Plea Bargaining when Juror Effort is Costly

Brishti Guha[1]

Abstract

This is the first paper to integrate plea bargaining with costly juror effort. Jurors care about achieving a correct verdict, but experience costs in processing trial-relevant information. There are no fully separating equilibria, where only innocent defendants go to trial, or pooling equilibria, where innocent defendants falsely plead guilty. The first result has been found in literature which does not incorporate costly juror attention, and is thus robust to the inclusion of this phenomenon. The second is new (barring schemes involving post-trial review by external bodies) and shows that laws restricting very lenient plea bargains are unnecessary; costly, unverifiable attention combined with the Cho-Kreps intuitive criterion rules such bargains out in equilibrium, regardless of prosecutor preferences. I characterize feasible semi-separating equilibria that a prosecutor can induce. I also characterize the optimum plea offer for different possible prosecutor preferences. There is a tradeoff between court costs, verdict accuracy and the length of plea sentences. The model generates novel testable implications, and helps to resolve a puzzle noted by legal scholars – that defendants going to trial overwhelmingly opt for jury trials over bench trials, while bench trials, in fact, have a higher rate of acquittal. I perform some robustness checks.

Keywords: Plea bargaining, costly attention, jury, free riding.

JEL Codes: K41. C72, D71, D91.

1. Introduction

A growing literature – theoretical as well as experimental – highlights the issue that jurors, despite being motivated to achieve a correct verdict, may also find it costly to analyze complex information, and may, like other economic agents, dislike effort or suffer from cognitive biases. For instance, in the law and economics literature, Mukhopadhyaya (2003) considers jurors who experience a small cost of paying attention, and shows that this generally results in small jury

[1] Centre for International Trade and Development, School of International Studies, Jawaharlal Nehru University, New Delhi 110067. Email: brishtiguha@gmail.com. I would like to thank an anonymous Associate Editor and two anonymous referees for valuable feedback.

panels making more accurate decisions than larger ones. Guha (2018) models costly juror effort where jurors vote secretly without deliberation, and the verdict is arrived at via a voting rule, unlike in Mukhopadhyaya's model where jurors deliberate. Guha (2020a) highlights how jurors that experience a small cost of effort experience a free riding problem not only with respect to each other's efforts, but also with respect to their collective priors, to the extent that a more informative prior ends up actually worsening juror verdicts over a large parameter range – a range which is larger, the smaller the jurors' effort cost.[2]

In conjunction with this, there is an empirical psychology literature on juror decision-making. Bornstein and Greene (2011) discusses how this literature highlights jurors' cognitive limits. They find that jurors resort to heuristics often, especially if the trial-relevant information is complex, and are subject to biases like the availability bias (which relates to the ease of recalling information), the hindsight bias, and the representativeness bias. It can be taxing to overcome the effects of ingrained psychological biases, providing a reason why jurors might find it costly to pay attention during a trial. In addition, jurors may also display what is sometimes called "confirmation bias", in the sense that if they have extreme priors (strongly skewed towards either guilt or innocence), they ignore ambiguous or contradictory evidence.[3,4]

In this paper, I examine how costly juror effort interacts with another institution which has been much studied in the law and economics literature – plea bargaining. The existing literature on plea bargaining has either modeled it as a game played between a prosecutor and a defendant, or acknowledged the role of judges or jurors, but has never considered how the fact that jurors may experience a small effort cost feeds into plea bargaining. In my model, jurors' priors, were plea bargaining not present, would be based on factors like the efficacy of the police. However, when plea bargaining exists, jurors realize that guilty and innocent defendants may differ in their tendency to accept or reject a plea, and hence, they use Bayes' rule to update their priors about the likelihood that a defendant who comes up for trial is guilty. These updated priors (beliefs) can, in

[2] Guha (2020b) also models costly juror effort, and focuses on the problem of whether it is better (in the interests of a more accurate verdict) to encourage jurors to vote with or without deliberations. She finds that if the jury panel is small enough, barring deliberations might be better.

[3] See https://www.plaintiffmagazine.com/recent-issues/item/the-psychology-of-jurors-decision-making ("The Psychology of Jurors' Decision Making", Plaintiff Magazine, January 2018).

[4] Interestingly, while I do not assume confirmation bias explicitly in my model, it turns out that an implication of allowing jurors to experience an effort cost is that they stop paying attention if priors are very extreme.

turn, drastically change the likelihood that jurors pay attention during a trial, affecting the probability of the verdict being correct. This is, moreover, anticipated by defendants and prosecutors, who know that jurors experience effort costs, and also know that jurors' updated priors can be affected by defendants' decisions to accept or reject plea offers. Thus, juror decision-making and plea bargaining mutually influence each other. Given that the assumption of jurors experiencing a small effort cost is a realistic one, as discussed above, the paper extends the literature by examining the extent to which some previous results in the plea bargaining literature remain robust, as well as by exploring differences in the implications caused by accommodating this new factor.

I show that with costly attention, jurors rely overly on their updated priors to the extent that when these priors are extreme, jurors stop paying attention and make collective decisions on the basis of these updated priors. I find that fully separating equilibria are not possible. If there were full separation at the plea bargaining stage, with only innocent defendants rejecting plea bargains, jurors would always acquit. Given this, the guilty would find it profitable to refuse the plea bargain in the first place.

I also find that there is no pooling equilibrium. Intuitively, this is an outcome of the innocent always being more ready than the guilty to opt for a trial over a given plea offer. This, in turn, is an outcome of jurors being more likely to be correct in their collective verdicts – given that costs of paying attention are small – than wrong. Thus, an innocent defendant opting for trial faces a larger probability of being acquitted than a guilty one. Jurors know that the same plea offer is more likely to be accepted by guilty, rather than innocent defendants. So, if they observe someone rejecting such a plea bargain, they believe the deviation was made by an innocent rather than a guilty defendant, and acquit such a defendant without incurring the costs of attention. But then, this prompts innocent defendants to refuse the plea offer. (Formally, this is shown by using the Cho-Kreps intuitive criterion).

I characterize semi-separating equilibria that the prosecutor can induce. I focus on a mid-range of plea offers that give rise to interior equilibria, where all innocent defendants and some guilty ones opt for trial, with the proportion of guilty defendants being high enough to induce jurors to pay attention at trial with positive probability, but low enough so that the jurors' default option, in case no one ends up paying attention, is acquittal. Focusing on this mid-range enables me to obtain

interesting comparative static results later. Which of these semi-separating equilibria he induces will depend on the prosecutor's preferences (I discuss more than one possible preference function that the prosecutor can have). Note that juror attention is also unverifiable, like effort in many other contexts; there is no external entity that is capable of measuring juror effort or involved in enforcing it at a certain level.

I also show that there is generally a tradeoff between greater verdict accuracy, and greater court costs and harsher plea sentences. A prosecutor might opt for relatively lenient plea sentences which are accepted by a majority of guilty defendants; this would keep court costs low, but would result in a verdict which is not very accurate, in the case of those guilty defendants who do go to trial.

My results speak to theoretical issues in the plea bargaining literature. For instance, some of the literature has been concerned with the possibility that innocent defendants may plead guilty (pooling equilibria where both innocent and guilty defendants plead guilty are modeled in Grossman and Katz (1983), and Bjerk (2021) among others) and has suggested that plea discounts be legally restricted so that plea bargains are not lenient enough to attract innocent defendants (Bar-Gill and Gazal Ayal 2006). I find, in contrast, that such restrictions are unnecessary: even without them, plea bargains sufficiently lenient to attract innocent defendants will not be offered. In a later section, I discuss how my results are affected if I allow for heterogeneity in risk attitudes within each defendant type (guilty or innocent) with risk attitude being uncorrelated with innocence or guilt. My results are preserved, some without any additional assumptions, and others subject to an upper bound on the extent of risk aversion, and to jurors' attention costs not being too high. I also contrast my results there with Mungan and Klick (2016), which suggests a way of avoiding a pooling equilibrium, but relies on a post-trial mechanism, review and subsequent exoneration, to achieve this.

I also find interesting empirically testable implications, such as the prediction that, for a fixed punishment if convicted at trial, a larger jury panel or a more complex trial entails either more lenient plea bargains (higher plea discounts) or smaller rates of take-up of plea offers. A given plea bargain has a higher rate of take-up, the more serious the crime. In a later section, I discuss to what extent the existing empirical literature matches these predictions.

My results also suggest a way of resolving a puzzle noted by legal scholars. Leipold (2005) noted that most defendants who opt to go to trial opt for a jury trial instead of a bench trial. He finds this counter-intuitive, as in fact bench trials have a higher acquittal rate than jury trials. Moreover, Kalven et al (1966) and Eisenberg et al (2005) find that *for the same set of cases* – cases actually resolved through jury trials – judges, when asked if they agreed with the jury's verdicts, were usually inclined to be stricter and to choose to convict more often. I explain, in the discussion section, how my model and results are compatible with all of these findings and help resolve the apparent puzzle. This happens via a selection effect induced by plea bargaining.

Thus, this paper integrates two strands of the literature – that on costly juror effort, and that on plea bargaining. While I have already discussed the former, the papers on plea bargaining that I consider most relevant to my paper include Grossman and Katz (1983), Reinganum (1988), Baker and Mezzetti (2001), Bar-Gill and Gazal Ayal (2006), Bjerk (2007), Kim (2010), Lee (2014), and Bjerk (2021).

Grossman and Katz (1983) departed from Landes (1971)'s earlier presumption that all defendants must be guilty, and derived conditions for a separating and a pooling equilibrium to exist. In their paper, an uninformed prosecutor makes a single offer to all defendants, and the defendant's guilt is private information (my model shares this feature). They, however, do not model the trial stage, so jurors do not feature as players. Reinganum (1988) assumes two-sided asymmetric information; the defendant knows whether she is guilty or innocent, but the prosecutor knows the strength of the defendant's case. She derives equilibria – including fully separating equilibria – under both limited prosecutorial discretion (a prosecutor must make the same offer to all defendants) and under unlimited discretion, and comes to the surprising conclusion that sometimes it is better to restrict prosecutorial discretion. Like Grossman and Katz, she does not model the trial stage either. Baker and Mezzetti (2001) are concerned with the implication of fully separating equilibria; though they do not model judicial decision-making, either, they argue that equilibria where only innocent defendants go to trial may be hard to sustain as prosecutors might then wish such cases to be dismissed. They discuss how prosecutors can make the threat to go to trial credible, by committing to spend resources to gather evidence against defendants who reject plea offers. Bar-Gill and Gazal Ayal (2006) are concerned with the adverse implications of pooling equilibria where plea offers are accepted not only by guilty but also by innocent defendants, and

argue for laws restricting leniency of plea offers. As already mentioned, I show that pooling equilibria with false guilty pleas will not emerge, so it would be unnecessary to use a law to restrict the leniency of plea offers.

Bjerk (2007), unlike the previous papers mentioned, models the jury decision-making stage. In his model, prosecutors receive a signal about defendant guilt in the pre-plea stage. On this basis, prosecutors choose a cutoff signal and only make plea offers to defendants whose signal falls below the cutoff (ie is less indicative of guilt). They send the other defendants to trial. Jurors do not (unlike in my model) face any cost to exerting effort; they receive a signal about defendants who go to trial, and this signal is more precise than the prosecutor's signal. They then use this signal, as well as their knowledge of the prosecutor's cutoff, to assess the likelihood of defendant guilt. They end up choosing a cutoff signal above which to convict. The author derives a continuum of possible equilibria. Some involve a low guilt threshold, so that the prosecutor can only offer the plea bargain to a few defendants. Most end up in trial and have a high probability of conviction, as the jurors also use a low guilt threshold for conviction. Hence, these equilibria also carry harsh plea sentences. There are other equilibria, in which more defendants are offered a plea, the guilt threshold is high, defendants that go to trial face a low conviction probability, and plea sentences are lenient. Bjerk (2021) further refines this setup and also considers pooling equilibria – where both innocent and guilty defendants accept the plea – and discusses issues of social optimality.

