxed.

The purpose of this section is to show that plea bargaining with two verdicts can do poorly when the guilty sentence is excessively high, relative to a three verdict system that splits the guilty verdict and keeps the excessively high sentence for strong evidence.

Inefficiently high sentences

Owing to 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 plea, it is optimal to punish the defendant for either type of evidence (weak or strong), and the optimal sentence is $s_1 = 16$;
- In a two-verdict system with plea, the same as above holds, and only the guilty defendant takes the plea;
- If, however, the 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 set at $s^b = 0.8 * s_2 = 16$, and both guilty and the risk averse defendant 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 plea, with a high sentence of $s_2 = 20$.

This result illustrates that the introduction of an intermediate verdict with a lower sentence may be more efficient than a plea to counteract the requirement that the highest sentence be set so as to adequately punish a guilty defendant. This again 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, the guilty sentence would be reduced (here, to 16) and pleas may be optimal. However, if the guilty sentence is set too high from a utilitarian

perspective (which may be hard to modify due to political considerations), plea bargaining not be the right way.

Erroneous beliefs

In this example, the social welfare conditional on facing an innocent defendant coincides with the utility of a defendant. Thus, if an innocent defendant optimally takes the plea, it means that is socially optimal for him to do so. Suppose, however, as suggested by some observers, that defendants are sometimes scared into believing that the trial outcome is more likely to be bad than it really is. For example, suppose that the risk averse innocent defendant erroneously believes that the probability of weak evidence being found against him is higher than it is, say equal to 75%. Then he may prefer to take the plea 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.

8 Incentives for Evidence Formation

Previous sections have taken as given the technology generating evidence in favor of or against the defendant. However, searching 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 possibility of biases in these agents' behavior, the socially optimal amount of information to be acquired in a case clearly depends on the verdict structure employed for that case. For example, a trial system in which a single verdict is given regardless of the evidence produced would clearly destroy all incentives for gathering evidence. In fact, this criticism has made against plea bargaining: since so many defendants take a plea, this reduces incentives for information acquisition.

In this section, we wish to compare the impact on evidence formation of switching from two verdicts to three verdicts. For simplicity, we focus on the situation presented in Section **??** in which the guilty verdict of the two-verdict system is split into two distinct verdicts. When all the evidence of a case has been presented, the jury or judge returns a verdict which depends on

the posterior belief p that the defendant is guilty.

A verdict system implies a welfare level

$$w(p) = pW(s(p), g) + (1 - p)W(s(p), i), \tag{10}$$

where $p \mapsto s(p)$ is a step function starting at zero with only two levels in the 2-verdict case, and three levels for a three-verdict system. As mentioned, the shape of this function affects the incentives to acquire information. From (10), the welfare function $w(p)$ is piecewise linear. Since $W(0, i) = s(0) = 0$, w starts at 0, decreases as the probability p of letting the guilty go free increases, and then has a kink when the sentence jumps from 0 to a positive level. Figure 1 represents the welfare function w_2 for the optimal 2-verdict system when $W(\cdot, g)$ and $W(\cdot, i)$ are quadratic, for parameters given in the appendix.

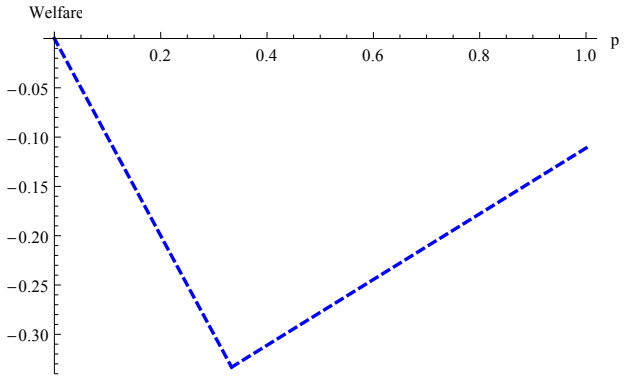

Figure 1: Welfare function, 2 verdicts.