Kim (2010) considers the issue of credible plea bargaining, and shows that if semi-separating equilibria exist, then the prosecutor does not need to be able to commit to spend resources to investigate defendant guilt (unlike Baker and Mezzetti 2001). Jury decision-making is not modeled. Lee (2014) does model jury decision-making. However, he uses a different model, where jurors do not face any effort cost, but vote secretly (without deliberation) in accordance with a private signal, and the voting rule – unanimity or super-majority – becomes important. He shows that a prosecutor who is more concerned about wrongful convictions than wrongful acquittals can then use plea bargaining to influence juror decision-making. As in my main model , all agents in his model are risk neutral, and the prosecutor is uninformed about defendant guilt. In his model, a juror assesses if she is pivotal based on whether she expects her vote to change the aggregate decision, given the voting rule: this determines whether the juror votes "sincerely" (according to her signal) in the spirit of Austen-Smith and Banks (1996) or Feddersen and Pesendorfer (1998).

There has been a more recent literature on judicial mechanism design which integrates plea bargaining with juror decision-making. This includes Siegel and Strulovici (forthcoming) who find that plea bargaining emerges endogenously as a feature of optimal mechanism design. Silva (2019) considers a model with a number of possible suspects; each suspect's guilty plea has the externality that it exonerates all other suspects.

The rest of the paper is organized as follows. Section 2 contains the model and results. Starting with setting up a costly juror attention model and deriving some preliminary results in the absence of plea bargaining, I then integrate plea bargaining into the model and derive results relating to possible outcomes which a prosecutor can or cannot feasibly induce. Next, I turn to possible prosecutor preferences to derive further results about which outcomes are actually going to be induced by a prosecutor, and also which outcomes are socially optimal. Thus far in the model, I have implicitly assumed that it is credible for a prosecutor to take any given defendant to trial if the defendant rejects his plea offer. I then explicitly discuss this, showing that for some prosecutor preferences it follows automatically from the rest of the model, while for others, it requires an additional restriction which does not conflict with our other assumptions (at the same time I discuss data which shows that actual dismissal in such cases has become increasingly rare). Finally, I show that nothing changes if we allow jurors the right to veto a plea bargain.

In Section 3, I first discuss the testable implications of the model. I then explain how it resolves the puzzle noted earlier, that defendants overwhelmingly prefer jury to bench trials even though bench trials have a higher acquittal rate. Next, I discuss some robustness issues. Specifically, the main model assumes, in the interests of tractability, that if a juror does pay attention, she receives a perfectly accurate signal about defendant guilt or innocence. Therefore, as a robustness check I discuss the effect of having attentive jurors receive (i) imperfect but perfectly correlated signals (each attentive juror receives exactly the same information, eg the evidence at trial, but this is not perfectly informative), and (ii) imperfect independent signals (traditionally considered in the Condorcet jury theorem literature). I also briefly compare results if jurors do not experience any cost of attention, but do receive imperfect independent signals, to highlight that the results of my main model stem from costly attention rather than from perfect signal accuracy. I then discuss, as a robustness exercise, the effect of allowing heterogeneity in risk attitudes within defendant types. Finally, I elaborate on the implications of the unverifiability of costly juror effort, in contrast to

requirements (imposed in some other work) that jurors must bind themselves to exert effort. Section 4 concludes. The appendix contains additional proofs and a couple of remarks.

2.The Model and Results

A game G is played by risk neutral players. A unit measure of defendants is arrested: a proportion *r* of these are guilty and the rest are innocent. I assume $r>0.5$; police are at least somewhat efficient, so that anyone arrested is more likely to be guilty than innocent. All defendants are then offered a single plea bargain – carrying penalty *P* – by a prosecutor who only knows *r*, but does not know which defendant is guilty. The prosecutor chooses *P*, while in response the defendant chooses a probability λ, $0\leq\lambda\leq1$, of rejecting the plea and proceeding to trial.[5] This probability can vary by defendant type. If the plea is accepted, the game ends. If it is rejected, a panel of *n* jurors hears the evidence. If convicted, a defendant receives an exogenous punishment of *F*, and 0 if he is acquitted.

While in the absence of plea bargaining, the jurors' common prior about a defendant's guilt would have been *r*, it is updated using Bayes' rule in accordance with defendant behavior in the plea bargaining stage.[6] Jurors are symmetric, and each obtains a utility of 1 from a correct verdict. However, each juror may also choose whether to exert costly effort and pay attention during the trial. Effort entails a cost $c<0.5$[7] (typically, the cost of paying attention is small relative to the utility a judge gets from a correct verdict, so *c* will usually be much smaller, e.g 0.1 or 0.05). This cost might represent the complexity of a trial, and is higher for longer and more complex trials. If a juror chooses to pay attention, he incurs *c* but also receives a sure signal about the defendant's guilt or innocence. If he does not, he saves on *c* but does not receive the signal, either. Each juror i chooses his own probability of paying attention, σ_i, the trial occurs, and all jurors deliberate. If even one juror has paid attention, he can ensure the correct verdict by truthfully sharing his signal

[5] Whether it is always credible for the prosecutor to proceed to trial rather than seek dismissal is discussed in Section 2.4, where it is shown that depending on the prosecutor's preferences, credibility may follow automatically, or require a parameter restriction which is not in conflict with the rest of the model. For now we treat it as credible.

[6] Lee (2014) also allows jurors to update their priors based on differing proportions of acceptance of plea bargains among guilty and innocent defendants. In my model, it is easy for jurors to compute these proportions, as long as they know the terms of the plea bargain offered, because, as formally shown later, these proportions are a function of the plea offer. Knowledge of the terms of a plea offer is realistic because in practice the law requires disclosure of the terms of any plea offer in open court, even if the defendant ends up rejecting the plea offer (see https://www.mass.gov/rules-of-criminal-procedure/criminal-procedure-rule-12-pleas-and-plea-agreements).

[7] This threshold is needed for the existence of a symmetric mixed strategy equilibrium in the trial stage.

with the others, who also want a correct verdict and hence vote according to this signal. If it turns out that no one has paid attention, the jurors collectively decide to vote according to their updated priors about defendant guilt or innocence.

Following Guha (2020a) – which contains a costly juror effort model, but no plea bargaining – as well as earlier literature on costly juror models (Mukhopadhyaya 2003, Guha 2018) – I focus on the case where all jurors choose a symmetric mixed strategy, so that $0<\sigma_i = \sigma <1$. Though other equilibria – such as pure strategy ones where only one juror pays attention with probability 1 – exist, they all encounter a coordination problem, as pointed out by Mukhopadhyaya (2003). Since all jurors are alike, and do not communicate before the trial, there is no basis for deciding which of them is to pay attention, and nor are there any grounds to justify asymmetric behavior.

Each juror's expected utility is given by

$$U_J = p - \sigma c \tag{1}$$

where p is the probability (derived endogenously) of the verdict being correct. The probability of a correct verdict is multiplied by 1 – the benefit of a correct verdict – and expected costs of paying attention are deducted.

2.1 A jurors'game for any given belief r'

Next, I replicate some key results based on Guha (2020a) which show how jurors' probability of paying attention, and the probability of a correct verdict, change according to their beliefs about defendant guilt. In that model, priors are not updated, since there is no plea bargaining. However, here we denote jurors' belief about defendant guilt by r' ; this may be different from r as jurors will update their initial priors based on the fact that any defendant coming up for trial must have rejected the plea bargain offered by the prosecutor. We will explore the updating process in detail in subsequent sub-sections. Here, we focus on the jurors' game for a given belief.

First, we examine how σ is determined in such a model. To randomize between paying attention and not doing so, each juror must exactly equate his expected benefit from paying attention to his cost of doing so. He benefits from paying attention only if (i) the other n-1 jurors are not paying attention (if even one were, that juror would have shared their signal

truthfully, ensuring a correct verdict) *and* (ii) if the collective judgment based on the prior, in the event of no one paying attention, were wrong. Thus, the probability of him benefitting from paying attention is $(1-r')(1-\sigma)^{n-1}$, where the first bracketed term is the probability of an uninformed collective guess being wrong (for the case where r'>0.5), and the second is the probability of all the other jurors being inattentive, where the probability of any one juror's being inattentive is 1-σ). This gives us[8]

$$(1-r')(1-\sigma)^{n-1} = c$$

or

$$\sigma = 1 - \left(\frac{c}{1-r'}\right)^{1/(n-1)} \tag{2}$$

From (2), we can see that for positive σ, we require that c<1-*r'* or equivalently, *r'*<1-c. If *r'* becomes too high, the probability of the collective guess of inattentive jurors being wrong is very low, and jurors rely entirely on their beliefs and stop paying attention.

Next, the probability *p* of a correct verdict is one minus the probability that (i) no one pays attention *and* (ii) the collective guess in that event is wrong. In all other cases, the verdict is correct. Thus, we have

$$p = 1 - (1-r')(1-\sigma)^{n} \tag{3}$$

From (2) and (3),

$$p = 1 - c^{n/(n-1)}(1-r')^{-1/(n-1)} \tag{4}$$

Finally, note that equations (2) to (4) apply to *r'≥0.5.* Here, defendants are more likely to be guilty than innocent, so in the event that none of the jurors ends up paying attention, they vote for guilt – a verdict which is wrong if the defendant is innocent (however, this does not happen if even one juror pays attention). In case defendants were more likely to be innocent than guilty, i.e for *r'*<0.5, the error in the case of all jurors being inattentive would involve wrongful

[8] While I have derived σ here using the concept of a symmetric mixed strategy equilibrium, we get the same answer if we use (1) to study the optimization exercise carried out by each juror. This is shown in the appendix.

acquittal rather than wrongful conviction. I consider this possibility, because as we will see, updated priors once plea bargaining is considered may very well be in favor of acquittal. The error of wrongful acquittal happens only if (i) no one was attentive, and (ii) if the defendant was actually guilty. Then, equations (2) to (4) get modified as follows:

$$\sigma = 1 - \left(\frac{c}{r'}\right)^{1/(n-1)} \tag{2'}$$

$$p = 1 - r'(1-\sigma)^n \tag{3'}$$

$$p = 1 - c^{n/(n-1)}(r')^{-1/(n-1)} \tag{4'}$$

Lemma 1. *Let r' be the jurors'belief that a defendant who comes up for trial is guilty. Then*

(i)*If r'<c, σ=0 and jurors automatically acquit defendants, with p=1-r'.*

(ii)*If c≤r'<0.5, σ is given by (2') and we have* $\frac{\partial p}{\partial r'} > 0$

(iii) *If 0.5≤ r'<1-c , σ is given by (2) and we have* $\frac{\partial p}{\partial r'} < 0$ *,that is, the probability of a correct verdict declines in the jurors' beliefs over this range.*

(iv)*If r'≥1-c, σ=0 and jurors automatically convict defendants, with p=r'.*

Proof. (i) From (2'), it is clear that σ cannot be positive if *r'*<c, as the a priori probability of innocence would be too high to justify paying attention, so we would have σ = 0. Substituting this in (3') gives us that *p* = 1-*r'* in this range of *r'*.

(ii) If c≤r'<0.5, we can differentiate (4') with respect to *r'* to get

$$\frac{\partial p}{\partial r'} = \frac{1}{n-1}\left(\frac{c}{r'}\right)^{\frac{n}{n-1}} > 0 \tag{5'}$$

Thus, p increases in r' over this range, and σ is positive and is given by (2').

(iii) If 0.5≤ r'<1-c , we can differentiate (4) with respect to *r'* to get

$$\frac{\partial p}{\partial r'} = -\frac{1}{n-1}\left(\frac{c}{1-r'}\right)^{\frac{n}{n-1}} < 0 \tag{5}$$

Note that the interval [0.5, 1-c] is necessarily non-empty given c<0.5. Thus p is decreasing over this interval.

(iv) If r'≥1-c, jurors stop paying attention. Then, we have $p=r'$ as can be checked by substituting σ = 0 in (3). *QED*

The intuition behind Lemma 1 is as follows. As beliefs become stronger, so do jurors' temptations to take advantage of an informative belief by reducing their probability of paying attention. This can be confirmed by differentiating (2) and (2') with respect to *r'*. We find that $\frac{\partial\sigma}{\partial r'} = -\frac{1}{n-1}c^{\frac{1}{n-1}}(1-r')^{-\frac{n}{n-1}} < 0$ when $0.5<r'<1-c$ (each juror's probability of being attentive shrinks as beliefs in favor of guilt strengthen) while $\frac{\partial\sigma}{\partial r'} = \frac{1}{n-1}c^{\frac{1}{n-1}}(r')^{-\frac{n}{n-1}} > 0$ when $c<r'<0.5$ (each juror pays less attention as *r'* falls in this range, i.e as beliefs in favor of innocence strengthen). When this effect is compounded for all jurors in the panel, the indirect effect of a stronger belief – the reduction in the probability of juror attention – overpowers the direct effect that the belief will be correct more often. However, for very extreme beliefs, jurors stop paying attention altogether, and at this point, an even more extreme belief improves the probability of a correct verdict, since jurors anyway solely vote based on this belief.[9]

Remark 1. *The probability of a correct verdict is weakly greater than 1-c.*

Remark 1 directly follows from Lemma 1; we can see that *p* hits this minimum value, 1-c, at r'=c and at r' = 1-c, and is higher than this for all other values of *r'*.

2.2Integrating the Plea Bargaining Stage

We now examine the plea bargaining stage prior to jury trial, which causes Bayesian jurors to update their priors and accounts for the difference between *r* and *r'*. The updated beliefs would depend not just on *r* but also on the values of λ chosen by guilty as well as innocent defendants, in a manner to be made explicit shortly. Jurors then make their decision about σ, the probability of paying attention, subsequent to the updating process. Defendants and the prosecutor are aware about how jurors update priors in response to choices made during the plea bargaining stage.

[9] Interestingly, this effect of jurors relying overly on informative beliefs would obtain even if each juror had been "socially minded" in the sense of internalizing all the other jurors' benefits. Consider the case where beliefs favor guilt. Then the line before (2) would have changed to $n(1-r')(1-\sigma)^{(n-1)}=c$, yielding $\sigma=1-(c/n(1-r'))^{1/(n-1)}$. It can be easily verified that this does not change the sign of $\partial\sigma/\partial r'$. Similar conclusions can be drawn for the case where beliefs favor innocence.

For the time being, we do not specify prosecutor preferences. We simply look at feasible ranges of prosecutor choice of plea bargains P that could induce subgames with combinations of (λ_G, λ_I, σ) – the first two terms denoting the proportion of guilty and innocent defendants, respectively, that reject plea bargains, and the third term denoting the probability of an individual juror paying attention – that could constitute part of an equilibrium. In a subsequent sub-section, we specify a welfare function, as well as possible prosecutor preferences, and discuss some implications about which of these feasible equilibria a prosecutor is more likely to induce. (In case his preferences tally with the social planner's, his choice will be socially optimal). Thus, we characterize prosecutor's choice of equilibrium P after characterizing equilibrium behavior in subgames induced by various feasible choices. Our equilibrium concept is perfect Bayesian equilibrium with a refinement, the Cho-Kreps intuitive criterion.

Definition: *A perfect Bayesian equilibrium in this game G is a quartet {* λ_G, λ_I, P, σ*} such that given P and σ, neither guilty nor innocent defendants can benefit by changing their probability of plea rejection. Jurors update their priors of defendant guilt using Bayes' rule where possible.*

Jurors' beliefs off the equilibrium path are restricted by the Cho-Kreps intuitive criterion, which specifies that zero weight be placed on the belief that an off-equilibrium deviation has been made by a type that would be worse off after the deviation than in the posited equilibrium.[10]

We are now in a position to derive some results.

Lemma 2. *The game G has no fully separating equilibrium where only innocent defendants reject the plea, while all guilty defendants accept it.*

Proof. Suppose that such a separating equilibrium exists. Then, we have $\lambda_G = 0$. Jurors realize that all defendants who proceed to trial are innocent, and hence they update their priors about defendant guilt to 0. From Lemma 1, $\sigma = 0$ would then be chosen, as it is optimal not to pay

[10] See Cho and Kreps (1987). Suppose a defendant of type j, j=I,G obtains a utility of u*(j) at some posited equilibrium. Then, if jurors observe a defendant deviating from the prescribed equilibrium strategy (say, by sending a message m – for instance, rejecting a plea when they are expected to accept it), then if the maximum possible utility type j derives from such a deviation, given the jurors' beliefs, is $\max_{r'} u(j, m, r')$, then if u*(j)> $\max_{r'} u(j, m, r')$, jurors believe that a type other than j made the deviation.

attention until jurors' assessment of the probability of defendant guilt hits at least *c*. For probabilities below this, jurors automatically acquit everyone. But then, this also implies that guilty defendants have an incentive to reject the plea bargain and come up for trial, as they know that they will be automatically acquitted, while the plea bargain subjects them to a penalty. Thus the candidate equilibrium breaks down. *QED*

Lemma 3. *The guilty are always more ready to accept a given plea bargain than the innocent.*

Proof. For an innocent defendant to accept a plea bargain, the penalty offered by the plea bargain must be weakly less than the innocent defendant's expected punishment from going to trial. Now, an innocent defendant would only be punished at trial if the verdict were wrong. Hence, the condition for an innocent defendant to accept a plea bargain P is

$$[1-p(\lambda_G,\lambda_I,r)]\ F \geq P \tag{6}$$

The condition for a guilty defendant to accept the same plea bargain is

$$p(\lambda_G,\lambda_I,r)F \geq P \tag{7}$$

The difference stems from the fact that a guilty defendant will be punished at trial whenever the verdict is correct.

Note from Remark 1 that p is necessarily greater than 0.5, as $c<0.5$. Thus, the LHS of (7) is greater than the LHS of (6), while the RHSs of both inequalities are the same. Thus, (7) is more likely to hold than (6) is. *QED*

Lemma 4. *There is no pooling equilibrium where both guilty and innocent defendants accept a plea bargain.*

Proof. Suppose such a pooling equilibrium exists, such that defendants of both types accept a given plea bargain P. From Lemma 3, this implies that guilty defendants strictly prefer the plea to going to trial – as even innocent defendants, who are less willing to accept any given plea, are willing to accept this one. Thus, if any defendant deviates and goes to trial, jurors appeal to the Cho-Kreps intuitive criterion and assume that she is innocent rather than guilty. But then, jurors stop paying attention and automatically acquit such a deviant defendant. Given this, an innocent defendant has an incentive to deviate from the pooling equilibrium, going to trial

since sure acquittal is more attractive than any feasible plea bargain that offers a nonzero punishment. Thus, the pooling equilibrium breaks down. *QED*

While Lemma 2 shows that there is no fully separating equilibrium where only the innocent reject the plea bargain, Lemma 3 implies that there cannot be a fully separating equilibrium where only the guilty go to trial, because if the guilty rejected a plea bargain, so would the innocent. Thus these two Lemmas together show that there is no fully separating equilibrium, while Lemma 4 shows that there is no pooling equilibrium where both types accept a given plea.

The fact that there is no separating equilibrium eliminates an issue that arises, for instance, in the models of Grossman and Katz (1983), or Reinganum (1988). Neither of these models jury beliefs, but they have equilibria in which only the innocent go to trial. Subsequently, other authors have pointed out either that prosecutors may not wish to send defendants to trial if they are sure of their innocence (Baker and Mezzetti 2001, Kim 2010). Others, like Bjerk (2007, 2021), also raise the issue, arguing that rational jurors would be so lenient to defendants in a separating equilibrium, strongly believing them to be innocent, that only very lenient plea bargains would be accepted. I show that, in fact, an equilibrium where only innocent defendants go to trial cannot be supported, as jurors would stop paying attention altogether in this hypothetical equilibrium. But that would distort guilty defendants' incentives, causing them to opt for trial, so that this equilibrium would break down.

The fact that there is no pooling equilibrium where all defendants accept the plea also speaks to important issues in the literature – namely the concern that innocent defendants may knowingly plead guilty, accepting a lenient plea bargain (Bar-Gill and Ayal 2006, Lundberg 2018). Some papers (eg Grossman and Katz 1983, Bjerk 2021) do have such pooling equilibria, while others (eg Bar-Gill and Gazal Ayal 2006) explicitly suggest restricting plea discounts so that prosecutors are not permitted to offer pleas lenient enough to attract the innocent. However, in my model, restricting plea discounts is not necessary, as a pooling equilibrium does not obtain.

Thus, in what follows we focus on the possibility of semi-separating equilibria, in which all of the innocent defendants and some of the guilty defendants reject the plea bargain and go to trial, while everyone accepting the plea is guilty.

In such equilibria, we have $\lambda_I = 1$, while $0<\lambda_G = \lambda<1$, where we drop the subscript G on λ_G. Jurors then update their priors according to Bayes' rule. Given a prior probability $r \geq 0.5$ of an arrestee being guilty, the jurors' updated probability estimate of a defendant's guilt is $\frac{\lambda r}{\lambda r+(1-r)} = r'$. Values of σ and *p* may then be obtained by using the appropriate equations. For the rest of the analysis, we assume that the proportion of guilty arrestees is not so high that, in the absence of plea bargaining, jurors would have stopped paying attention. That is, we assume that

A1: *r<1-c*.

Next, we characterize semi-separating equilibria in subgames defined by plea offers in different feasible ranges that the prosecutor may offer.