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 w_3 for the optimal 3-verdict system obtained by splitting the guilty verdict in the two verdict system 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 w_3 is discontinuous at p_1: this reflects the fact that p_1 is not chosen optimally, but rather inherited from the 2-verdict system. By contrast, because p_2 is chosen optimally, w_3 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, and take the viewpoint of a social planner wishing to aggregates information until a verdict is reached.

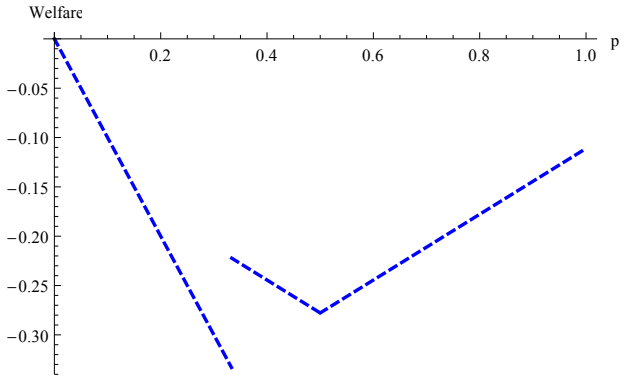

Figure 2: Welfare function, 3 verdicts.

The tradeoff at the heart of this task is clear: more effort spent searching for evidence means more cost for society but more precise information about the defendant's guilt. There are many ways of modeling this tradeoff; we discuss two: an elementary one with a one-shot decision, and which already captures the rough intuition for why two verdict and three verdict systems are not comparable in terms of information acquisition. The second model is more sophisticated and provides a visually more appealing representation of the impact of a third verdict on information acquisition.

8.1 One-shot information acquisition

The simplest model of information acquisition, which suffices to provide the intuition, is a one-shot model: the judge decides whether to search for 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 posterior p' has a mean equal to $1/2(p + \Delta) + 1/2(p - \Delta) = p$, i.e., the prior.

When is information acquisition desirable? Suppose first that the prior lies close to zero, so that the posterior p' will like on the first branch of the graph of w_2. Then, the value of information 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 lie in on the second branch of the w_2, regardless of the outcome, the value of information is zero: intuitively, the information is not enough to change the verdict and hence is valueless.

Consider now the case of three verdicts. For p slightly above a $p_1 + \Delta$, information 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.

However, for slightly p less than p_1, the value of information is large, because a positive belief update triggers a high improvement in welfare. Similarly, for p in a neighborhood of p_2, the value of information is positive whereas it is equal to zero (for Δ small enough) in the two-verdict case.

8.2 A continuous model of information acquisition

To better visualize the difference between two and three verdict systems, the model is modified to allow a continuous amount of search for evidence. the model is conceptually similar: as long as evidence is gathered, a cost is incurred at a constant rate of c. During that 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: a higher value of D means that p evolves faster towards the true state of the world. At some time T, the trade-off is reversed. The evidence formation process is then interrupted and the verdict is chosen based on the posterior p_T at time T, resulting in a social welfare $w(p_T)$.

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

$$0 = \max\{w(p) - v(p); -rv(p) - c + D^2p^2(1-p)^2v''(p)\}, \tag{11}$$

where r is a discount rate capturing the idea that longer trials are penalizing for the all parties involved. The first part of the equation implies that $v(p) \geq w(p)$, which simply means that the value function always exceeds the welfare obtains if one stops 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:

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

As it turns out, all solutions to this equation have a closed form solution when $D^2/r = 3/4$:

$$v(p) = -\frac{c}{r} + \left(A_1 + A_2 \left(p - \frac{1}{2} \right) (1-p)^{-2} \right) p^{-\frac{1}{2}} (1-p)^{\frac{3}{2}}, \tag{12}$$

where A_1 and A_2 are free integration constants. For simplicity, we set $r = 1$ and $D^2 = 3/4$ and vary the search cost c.[13]

The search region and value functions are then 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 2-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 further search has low value. Near w's kink (i.e., the threshold p^* at which the sentence switches), however, there is a high value of information and 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 (the so-called value matching condition) and the derivatives must also match (smooth pasting condition).

This imposes four conditions (2 value matching and 2 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 (12). The result is represented on Figure 3.