Proposition 1. *Let A1 hold. Suppose the prosecutor chooses P in the interval* $P \in [(1-c)F, \left(1-2^{\frac{1}{n-1}}c^{\frac{n}{n-1}}\right)F]$. *Then the subgame defined by any P in this interval has a semi-separating equilibrium such that (i)* $\lambda_I = 1$, *(ii)* $\lambda_G = \lambda(P) = \frac{(1-r)c^n}{r\left[\left(1-\frac{P}{F}\right)^{n-1}-c^n\right]}$, *(iii)* $\sigma(\lambda) = 1 - \left(\frac{c\{\lambda r+1-r\}}{\lambda r}\right)^{\frac{1}{n-1}}$ *and (iv)* $p(\lambda) = 1 - c^{\frac{n}{n-1}}\left(\frac{\lambda r+1-r}{\lambda r}\right)^{\frac{1}{n-1}} = \frac{P}{F}$. *Moreover, in the event that no jurors end up paying attention, the default verdict is acquittal.*

Proof. To start with, we will assume that the default option of the jurors, in case no juror pays attention, is acquittal, and later we will prove that this assumption is valid. Now, it is clear that innocent defendants will always opt for trial, as they are acquitted both if at least one juror is attentive, as well as if none of the jurors are attentive. Thus, $\lambda_I = 1$. Next, note that for guilty defendants to randomize between accepting and rejecting a specific plea offer *P*, they must be indifferent between the two options. Thus, any specific choice of *P* by the prosecutor will generate a value of $\lambda_G = \lambda(P) = \lambda$ such that

$$p\big(\lambda(P)\big)F = P \tag{8}$$

Next, note that $\frac{\lambda r}{\lambda r+(1-r)} = r'$ represents the jurors' updated probability of defendant guilt given that all innocent defendants and a proportion λ of guilty defendants come to trial. Moreover, by assumption we have $r' \leq 0.5$ as the default option is acquittal. Thus, we can apply equation (4') to obtain $p(\lambda) = 1 - c^{\frac{n}{n-1}}\left(\frac{\lambda r+1-r}{\lambda r}\right)^{\frac{1}{n-1}}$. Substituting in (8), we get

$$1 - c^{\frac{n}{n-1}}\left(\frac{\lambda r + 1 - r}{\lambda r}\right)^{\frac{1}{n-1}} = \frac{P}{F}$$

Some manipulations yield

$$\lambda(P) = \frac{(1-r)c^n}{r\left[\left(1-\frac{P}{F}\right)^{n-1} - c^n\right]} \tag{9}$$

We will now show that the assumption of acquittal as the default option in case no juror pays attention is a valid one. The statement of the Proposition specifies that

$$P \leq \left(1 - 2^{\frac{1}{n-1}}c^{\frac{n}{n-1}}\right)F$$

Manipulating this inequality, we get

$$c^n \leq \left(1 - \frac{P}{F}\right)^{n-1} - c^n \tag{10}$$

(9) and (10) together imply that

$$\lambda \leq \frac{1-r}{r}$$

Or,

$$\frac{\lambda r}{\lambda r+(1-r)} = r' \leq 0.5 \tag{11}$$

Thus, the original assumption is justified. (9) determines a unique value of λ, the probability of rejecting a plea by a guilty defendant. Note that the statement of the Proposition specifies that $P \geq (1-c)F$, which, from (9), implies that $\lambda \geq \frac{c(1-r)}{r(1-c)}$, which can be rewritten as $\frac{\lambda r}{\lambda r+(1-r)} = r' \geq c$. Moreover, given any choice of *P*, a guilty defendant has no incentive to choose any $\lambda' > \lambda(P)$ (where λ(P) is given by (9)). To see this, note that from Lemma 1, *p* is strictly

increasing in *r'* in the interval specified, since *r'* must lie between c and 0.5. It is also easy to see that *r'*, in turn, is increasing in λ, and thus we have that *p* is increasing in λ in this range. Thus, if a guilty defendant were to choose any λ'>λ(P), this would increase the defendant's expected penalty at trial, making the plea bargain more attractive, so that the probability of rejecting the plea falls back to λ(P). On the other hand, a deviation in the opposite direction is not profitable either. If a guilty defendant were to choose any λ'<λ(P), thus accepting the plea with greater probability, p would fall and so would the expected penalty at trial, increasing the attractiveness of rejecting the plea, so that the probability of plea rejection rises back up to λ(P). Finally, as equations (2') and (4') hold, jurors are choosing σ and *p* optimally given λ – and hence *r'.* Moreover, since *r'*≤0.5, all jurors vote for acquittal in the event that all turn out to be inattentive[11] (which happens with probability $(1-\sigma(\lambda))^n$). ***QED***

Proposition 1's focus on a middle range of feasible plea offers by the prosecutor is connected to the desire to focus on an interior solution, where jurors do pay attention with some probability, and where prosecutors offer plea discounts – the advantage being that it is possible to generate interesting comparative statics in this range. Proposition 1 also highlights that it is possible to use plea bargaining to change r' to, say, 0.5, which would (from Lemma 1) maximize the probability of a correct verdict. Thus, plea bargaining offers a tool to increase the probability of a correct verdict. This is elaborated on in a couple of remarks in the appendix.

We now come to Proposition 2, which deals with comparative statics.

Proposition 2. *Suppose P is in the range considered in Proposition 1. In any semi-separating equilibrium, an increase in (i) the number of jurors n, or (ii) the cost of paying attention, c will result in a larger fraction of guilty defendants going to trial for a given plea offer P.*

Proof. In any semi-separating equilibrium induced by a choice of $P\in[(1-c)F, \left(1-2^{\frac{1}{n-1}}c^{\frac{n}{n-1}}\right)F]$., we have *P=p(λ)F.* Note that $p(\lambda)=1-c^{\frac{n}{n-1}}\left(\frac{\lambda r+1-r}{\lambda r}\right)^{\frac{1}{n-1}}$ and thus, it can be checked that *p* decreases in both c and *n.* For a fixed P, now trial becomes more attractive as the probability of conviction goes down, resulting in a greater likelihood of guilty defendants

[11] Here, I assume that at the point where r'=0.5, that is, jurors' updated priors are not biased in any direction, they vote for acquittal if they all turn out to be inattentive. Interestingly, the practice in ancient Greece was to acquit a defendant for whom votes were exactly equally split.

rejecting the offered plea bargain, until λ increases to the point where the probability of conviction goes up by enough to reduce the expected attractiveness of trial to that of the plea bargain. ***QED***

Proposition 2 implies that if n or c rise, and if the prosecutor wishes to maintain the rate of plea bargain take-up at its old rate, he has to offer a more lenient plea bargain.

Intuitively, a higher cost of juror attention increases jurors' temptation to free ride. A high number of jurors on the panel also increases each individual juror's tendency to free ride; he thinks it likely that at least one of the numerous other jurors would have paid attention, guaranteeing a correct verdict. Thus, given a fixed sentence if convicted at trial, guilty defendants facing a plea bargain expect a relatively small penalty from trial because they face a relatively small probability of conviction. This, in turn, implies that the plea bargain that makes them indifferent between accepting it and going to trial needs to be more lenient.

Remark 2. *An increase in F (the sentence if convicted at trial) results in a higher take-up of a given plea bargain.*

Remark 2 is intuitive because of the need to equalize expected penalties between trial and a plea offer for guilty defendants in a semi-separating equilibrium. A rise in F leads to greater acceptance of any plea offer, to the point that the resulting fall in λ reduces jurors' probability of paying attention at trial to the point where the probability of conviction falls enough to offset the rise in F.

2.3 Welfare Properties and other related results

Suppose a social planner would like to minimize welfare loss. For convenience, we assume that the exogenous penalty, F, is already set at the level that the planner would consider socially optimal for the crime; thus, if a guilty defendant is convicted, he is given the socially optimal punishment. Social loss then stems from three sources (i) court costs, (ii) losses from wrongly convicting innocent defendants, and (iii) losses from not punishing guilty defendants enough (either due to errors by jurors, or due to their accepting plea bargains which hand out smaller punishments).

More precisely, let the loss function *in the absence* of plea bargaining be

$$L_0 = 1 + [1 - p(r)]F \tag{12}$$

Here, $p(r)$ is given by (4). The first term in the loss function normalizes court costs per defendant to 1, and allows for the fact that the entire unit mass of defendants goes to trial, as there is no plea bargaining. The second term shows that a number 1-$p(r)$ of this unit mass of defendants is judged wrongly; the wrongly judged defendants are all innocent (given r>0.5, the default verdict for inattentive jurors is conviction) and this term captures the loss from mistakenly convicting innocent defendants. I assume that the loss from mistakenly punishing an innocent defendant is simply equal to the extent of the penalty. This is tractable, however, the results below would be robust to a positive monotonic transformation. Moreover, F should be more properly thought of as the ratio of the socially optimal penalty to the court costs (normalized to 1).

The loss function *with* plea bargaining, for the class of semi-separating equilibria discussed in Proposition 1 is

$$L_1 = (1 - r + \lambda r) + (1 - p(\lambda))F \tag{13}$$

The last term in the loss function above represents losses both on those guilty defendants who accept a plea (noting that since $P(\lambda)=p(\lambda)F$, the difference between the optimal punishment and the punishment in a plea bargain is $F-P(\lambda) = (1-p(\lambda))F$) *and* losses on those defendants who opt for trial but receive a wrong judgment. The first term represents court costs which are incurred for defendants who opt for trial.

Note that a prosecutor who wants to minimize social losses will not want to offer a plea bargain harsher than the harshest plea bargain considered in Proposition 1. If he did so, it would escalate court costs by prompting even more guilty defendants to opt for trial. Secondly, it would increase r' above 0.5, which, as shown in Lemma 1, would actually reduce the likelihood of correct verdicts, and would lead to a larger proportion of wrong convictions (as the default option for jurors in this range would change to conviction, if they failed to pay attention).

Observation 1 characterizes choices of P made by the prosecutor under three different feasible preferences – the first, where the prosecutor shares the social planner's objectives, the second where he does not internalize court costs, but wants to maximize the probability of a correct verdict, and the third, where he only cares about court costs.

Observation 1. (i) A prosecutor who shares the social planner's preferences optimally offers P* such that $P^* = p\big(\lambda(P^*)\big)F$, where λ(P) is given by (9), λ satisfies the first order condition $r = \frac{\partial p}{\partial \lambda}F$.

(ii) *A prosecutor who wants to maximize p, wants to induce* $\lambda=\frac{1-r}{r}$*, and hence chooses* $P=\left[1 - c^{\frac{n}{n-1}}(2)^{\frac{1}{n-1}}\right]F$.

(iii)*A prosecutor who wants to minimize the number of people going to trial wants to induce* $\lambda=\frac{c(1-r)}{r(1-c)}$*, and hence chooses P=[1-c]F.*

Proof. (i) Differentiating (13) with respect to λ, and simplifying, we get the first order condition

$$r = \frac{\partial p}{\partial \lambda}F \tag{14}$$

The result follows.

(ii) From Lemma 1, it is clear that *p* is maximized at *r'*=0.5, which only happens at $\lambda=\frac{1-r}{r}$. Thus, this is preferred by a prosecutor who wants to maximize the probability of a correct juror verdict. This can be induced by setting $P = \left[1 - c^{\frac{n}{n-1}}(2)^{\frac{1}{n-1}}\right]F$, as the term in square brackets is the probability of a correct verdict when jurors are aware that guilty and innocent defendants are represented equally among the defendants that come up for trial (we get the square bracketed expression by substituting 0.5 into the expression for *r* in (4')).

(iii)The number of defendants who reject the plea and go for trial is 1-*r*+λ*r*, which is increasing in λ. Thus, a prosecutor who wants to keep this number low – attracting more guilty defendants towards a plea bargain – will try to induce the value of λ which constitutes the lower limit of the class of semi-separating equilibria, that is, $\lambda=\frac{c(1-r)}{r(1-c)}$. As *p* at this value reaches its lower bound of 1-c, the prosecutor can induce this equilibrium by offering *P=[1-c]F. QED*

Parts (ii) and (iii) highlight the tension between maximizing verdict accuracy and minimizing court costs.

Example 1. Consider r = 0.7, c = 0.1, n = 3, and F = 12. For these parameters, a prosecutor who shares the social planner's preferences induces λ=0.4, which results in p = .9545. He does this by offering P = .9545(12) = 11.454. On the other hand, a prosecutor whose preferences are described by (iii) will offer P = .9(12) = 10.8, inducing λ = 1/21, and p = .9. A prosecutor whose preferences are described by (ii) will offer P = .9553(12) = 11.4636, inducing λ = 3/7, and p at its maximum feasible value for these parameters, .9553.

2.4 On the credibility of going to trial

So far, we have assumed that it is always credible for a prosecutor to threaten to go to trial, rather than dismiss a case, if a plea bargain is rejected. Interestingly, Leipold (2005) mentions that partly due to changes in the law which make it difficult to dismiss cases, the percentage of pretrial dismissals has been falling over the years; it fell to 9% in 2002 from 15% in 1987. The percentage of criminal cases dismissed was about 8% in 2021[12]. We now examine what conditions, if any, are needed to ensure that not dismissing cases is credible for the prosecutor.

First, consider a prosecutor who does not internalize court costs, but cares about wrongful convictions and wrongful acquittals. It is easy to show that no additional assumptions are needed for such a prosecutor to have a credible threat not to dismiss a rejected plea offer; he will always proceed to trial. To see this, note that such a prosecutor will offer a plea offer P that induces a semi-separating equilibrium in the category of Proposition 1. If this is rejected by a defendant, the prosecutor assigns a probability $\frac{\lambda r}{\lambda r+(1-r)} = r'$ to the defendant being guilty, where λ=λ(P) is given by (9). Then, the prosecutor's expected disutility from dismissing this case is $\frac{\lambda r}{\lambda r+(1-r)} F$ – his expected disutility that a guilty defendant escapes punishment. On the other hand, if he proceeds to trial, he expects to incur loss if the defendant is guilty *but is wrongly acquitted* (which happens with probability 1-p(λ)) and this expected disutility is $\frac{\lambda r}{\lambda r+(1-r)}(1 - p(\lambda))F$ (note that wrongful convictions do not happen in this range, all innocent defendants are acquitted even if jurors end up not paying attention). Thus, clearly the

[12] See https://www.aerlawgroup.com/blog/these-are-the-easiest-ways-to-get-your-criminal-case-dismissed

prosecutor will find it better to go to trial rather than dismiss, and no additional assumptions are needed to render this threat credible.

Next, consider a prosecutor who does internalize court costs, and shares the social planner's preferences. Then, we do need additional parameter restrictions to bolster the threat of proceeding to trial on rejection of a plea offer. Taking court costs into consideration, we require

$$\frac{\lambda r}{\lambda r + (1-r)} F > 1 + \frac{\lambda r}{\lambda r + (1-r)} \big(1 - p(\lambda)\big) F$$

Or

$$p(\lambda) F > 1 + \tfrac{1-r}{\lambda r} \tag{15}$$

Note that the LHS of (15) is increasing in λ over the relevant range (of semi-separating equilibria described in Proposition 1), while the RHS is decreasing in λ. Therefore, a sufficient condition for (15) to hold is if it holds at the lowest limit of λ in the permissible range, that is, for $\lambda = \frac{c(1-r)}{r(1-c)}$. Noting that for this value of λ, we have p = 1-c, and simplifying, this gives us

A2. $\boldsymbol{c(1-c)F > 1}$

Thus, subject to A2, a prosecutor who shares the social planner's preferences prefers to go to trial rather than dismiss a case, if his plea offer is rejected, rendering his threat credible. Essentially, we need the social cost of punishing the guilty to be sufficiently large relative to the cost of a trial. Note that A2 is satisfied by Example 1.

2.5 Juror discretion in fixing plea bargains

In the analysis so far, the prosecutor decides on *P*. Initially, we discussed the set of feasible equilibria that a prosecutor could conceivably induce in the game, and then we briefly discussed the welfare properties of some such equilibria.

What if jurors had the discretion to determine *P*? It turns out that if the jury, or one of their representatives, were in charge of determining *P*, they would be indifferent among the entire set of feasible equilibria, and thus would have no incentive to zone in on a specific *P*, as they do not wish to induce a specific λ, or *r'*.

Proposition 3. *Jurors are indifferent among all possible equilibria of the game G. Their utility in all of them is equal to 1-c.*

Proof. A juror's utility is given by (1), equal to p-σc. Substituting for p and σ in this expression, using either (2) and (4) or (2') and (4'), and solving, we get 1-c (which is also the utility that would have been obtained by a juror who paid attention with probability 1). This utility is independent of r and thus, of r' and λ. A juror's utility does not depend on what proportion of guilty defendants reject plea bargains, and hence, jurors do not care about plea sentences either. *QED*

3 Discussion

The results speak to some theoretical issues in the literature, and also have some empirically testable implications. They also provide a way of resolving a puzzle noted by some legal scholars. I discuss all these in turn, starting with the empirically testable implications.

Proposition 2 shows that for a given sentence on conviction at trial, high costs of juror attention or a large jury panel would result either in a lower take-up of plea bargains, or in more generous plea discounts. This would be supported if for crimes in the same category, plea discounts were positively correlated to jury size (e.g larger for a 12-member jury than a 6-member one, or larger for a jury trial than a bench trial with 1, 2 or 3 judges). Although a large literature in criminology looks at the correlates of plea discounts, and shows that there is substantial variation in plea discounts across jurisdictions for the same type of crime, the issue of correlation with the size of the panel has not, to my knowledge, been studied there yet. I am also unaware of any work testing whether plea bargaining take-up rates are negatively related to the size of the jury panel.

Interestingly, there is some empirical support for the other part of Proposition 2, that an increase in c is associated with more lenient plea bargains. Smith (1986) uses a large dataset on prosecutions and plea bargains and finds that the major beneficiaries of plea bargaining – those who obtain the biggest plea discounts – are those defendants who have very few or no eyewitnesses, or those charged with relatively less serious crimes. Now, both of these factors can be interpreted in terms of proxies for the cost of jurors' paying attention. For a case with a

lot of eyewitness testimony, jurors already have so much obvious evidence confronting them, that their own cost of analyzing the case is lower than for a case with very little or no eyewitness testimony, which would entail higher *c*. Similarly, for very serious crimes, jurors might find it less onerous psychically to pay attention, as it may seem more crucial to catch, say, a murderer as compared to a petty thief. Then, both less serious crimes, as well as crimes with no eyewitnesses, would be associated with higher *c*, and it is for these crimes that the defendants benefit the most in terms of plea discounts. Thus, this interpretation is consistent with my prediction. Interestingly, crimes that were the easiest to prove (low c) had the highest rates of take-up of plea bargains (Subramanian et. al 2020), which is in line with the predictions of Proposition 2. According to the same source, increases in penalties, F (such as the reinstatement of the death penalty for murder charges) also resulted in greater rates of take-up of plea offers (by 20-25%), which is in line with Remark 2.

We may also discuss the model in view of the findings of Leipold (2005) who points out a puzzle. Leipold discussed how, over time, the rate of conviction in bench trials had become much lower than the rate of conviction in jury trials; but counter-intuitively, most defendants preferred to be tried by jury. While the set of cases facing bench trials differs from the set of cases facing jury trials in Leipold's paper, Kalven et. al (1966), and Eisenberg et. al (2005) (who partially replicate their results) find that for *the same* cases, judges were much less inclined to acquit than were jurors (these cases actually came up for jury trial, but judges were asked if they agreed with the jurors' opinions). Interestingly, my model is compatible both with Kalven et al's findings, as well as Leipold's, and helps explain the apparent puzzle through a selection effect induced by plea bargaining. I explain this below.

As Proposition 2 points out, the probability of a correct verdict decreases with the size of the panel. Now, if plea bargaining were absent or insignificant, incorrect verdicts would overwhelmingly comprise of wrongful convictions, as the default for inattentive jurors would be to convict, according to their priors. In a bench trial, with only one decision-maker, wrongful convictions would not occur in the model, because the judge would pay attention as he was sure to be pivotal. Therefore, if plea bargaining were negligible or absent, innocent defendants would actually prefer a bench trial, while guilty defendants would be indifferent between bench trials and jury trials. However, the situation would change dramatically with the increased

importance of plea bargaining. *With* plea bargaining, incorrect juror verdicts would comprise of wrongful acquittals. Thus, guilty defendants would now overwhelmingly opt for jury trials, while innocent defendants would be indifferent between jury trials and bench trials as they would be acquitted in both. Now, as documented by Johnson (2019), plea bargaining has become increasingly dominant over time. Thus, over time guilty defendants would start overwhelmingly opting for jury trials, rejecting bench trials, while innocent defendants may still opt for a bench trial. Assuming $r>0.5$ (that more defendants are guilty than innocent), therefore, the proportion of guilty defendants among those opting for jury trials would go up, while the proportion of innocent defendants among those opting for bench trials would also go up – as the guilty defendants would opt out of bench trials. Hence, this would explain Leipold's findings that the actual rate of acquittal in bench trials increased, while jury trials actually had a higher proportion of convictions than bench trials. At the same time, it would be consistent with the findings of Kalven et al and Eisenberg et al, because with plea bargaining, a jury is more likely to make a wrongful acquittal than a judge. It also explains Leipold's other finding about how defendants overwhelmingly opt for jury trials, but resolves the apparent puzzle he noted.

The model also has implications for various theoretical issues. As mentioned earlier, it has been suggested that restricting the size of plea discounts is necessary so that prosecutors do not offer pleas that are so lenient that they are accepted by innocent defendants (Bar-Gill and Gazal Ayal 2006). However, I find that such laws are not needed. Prosecutors will not in any case be able to offer plea bargains that are lenient enough to attract innocent defendants. This stems from Lemma 4, which shows that there is no pooling equilibrium where both innocent and guilty defendants accept a plea bargain.

I briefly discuss the implications of having a very efficient police – a high r – for my model. A high r implies that $\left(\frac{1-r}{r}\right)^2$ is low, so from Remark 3 part (b) (please see the Appendix), there is more scope for plea bargaining to improve p, the probability of a correct verdict; it now improves even if the proportion of guilty defendants rejecting a plea is not too high (thus, verdict accuracy improves even if a lot of guilty defendants accept the plea bargain). Alternatively, this implies that even for plea bargains carrying relatively short sentences (large

plea discounts) it is feasible that the resulting equilibrium involves greater verdict accuracy, than would have been obtained in the absence of any plea bargaining.

It may be interesting to compare this with Bjerk (2007). In his model, prosecutors observe a signal of defendant guilt, while jurors later (if the case goes to trial) observe a slightly different but more accurate signal. In his paper, prosecutors choose a guilt threshold: if they observe initial signals above this, they do not offer the defendant a plea bargain at all; otherwise they offer the defendant a plea bargain. There is a continuum of such equilibria, but if the threshold chosen is low, then the plea is offered only to a few defendants, and those that go to trial are more likely to be guilty; thus jurors use a low threshold for conviction as well. This low threshold implies a high probability of conviction, which then ensures that the plea bargain offers harsh sentences (conversely, a high guilt threshold corresponds to lenient plea bargains offered to many). He finds that better policing lowers jurors' threshold for being willing to convict, and thus raises the probability of conviction, and hence, the sentence under plea bargaining. In contrast, I find that improvements in policing may actually result in even lenient plea bargains having a beneficial impact on juror decision-making accuracy.

Bjerk (2021) considers how a prosecutor may implement a pooling equilibrium more readily when policing is efficient, reasoning that a relatively small proportion of defendants who accept the pooling plea offer will be innocent. However, in my model, the use of the Cho-Kreps intuitive criterion rules out pooling equilibria irrespective of policing efficiency.

3.1 Robustness

In the model presented here, I assume that if a juror is attentive, she receives perfectly accurate information about the defendant's guilt or innocence. We might want to see if the results are robust to this information being wrong with a small probability.