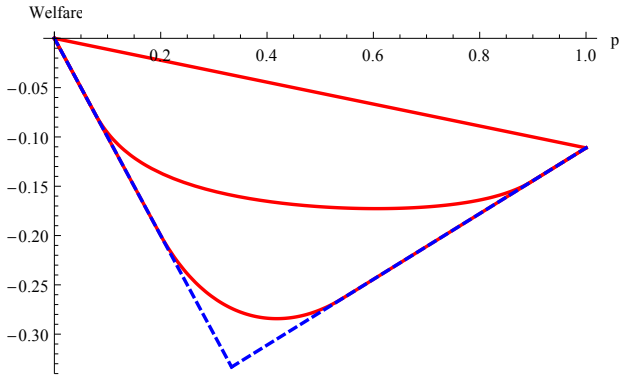

Figure 3: Value function, 2 verdicts, for varying cost levels.

The situation is more interesting in the three verdict case. Around the kink p_2, we still

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

have a two-way smooth connection between w and v, as in the 2-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, either $v(p_1) > \bar{w}$ (low search cost), and the situation looks exactly as in the two verdict case. Intuitively, the search cost is low enough that the intermediary verdict doesn't matter: evidence will be gathered until either the not guilty or the guilty verdict is reached. This is an ideal situation in which the trial technology is quite accurate and a two-verdict system suffices.

For larger search 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 search region now has two parts: when p is below p_1, there is a large incentive to look for evidence, because it can switch the verdict from a zero sentence to s_1, and s_1 was tailored so as to provide a much fairer sentence around p_1 than either 0 or s_2. This also implies that not search on a right-neighborhood of p_1 is optimal. The second search region is around p_2, as before.[14]

Because the first search 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 search 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 (12) for that region. The result is represented on Figure 4.

Importantly, because the welfare w_3 is always higher than the welfare w_2, the value function v_3 in the three-verdict case is also (weakly) higher than the two-verdict value function v_2 – this fact is straightforward to establish. This difference only matters in high enough search costs, 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 search 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 that the right cutoff $\tilde{p}_2$ of the second search region in the

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

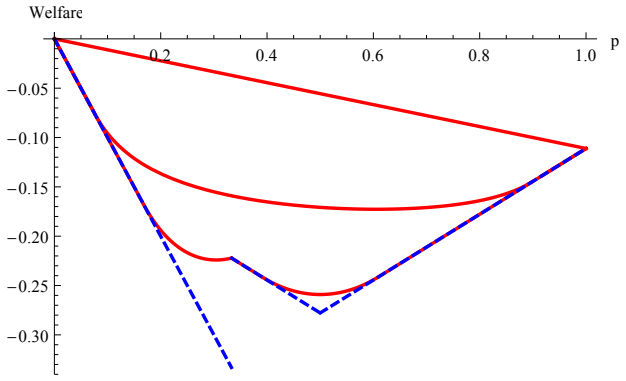

Figure 4: Value function, 3 verdicts, for varying cost levels.

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 search cost. When the trial technology is very accurate, the posterior is unlikely to end up in the middle region and this intermediary verdict has no impact. When finding new evidence is very costly, however, the posterior may end up in the middle region. In that case, the third-verdict system increases the incentive for evidence formation in two regions: before p_1 and around p_2, and decreases the incentive just to the right of p_1. Overall, because $\tilde{p}_0 < \hat{p}_1$ and $\tilde{p}_2 > \hat{p}_2$, the three-verdict system results in searching at more extreme values of beliefs, where in the two-verdict system search has already been given up.

A 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.

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

Proposition 11 *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 supermod-

ular. 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 $W'(s_1^*(p_1,p_2),g) > 0 > W'(s_1^*(p_1,p_2),i)$. ■

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

Proposition 12 *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.

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

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 $\frac{d^2p_2^*(s_1)}{ds_1^2} < 0$.

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

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 $s_1 \in [0, s_1^*]$, $\frac{dp_2^*(s_1)}{ds_1} > 0$. ■

A.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.

B 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

B.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.

B.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.

C Parameters for the welfare functions of Section 8

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

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

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

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

[5] 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.

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

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

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

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