3.1.1 An imperfect but perfectly correlated signal

Guha (2020a) allows attentive jurors to receive perfectly correlated signals (so that each attentive juror receives the same information) but these signals are not perfectly accurate. Signal accuracy is $q<1$, but the signal is still precise enough so that $q>c+0.5$. The paper then shows that, for the case of $r>0.5$, we have

$$\sigma = 1 - \left(\frac{c}{q-r}\right)^{\frac{1}{n-1}}$$

This would be replaced by $\sigma = 1 - (\frac{c}{q-(1-r)})^{1/(n-1)}$ in the case of r<0.5. Now, in case we had r>q-c (in the case of r>0.5), or in case we had r<1-q+c (in the case of r<0.5), jurors choose σ = 0 because the probability of an incorrect collective decision when no one pays attention is too small to warrant incurring the cost of paying attention. That paper also shows that Lemma 1 holds in the sense that the probability of a correct verdict (now given by $p = q - c^{\frac{n}{n-1}}(q-r)^{-1/(n-1)}$ for *r>0.5* and $p = q - c^{\frac{n}{n-1}}(q-(1-r))^{-1/(n-1)}$ for *r<0.5*) decreases as priors become more extreme. The parallel to Remark 1 is that the minimum possible value of *p* is now *q-c* (which by assumption is more than 0.5). Then, we can check that Lemma 3 holds (the innocent always prefer trial relative to the guilty for a given plea offer), and using this and the Cho-Kreps criterion, Lemma 4 (absence of a pooling equilibrium) also holds. Lemma 2 (absence of a fully separating equilibrium) also holds; we may argue as follows. Suppose a fully separating equilibrium does exist, where a plea offer is only accepted by all innocent defendants, while all guilty ones reject it. Then, jurors assign probability 0 to a defendant's guilt if he comes up for trial. Accordingly, they are inattentive, as 0<1-q+c, which is the minimum probability of guilt at which jurors would start paying attention), and automatically acquit such defendants, creating an incentive for guilty defendants to deviate. The existence of a non-empty range where no jurors pay attention means that signal accuracy does not come into play (only attentive jurors receive signals anyway).

The above indicates that if any equilibria exist, they must be semi-separating. The range of plea offers inducing semi-separating equilibria in categories corresponding to Proposition 1 will change, but the qualitative results are similar. The change comes from the fact that though jurors acquit by default, wrongful convictions become possible if jurors pay attention but receive a wrong signal. The range of plea offers inducing this class of equilibria can be shown to change to $P \in [(q-c)F, \left(q - c^{\frac{n}{n-1}}(q-0.5)^{-1/(n-1)}\right)F]$. The change in limits reflects the change in the probabilities of conviction with the imperfect signal. For any P in this interval, the proportion of guilty defendants rejecting the offer is now given by $\lambda(P) = \frac{(1-r)[1-q+c^n(q-P/F)^{1-n}]}{r[q-c^n(q-P/F)^{1-n}]}$. We may verify that this proportion reduces to (9) for q=1. We may

also verify that given the assumption that q>c+0.5, the innocent always have an incentive to reject plea offers in this interval, even though they have a small probability of being wrongly convicted at trial (this would happen if jurors paid attention but received a wrong signal). Finally, the proportion of guilty defendants going to trial is small enough in this interval such that jurors choose to acquit by default if no one ends up paying attention.

Finally, we may note that Proposition 2 also holds, as the comparative statics properties of p with respect to n and c remain the same. Thus, an increase in the cost of jurors' paying attention, or an increase in the number of jurors, will both lower the rate of acceptance of a given plea bargain. Equivalently it may result in more lenient plea offers.

3.1.2 An imperfect uncorrelated signal

Guha (2022) describes how jurors in a costly attention setup decide on the probability with which to pay attention when attentive jurors may receive independent signals, each of precision $q<1$. In this setup, if different attentive jurors receive different signals, deliberation results in the jurors collectively voting in accordance with the majority of signals received. If no one has paid attention, everyone votes according to the prior, which is correct with probability r (the paper considers r = 0.5 but the reasoning is the same for other r). I replicate the expression for σ below (for the r>0.5 case); it solves

$$[q-r]\left[\sum_{j=1}^{\frac{n-1}{2}}(\sigma q)^j\big(\sigma(1-q)\big)^j(1-\sigma)^{n-1-2j}\frac{(n-1)!}{(n-1-2j)!\,j!\,j!}+(1-\sigma)^{n-1}\right]=c$$

The nth juror is pivotal either if (i) none of the other jurors pays attention, or (ii) if, out of the other attentive jurors, equal numbers have received opposing signals, in which case there is a tie. In both these events, if the nth juror sleeps, the others will collectively pick one of the outcomes, which will be correct with probability r, while the pivotal juror can ensure a verdict which is correct with probability q if he does pay attention. We get the above expression by equating the expected benefit of paying attention to the cost of doing so. On the LHS, q-r would change to q-(1-r) in the r<0.5 case.

As the expression in square brackets represents a probability of being pivotal, it cannot exceed 1. Thus, for jurors to pay attention at all we must have q-r>c or r<q-c (in the r>0.5

case), or q-(1-r)>c or r>1-q+c (in the r<0.5 case). Thus, just as in the correlated signal case, we will get ranges (r<1-q+c<0.5 and r>q-c>0.5) where none of the jurors are attentive. By the same logic as in the imperfect correlated signal case, we can thus argue that the fully separating equilibrium does not exist; jurors stop paying attention, and automatically acquit defendants opting for trial, prompting guilty defendants to deviate. This, moreover, does not involve the possibility of wrongful convictions (just as in the text) because jurors simply acquit without paying attention to signals. Thus signal accuracy does not come into play in this range. Moreover, we see from the expression above that an increase in r (for the r>0.5 case) is similar in effect to a rise in c, which, intuitively, decreases the probability of paying attention (and the probability of a correct verdict). The probability of a correct verdict is given by

$$\sum_{m=1}^{n}\sum_{k=\left\lfloor 1+\frac{m}{2}\right\rfloor}^{m}\frac{n!}{(n-m)!\,k!\,(m-k)!}(1-\sigma)^{n-m}(\sigma q)^{k}\big(\sigma(1-q)\big)^{m-k}$$
$$+\,r\left[\sum_{j=1}^{\frac{n-1}{2}}(\sigma q)^{j}\big(\sigma(1-q)\big)^{j}(1-\sigma)^{n-2j}\frac{n!}{(n-2j)!\,j!\,j!}+(1-\sigma)^{n}\right]=p$$

The first term on the LHS captures the probability of a correct verdict in the event of no ties. The verdict is correct if a majority of the attentive jurors receives the correct signal (when deliberation happens, attentive jurors share signals with the entire panel, and the decision is made according to the signal that the majority of attentive jurors have received). The second term captures ties, which lead to the correct verdict with only probability r (this is for the case r>0.5, and is replaced by 1-r if r<0.5). Ties occur either if an even number of jurors is attentive and equal numbers of attentive jurors receive contradictory signals, or if no one is attentive (then an outcome is picked by consensus). As max[r,1-r] exceeds 0.5, clearly the verdict is more likely to be correct than wrong even where jurors reach ties, or do not pay attention. Wherever they do not reach ties, a majority of attentive jurors are still more likely to be correct than wrong provided signal accuracy is greater than 0.5 (by assumption we have q>c+0.5 so this is always satisfied). Thus, p>0.5 and we can use this to show that Lemma 3 holds (the innocent prefer trial compared to the guilty) and can combine this with the Cho-Kreps intuitive criterion to show that Lemma 4 also holds (there is no pooling equilibrium).

While a full characterization of semi-separating equilibria under the assumption of uncorrelated imperfect signals lies outside the scope of the current paper, we may broadly use techniques similar to the perfectly correlated imperfect signal case to characterize such equilibria. However, comparative statics of jury size on p differ from the previous cases; we see from the table below (column 2) that p may first decrease in n, and later increase for even larger jury sizes. This is because, at larger panel sizes, the negative effect of greater free riding in larger panels is more than offset by the positive informational effects of pooling uncorrelated signals received by different attentive jurors. Despite this non-monotonicity, small panels are generally more accurate than large ones.

Finally, the table also illustrates another point. It contrasts the imperfect signal case *where jurors face no costs of attention* (column 3) with (i) the perfect signal case where jurors face a cost of attention (column 1), and (ii) the imperfect signal case where jurors face a cost of attention (column 2). I only look at odd panel sizes to rule out complications in computation due to tie breakers. I look at column 3 because one might speculate as to whether the main model's results are due to the signal being perfect, rather than to the costly attention feature. Note that, if jurors never face any cost of attention (c=0), they always pay attention, and the only factor which can lower verdict accuracy below 1 is then the imprecision of the signal. We may verify that p under these conditions would be

$$p(n) = \sum_{k=\frac{1+n}{2}}^{n} \frac{(n)!}{k!\,(n-k)!} q^k (1-q)^{n-k}$$

Table 1: Verdict accuracy for different jury panel sizes, r=0.5

n	Perfect signal(q=1), c = .02	q = .99, c = .02	q=.99, c = 0 (no costly attention)
3	.996	.9934	.9997
5	.9911	.9895	.9999
7	.9883	.9873	~1
9	.9866	.9856	~1
11	.9855	.9867	~1
13	.9847	.9872	~1
15	.9841	.988	~1

We see that, if we had an imperfect, uncorrelated but highly precise signal, and no costly attention problem, bigger panels would always reach better verdicts than smaller ones, and so, Proposition 2 would not hold. However, with a costly attention problem, even if the signal is imperfect, small panels generally make better verdicts than very large ones, despite the non-monotonicity (which is unlike the perfect signal case), and therefore, we may still find that when panels are very large, the rate of acceptance of a plea bargain is less than that with a much smaller panel.

Finally, we may note that as long as the jury panel is big enough so that p is smaller than what it would have been with a single judge (in the table above, for the imperfect signal case with costly attention, this would require n to be 5 or higher), then our previous conclusions about why plea bargaining has caused guilty defendants to prefer a jury trial to a bench trial, while bench trials have higher acquittal rates, would still hold.

3.1.3 Within-type heterogeneity in risk attitudes of defendants

In our model, all defendants, whether guilty or innocent, are risk neutral. While this assumption has been made in some previous work on plea bargaining (for example, Kim 2010 and Lee 2014), we might be curious about the consequences of allowing defendants to differ in their attitudes to risk. This is a within-type heterogeneity; risk attitudes vary within the sub-group of innocent defendants, as well as within the sub-group of guilty defendants; moreover, guilt or innocence are uncorrelated with risk attitude, so that a specific degree of risk aversion may be displayed by either an innocent defendant, or a guilty one. Note that this differs from across-type heterogeneity in risk attitudes (for instance, considered in Grossman and Katz (1983) where innocent defendants are more risk-averse than guilty ones). Within-type heterogeneity in risk attitudes is considered in Mungan and Klick (2016) (which we will briefly compare with our approach later on).

To derive concrete results, we assume that any defendant i draws a parameter β_i from a distribution with cdf $G(\beta_i)$, $\beta_i \in [1,\bar{\beta}]$, such that the disutility that she gets from a punishment F take the form $D_i(F) = F^{\beta_i}$. The distribution is common for guilty and innocent defendants. A draw of $\beta_i=1$ corresponds to risk neutrality, while higher values correspond to different degrees of risk aversion. Some of the results will hold even if there is no finite upper bound on risk aversion (that

is, β_i may be ∞), but as others (in particular the non-existence of a pooling equilibrium) require an upper bound, we assume a finite upper bound $\bar{\beta}$.

First, we will look at the feasibility of a semi-separating equilibrium, similar in spirit to Proposition 1. However, while in Proposition 1 all guilty defendants played the same (mixed) strategy, here guilty defendants may either accept or reject the offered plea bargain, depending on their degree of risk aversion. Nonetheless, a semi-separating equilibrium where all innocent defendants and some guilty ones go to trial can be supported.

Proposition 4. *Suppose that the prosecutor offers* $P\in [(1-c)^{1/G^{-1}\left(\frac{c(1-r)}{r(1-c)}\right)}F, \left(1-2^{\frac{1}{n-1}}c^{\frac{n}{n-1}}\right)^{1/G^{-1}\left(\frac{1-r}{r}\right)}F]$. *Define* β^* *as the solution to*

$$p(G(\beta^*)) = 1 - c^{\frac{n}{n-1}}\left(\frac{G(\beta^*)r+1-r}{G(\beta^*)r}\right)^{\frac{1}{n-1}} = \left(\frac{P}{F}\right)^{\beta^*} \tag{16}$$

Then a semi-separating equilibrium exists with (i) $\lambda_I = 1$, *(ii) a proportion* $G(\beta^*)$ *of guilty defendants reject the plea bargain, while the rest (those with* $\beta_i > \beta^*$*) accept it, (iii)* $\sigma\left(G(\beta^*)\right) = 1 - \left(\frac{c\{G(\beta^*)r+1-r\}}{G(\beta^*)r}\right)^{\frac{1}{n-1}}$ *and (iv)* $p\left(G(\beta^*)\right) = 1 - c^{\frac{n}{n-1}}\left(\frac{G(\beta^*)r+1-r}{G(\beta^*)r}\right)^{\frac{1}{n-1}}$.

Proof. First note that if all innocent defendants opt for trial, and if G(β*) proportion of guilty defendants opt for trial[13], then jurors update their priors of defendant guilt to r'= $\frac{G(\beta^*)r}{G(\beta^*)r+1-r}$. Now, note that $\left(\frac{P}{F}\right)^{\beta}$ is monotonically decreasing in β, given P<F, and thus the RHS of (16) is decreasing in β*. As for the LHS, we know from Remark 1 that p has a minimum value of 1-c, while its maximum value $\left(1-2^{\frac{1}{n-1}}c^{\frac{n}{n-1}}\right)$ obtains when jurors assign equal weights to guilt and innocence (requiring G(β) = (1-r)/r or $\beta = G^{-1}\left(\frac{1-r}{r}\right)$ in the LHS of (16) above). As $P<\left(1-2^{\frac{1}{n-1}}c^{\frac{n}{n-1}}\right)^{\frac{1}{G^{-1}\left(\frac{1-r}{r}\right)}}F$, we have $\left(\frac{P}{F}\right)^{G^{-1}\left(\frac{1-r}{r}\right)} < 1-2^{\frac{1}{n-1}}c^{\frac{n}{n-1}}$, while as $P>(1-c)^{1/G^{-1}\left(\frac{c(1-r)}{r(1-c)}\right)}F$, we

[13] Note that β* is a function of P.

also have $\left(\frac{P}{F}\right)^{G^{-1}\left(\frac{c(1-r)}{r(1-c)}\right)} > 1-c$. Thus, the RHS of (16) is higher than its LHS at the value $G^{-1}\left(\frac{c(1-r)}{r(1-c)}\right)$, while the RHS is lower than the LHS at the value $G^{-1}\left(\frac{1-r}{r}\right)$.

Given the LHS of (16) is strictly increasing in the range $\beta^* \varepsilon \left[G^{-1}\left(\frac{c(1-r)}{r(1-c)}\right), G^{-1}\left(\frac{1-r}{r}\right)\right]$ (corresponding to r' ε[c,0.5]) while its RHS is decreasing, and as the functions are continuous, an intersection must exist in between these two values. Thus, $G^{-1}\left(\frac{c(1-r)}{r(1-c)}\right)$ <β*<$G^{-1}\left(\frac{1-r}{r}\right)$, and hence $G(\beta^*)<\left(\frac{1-r}{r}\right)$. Thus, r'<0.5, and jurors' default option if no one ends up paying attention is to acquit. Then, no innocent defendant wants to deviate and accept the plea bargain. Doing so would give the innocent defendant a disutility of P^{β_i}, while opting for trial guarantees a sure acquittal (whether the jurors are attentive or not). Thus, it is optimal for all innocent defendants to reject the plea bargain. Next, given that G(β*) proportion of guilty defendants opt for trial, an individual guilty defendant's expected disutility from rejecting the plea bargain is $p(G(\beta^*))F^{\beta_i}$ as there is a probability $p(G(\beta^*)) = 1 - c^{\frac{n}{n-1}}\left(\frac{G(\beta^*)r+1-r}{G(\beta^*)r}\right)^{\frac{1}{n-1}}$ that the defendant is correctly convicted (note that given the continuity of the mass of defendants, an individual defendant cannot expect to impact the share of guilty defendants opting for trial) while his disutility from accepting the plea bargain is P^{β_i}. Given that P<F, $\left(\frac{P}{F}\right)^{\beta}$ is decreasing in β. Thus, for guilty defendants with β_i<β*, $p(G(\beta^*)) < \left(\frac{P}{F}\right)^{\beta_i}$, so that these defendants prefer to opt for trial as this gives them a lower expected disutility as compared to accepting P. Using similar logic, guilty defendants with β_i>β* accept the plea bargain. This results in exactly G(β*) proportion of guilty defendants opting for trial. Note that σ>0 as β*>$G^{-1}\left(\frac{c(1-r)}{r(1-c)}\right)$. Finally, note that r'<0.5, the appropriate substitutions using (2') and (4') ensure that jurors are choosing σ and p appropriately, so that we get $\sigma(G(\beta^*)) = 1 - \left(\frac{c\{G(\beta^*)r+1-r\}}{G(\beta^*)r}\right)^{\frac{1}{n-1}}, p\big(G(\beta^*)\big) = 1 - c^{\frac{n}{n-1}}\left(\frac{G(\beta^*)r+1-r}{G(\beta^*)r}\right)^{\frac{1}{n-1}}$.***QED***

Example 2. Consider r = 0.7, c = 0.1, n = 3 (the same parameters as in Example 1), and let G be a uniform distribution between the limits 1 and 11.5. Then, any P/F in the interval [.9318,.99] results in a semi-separating equilibrium as described above.

While there is a counterpart to the semi-separating equilibrium of Proposition 1, the comparative statics of Proposition 2 and Remark 2 are also similar. To see this, note that an increase in c or n still decrease p for a given initial β*, thereby shifting the LHS of (16) down, increasing the equilibrium value of β*. But this then increases G(β*), so that a larger fraction of guilty defendants now goes to trial, reducing plea bargaining take up rates. Equivalently, prosecutors need to make plea bargains more lenient if they wish to maintain the old value of β* (thereby maintaining plea take-up rates); this would shift the RHS of (16) down as well. Similarly, Remark 2 continues to hold, as an increase in F shifts down the RHS, decreasing β*, thereby increasing plea bargaining take-up rates.

Focusing on the proof of Lemma 2 – the absence of a fully separating equilibrium where only the innocent go to trial, while only the guilty accept a plea bargain – we can use the same logic as in the main model. Suppose such an equilibrium does exist. Then, jurors do not pay attention, and automatically acquit as r'<c. This creates an incentive for (even a risk-averse) guilty defendant to deviate, reject the plea bargain and go to trial, as there is no risk of conviction, and acquittal is certain, and is therefore superior to any plea bargain with a positive penalty. Thus, the posited equilibrium breaks down.

For the absence of a pooling equilibrium to hold as well, we need an additional restriction:

A3: $\boldsymbol{c < (1-c)^{\bar{\beta}}}$

This condition is easier to satisfy when c is small, and no one is too highly risk averse. For example, if c=.02, the condition holds even for $\bar{\beta}$ = 190. (If c is more moderate – though still low compared to the value jurors attach to a correct verdict – e.g c=0.1 – then $\bar{\beta}$ would be restricted to a smaller value, between 21 and 22. Note that this restriction is satisfied by Example 2).

Now, if A3 holds, so does the counterpart to Lemma 3; all innocent defendants are more likely to go to trial than all guilty defendants. Note that for an innocent defendant with risk aversion parameter β_i to opt for trial, we must have (1-p)$F^{\beta_i} < P^{\beta_i}$ or

$$1\text{-}p<\left(\frac{P}{F}\right)^{\beta_i} \tag{17}$$

(17) is hardest to satisfy for the maximum value of the parameter, $\bar{\beta}$, and for the highest possible value of 1-p (which is c, since 1-c is the lowest possible value of p). Thus, for all innocent defendants to prefer trial to the plea bargain, a sufficient condition is

$$c^{\frac{1}{\bar{\beta}}} < \frac{P}{F} \tag{18}$$

For a guilty defendant with risk aversion parameter β_i to opt for trial, we require $pF^{\beta_i} < P^{\beta_i}$, or

$$p < \left(\frac{P}{F}\right)^{\beta_i} \tag{19}$$

(19) is easiest to satisfy for p=1-c (the lowest possible value of p) and the lowest possible value of β, that is, 1. Thus, the necessary condition for at least some guilty defendant to opt for trial is

$$1 - c < \frac{P}{F} \tag{20}$$

Note that while (18) and (20) have the same RHS, the LHS of (20) is larger than that of (18) by A3, and is thus harder to satisfy. This establishes a counterpart to Lemma 3; even the most risk averse of the innocent are always more likely to reject a plea bargain and opt for trial than even the least risk averse guilty defendants. Using this, we can apply exactly the same steps as in Lemma 4 of the main model to establish the absence of a pooling equilibrium.

The other results do not make use of the risk attitudes of defendants, and thus continue to be valid.

Before closing this sub-section, we briefly discuss Mungan and Klick (2014, 2016). While Mungan and Klick (2014) does not focus on plea bargaining, it argues that it is difficult to make definitive conclusions about criminals' risk attitudes. Mungan and Klick (2016) considers a setup where risk attitudes are heterogeneous and uncorrelated to defendant guilt. Juror decision-making is not modeled explicitly (unlike in my treatment), but after trial, there is an exogenous probability of exoneration, greater for innocent defendants than for guilty ones. While they assume that the most risk-averse innocent defendants always accept a plea bargain, they also show that it is possible to minimize false guilty pleas (by innocent defendants). This can be done by simultaneously making plea offers more lenient, and by slightly increasing exoneree compensation (while lenient plea offers attract both guilty and innocent defendants, slight increases in exoneree compensation make innocent defendants disproportionately more likely to opt for trial, since their

exogenous probability of exoneration exceeds that for guilty defendants). They thus construct a method to avoid a pooling equilibrium where everyone pleads guilty.

My approach differs in that it does not require a mechanism over and above jury trials, and is not dependent on re-examination of past trials and possible exoneration. Instead, it endogenizes juror probabilities of paying attention at trial, using costly attention. Even if defendants are heterogeneous in risk attitudes, in a manner uncorrelated with guilt or innocence, I find that my results are preserved, though some require restricting the size of the upper bound on defendants' risk aversion, and requiring costs of attention to be small.

3.2 Further Remarks

Finally, we may note a feature of my model which, also, is related to the fact that no external mechanism monitoring the jury is assumed. While jurors, once selected, will attend a trial, they cannot in any way be compelled to exert effort to pay attention to trial-relevant information, if they do not find it optimal to do so. Effort is unverifiable, as in many other contexts. Thus, defendants and prosecutors correctly believe (based on the jurors' optimization exercise) that there are plea offers which could induce jurors not to pay attention, as their updated estimates of guilt or innocence would be too sharply skewed to make it worthwhile. The unverifiability of juror effort lends credibility to this, and contrasts with other work where judicial officers are committed to exert some effort (for instance, in reading reviews) and are assumed to rely on evidence rather than on priors. I have used costly juror effort, combined with the unverifiability of such effort, to relax this requirement. Interestingly, Emons and Fluet (2009), who consider a somewhat different context, that of an arbiter receiving reports made by two disputants, allow the arbiter to commit *not to read* the reports submitted to him, with some probability. In my model, even this sort of commitment is not required – inattention may (depending on the parameter ranges) follow optimally, and this possibility will be anticipated by the other players in the game.

4. Conclusion

This is the first paper, to the best of my knowledge, to integrate plea bargaining with costly juror attention. Jurors, though motivated to achieve a correct verdict, experience a small cost of exerting effort. This can create a tendency to rely overly either on the efforts of other jurors, or on an informative prior. When this effect is compounded over all jurors, there are large parameter ranges

over which an improvement in prior informativeness actually worsens verdict quality because of the increased likelihood that no juror ends up paying attention. The interplay of this behavior with the dynamics of plea bargaining generates all the results in my model. In particular, jurors update their priors about defendant guilt depending on the proportion of different types of defendants that accept a given plea offer. This, in turn, affects the defendants' actual willingness to accept such an offer, as it also changes the probability of being convicted at trial. Thus, plea bargaining and juror decision-making mutually affect each other.

I find that neither fully separating nor pooling equilibria exist. The prosecutor may induce semi-separating equilibria, and will do so within a subset where the default option in the event of juror inattentiveness is acquittal (thus, he will offer plea discounts). In this class of equilibria, there is a tradeoff between court costs and inducing a more accurate verdict. Indeed, more accurate verdicts can be obtained than in the absence of plea bargaining. However, there are also equilibria with less accurate verdicts, lower court costs, and more lenient plea sentences (larger plea discounts). Nonetheless, costly juror effort acts to rule out equilibria with very lenient plea sentences, such as those lenient enough to attract innocent defendants. Plea bargaining is always capable of achieving welfare gains, and therefore should not be banned. Moreover, it is not necessary to restrict plea discounts by law.

The analysis generates some empirically testable implications, such as the finding that fixing a punishment if convicted at trial, plea discounts will tend to be higher (or plea take-up rates will be lower) if the number of jurors is higher, or if considerable effort is needed to analyze the case. It also helps resolve the puzzle that defendants opting for trial mostly prefer jury trials to bench trials, while (seemingly paradoxically) bench trials have higher acquittal rates.

Appendix : *σ as a result of juror optimization*

Taking the $r' \geq 0.5$ case for simplicity, I show that the expression for σ in (2) can also be thought of as the result of an optimization exercise. Consider a juror i, who has to choose σ_i to maximize his expected utility (given by (1)), taking all the other jurors' choices of σ as given (I denote these by σ_j). Now, we will have

$$p = 1 - (1 - r')(1 - \sigma_i)(1 - \sigma_j)^{n-1}$$

The above equation differs from (4) only in that it specifies the probability of a correct verdict given that the other n-1 jurors are choosing σ_j, and if juror i is choosing σ_i. If all n jurors still end up being inattentive, and if the collective prior-based verdict is wrong, then that is the only occasion when the verdict will not be correct. Thus, substituting for *p* in (1), juror i chooses

$$\sigma_i = argmax\left[1-(1-r')(1-\sigma_i)\left(1-\sigma_j\right)^{n-1}-\sigma_i c\right]$$

Or

$$(1-r')(1-\sigma)^{n-1}=c$$

(taking the first order condition and substituting in $\sigma_j=\sigma$ for symmetry). But this is exactly the condition from which we get (2) in the text, and thus, we get the same solution for σ.

Remark 3. *Suppose* $P\in[(1-c)F,\left(1-2^{\frac{1}{n-1}}c^{\frac{n}{n-1}}\right)F]$*.Then*

(a) *Both σ and p are increasing in λ.*

(b) *For* $P\in[\left(1-(1-r)^{\frac{-1}{n-1}}c^{\frac{n}{n-1}}\right)F,\left(1-2^{\frac{1}{n-1}}c^{\frac{n}{n-1}}\right)F]$*, p is higher than it would have been for the same r in the absence of a plea bargaining stage. For lower P, the opposite is true.*

Proof. (a) We have already shown, in Proposition 1, that *p* is increasing in λ in the relevant range. We now show that the same is true for σ. Differentiating equation (2'), we obtain

$$\frac{\partial\sigma}{\partial r'}=\frac{1}{n-1}c^{1/(n-1)}(r')^{-n/(n-1)}>0 \qquad (21)$$

As shown earlier, *r'* in turn is increasing in λ, establishing the result.

(b)When $\lambda=\left(\frac{1-r}{r}\right)^2$, we find – substituting in for λ – that *r'*=1-*r*. Hence, at this value of λ, *p(r')=p(1-r)=p(r),* so that the probability of a correct verdict is exactly equal to what it would have been in the absence of plea bargaining, keeping the proportion of guilty arrestees constant. Using (9), this value of λ obtains when the prosecutor sets $P=\left(1-(1-r)^{\frac{-1}{n-1}}c^{\frac{n}{n-1}}\right)F$. (9) also shows that λ increases in P. Thus, it follows that for even higher *P* within the specified range, λ increases. Since *p* with plea bargaining is increasing in λ, so it follows that for even higher λ within the range, *p* would be strictly greater than what it would have been in the

absence of plea bargaining. Next, we can check that subject to A1, we always have that $\frac{c(1-r)}{r(1-c)} < \left(\frac{1-r}{r}\right)^2$. Since p is increasing in λ, it is therefore clear that for any $\lambda \in [\frac{c(1-r)}{r(1-c)}, \left(\frac{1-r}{r}\right)^2)$, $p(\lambda)<p(1-r)=p(r)$, so that in this range, the probability of a correct verdict would have been less than what it would have been without plea bargaining. Since λ increases in P, it follows that for $(1-c)F \leq P < \left(1-(1-r)^{\frac{-1}{n-1}}c^{\frac{n}{n-1}}\right)F$, p is less than it would have been in the absence of plea bargaining. *QED*

Remark 4. *Consider the class of semi-separating equilibria with plea bargaining characterized in Proposition 1. At least some equilibria in this class – for example, those characterized in Remark 3(b) as induced by* $P \in [\left(1-(1-r)^{\frac{-1}{n-1}}c^{\frac{n}{n-1}}\right)F, \left(1-2^{\frac{1}{n-1}}c^{\frac{n}{n-1}}\right)F]$, *achieve greater social welfare than what is achieved in the absence of plea bargaining.*

Proof. Comparing (12) and (13), clearly as λ<1, 1-r+λr<1. In addition, the subset of equilibria characterized in Remark 3 part (b) involved higher p than in the scenario without plea bargaining, so that p(λ)>p(r), and accordingly, (1-p(λ))F<(1-p(r))F. Thus, (13) will necessarily be less than (12). (Belonging to the category of equilibria characterized in Remark 3 part (b) is sufficient, but not necessary, for (13) to be less than (12)). *QED*

References

Austen-Smith, D. and J.S Banks (1996): "Information Aggregation, Rationality, and the Condorcet jury theorem", *American Political Science Review* 90:34-45.

Baker, S., and C. Mezzetti (2001): "Prosecutorial Resources, Plea Bargaining, and the Decision to go to Trial", *Journal of Law, Economics and Organization* 17:149-167.

Bar-Gill, O., and O. Gazal Ayal (2006): "Plea Bargains only for the Guilty", *Journal of Law and Economics* 49:353-364.

Bjerk, D. (2007): "Guilt shall not escape or innocence suffer? The limits of plea bargaining when defendant guilt is uncertain", *American Law and Economics Review* 9:305-329.

Bjerk, D. (2021): "Socially Optimal Plea Bargaining with Costly Trials and Bayesian Juries", *Economic Inquiry* 59:263-279.

Bornstein, B.H. and E. Greene (2011): "Jury decision making: implications for and from psychology", *Current Directions in Psychological Science* 20:63-67.

Cho, I-K. and D.M. Kreps (1987): "Signaling Games and Stable Equilibria", *Quarterly Journal of Economics* 102:179-222.

Eisenberg, T., P.L. Hannaford-Agor, V.P. Hans, N.L. Waters, G.T. Munsterman, S.J. Schwab and M.T. Wells (2005): "Judge-jury agreement in criminal cases: a partial replication of Kalven and Zeisel's The American Jury," *Empirical Legal Studies* 2:171-207.

Emons, W. and C. Fluet (2009): "Accuracy versus falsification costs: the optimal amount of evidence under different procedures," *Journal of Law, Economics and Organization* 25:134-156.

Feddersen, T. and W. Pesendorfer (1998): "Convicting the Innocent: the inferiority of unanimous jury verdicts under strategic voting", *American Political Science Review* 92:23-35.

Grossman, G. and M. Katz (1983): "Plea bargaining and social welfare", *American Economic Review* 73:749-757.

Guha, B (2018): "Secret ballots and costly information gathering: the jury size problem revisited", *International Review of Law and Economics* 54:58-67.

Guha, B. (2020a) "Pretrial beliefs and verdict accuracy: costly juror effort and free riding," *B.E Journal of Theoretical Economics* 20(2): article number 20180020.

Guha, B. (2020b): "Should Jurors Deliberate?" *Review of Law and Economics* 16(2), 2018011.

Guha, B. (2022) "Ambiguity aversion, group size and deliberation: costly information and decision accuracy," *Journal of Economic Behavior and Organization* 201:115-133.

Johnson, B.D. (2019): "Trials and Tribulations: The Trial Tax and the Process of Punishment," *Crime and Justice* 48:313-363.

Kalven, H., H. Zeisel, T. Callahan and P. Ennis (1966): *The American Jury.* Boston: Little, Brown.

Kim, J-Y (2010): "Credible plea bargaining", *European Journal of Law and Economics* 29:279-293.

Landes, W. (1971): "An economic analysis of the Courts," *Journal of Law and Economics* 14:61-107.

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

Leipold, A. (2005): "Why are federal judges so acquittal prone?" *Washington University Law Review* 83:151-227.

Lundberg, A. (2018): "When do the Innocent Plead Guilty?" Working Paper.

Mukhopadhyaya, K (2003): “Jury size and the free rider problem”, *Journal of Law, Economics and Organization* 19:24-44.

Mungan, M.C. and J. Klick (2014): “Forfeiture of illegal gains, attempts, and implied risk preferences,” *Journal of Legal Studies* 43:137-153.

Mungan, M.C. and J. Klick (2016): “Reducing false guilty pleas and wrongful convictions through exoneree compensation,” *Journal of Law and Economics* 59:173-189.

Reinganum, J. (1988): “Plea Bargaining and Prosecutorial Discretion,” *American Economic Review* 78:713-728.

Siegel, R. and B. Strulovici: “Judicial Mechanism Design,” *American Economic Journal Microeconomics*, forthcoming.

Silva, F. (2019): “If we confess our sins,” *International Economic Review* 60:1389-1412.

Smith, D.A. (1987): “The plea bargaining controversy”, *Journal of Criminal Law and Criminology* 77:949-968.

Subramanian, R., L. Digard, M. Washington II, and S. Sorage (2020): “In the Shadows: A Review of the Research on Plea Bargaining.” Vera Institute of Justice.