American Political Science Review (2024) 1–15
doi:10.1017/S0003055424000765

Guilt and Guilty Pleas

ANDREW T. LITTLE *University of California, Berkeley, United States*
HANNAH K. SIMPSON *Texas A&M University, United States*

Plea bargaining figures heavily in criminal justice systems in the United States and, increasingly, around the globe. Conventional wisdom holds that plea bargaining generates efficiency gains for all parties, while sorting the guilty from the innocent. We build a series of formal models to consider the relationship between a defendant's guilt and her likelihood of pleading guilty. In an inversion of the conventional wisdom, we show that under a range of empirically plausible scenarios—for example, if criminals are more risk-seeking than the wrongfully accused, or if prosecutors derive a career benefit from trial wins—the innocent are more likely than the guilty to plea bargain.

INTRODUCTION

In the United States, guilty pleas may constitute as much as 90% to 95% of all convictions (Devers 2011, 1; Hollander-Blumoff 1997, 116). Although rates of conviction by guilty plea tend to be lower in other countries, they are increasing, as jurisdictions worldwide adopt the plea bargain as a way to manage slow court processes and growing caseloads (e.g., Langer 2004, 37; Turner 2010). The growing ubiquity of the institution has led to spirited debates about its fairness, efficacy, and importance. The conventional political economy view is that plea bargaining benefits all participants in the criminal justice system. It assists state officials by sorting the guilty from the innocent (e.g., Baker and Mezzetti 2001; Grossman and Katz 1983), spares both parties the expense of a trial (e.g., Landes 1971), and rewards defendants who waive their right to a trial (or, in the terms used by the Federal Sentencing Commission, who demonstrate "acceptance of responsibility") with milder sentences (Grossman and Katz 1983; King et al. 2005, 961). However, some scholars and activists have argued that the high collateral costs to criminal litigation—for example, the cost of pretrial detention—weaken the institution's efficacy as a sorting mechanism by creating pressures on some innocent defendants to plead guilty (Blume and Helm 2014; Hollander-Blumoff 1997). As a result, there have been calls to remove these costs through policies like bail reform and speedy trial requirements.

In this article, we address this issue from a different angle, asking whether the assertion that plea bargaining can function as a sorting mechanism is necessarily true in the first place. In particular, we ask how several well-known features of criminal defendants and criminal justice officials affect the relationship between guilt and pleading guilty. We focus on the risk preferences of criminal defendants, and prosecutors' career incentives to take cases to trial. To isolate the effect of each on plea bargaining outcomes, we study them separately, albeit using models that share many features. We first ask how personal traits like defendants' risk preferences might affect who pleads guilty, taking plea offers as exogenous. We then investigate the incentives of trial-oriented prosecutors—that is, prosecutors who derive a benefit from winning cases at trial—to offer acceptable plea bargains in the first place, taking the pool of defendants as exogenous.

Our central finding is that *either* the individual risk preferences of criminal defendants *or* the strategic incentives of a trial-oriented prosecutor is sufficient, under a range of empirically plausible conditions, to generate situations in which the innocent not only plead guilty but do so *at higher rates than the guilty*. In other words, we show that the practice of plea bargaining may often lead to "perverse sorting" in which the innocent are more likely to plead guilty, while the guilty are more likely to opt for trial.

We begin by solving a decision-theoretic model in which a citizen characterized by risk attitudes and the extent to which she would benefit from criminal activity decides whether to engage in crime. Whether she commits a crime or not, with some probability (which is higher when she is truly guilty), she is arrested and charged. She then must choose whether to accept a plea bargain or risk a harsher sentence after the lottery of a trial. The probability of being convicted at trial is higher when she is guilty.

Holding risk preferences fixed, we show that—in line with the conventional view—a guilty citizen is more likely to accept a plea bargain than an innocent citizen due to her higher probability of conviction at trial. However, citizens with different risk preferences choose to commit crimes at different rates. In particular, consistent with a large empirical and formal literature

Andrew T. Little, Associate Professor, Department of Political Science, University of California, Berkeley, United States, andrew.little@berkeley.edu
Corresponding author: Hannah K. Simpson, Assistant Professor, Department of Political Science, Texas A&M University, United States, hannah.simpson@tamu.edu

Received: November 06, 2023; revised: April 21, 2024; accepted: May 23, 2024.

(Becker 1968; Block and Gerety 1995; Block and Lind 1975; Ehrlich 1973; Engel and Nagin 2015; Grogger 1991; Mata et al. 2018, see also Polinsky and Shavell 1999, 12),[1] more risk-accepting citizens are more likely to commit crimes. As a result, guilty citizens tend to have higher levels of risk acceptance than innocent citizens, and thus are more likely than their innocent counterparts to view the lottery of a trial with equanimity. If there are sufficient rates of error among either convictions or acquittals (or both), these systematic differences lead the innocent to plead guilty at higher rates than the guilty. In fact, if the probability of conviction at trial is sufficiently similar among guilty and innocent individuals, it can even be the case that *acquitted* individuals are more likely to be guilty than those who accept plea bargains.

Risk preferences are not the only individual characteristic that may generate this type of perverse sorting. In a variation on the model above, we demonstrate that one of the most widely documented behavioral biases, overconfidence (e.g., Moore and Healy 2008), can generate a similar result. Specifically, like risk-acceptant individuals, overconfident individuals may be more apt both to go to trial (because they overestimate the probability that they will be acquitted) and to commit crime (because they underestimate the probability that they will be caught). If there are sufficient rates of error in either convictions or acquittals, overconfidence, like risk acceptance, may cause the emergence of perverse sorting, where the innocent accept plea bargains, while the guilty risk trial.

We turn next to an examination of how prosecutors' incentives might affect their choice of plea offers, and how these choices shape the pools of defendants who plead guilty and go to trial, respectively. Existing literature on strategic prosecutors has tended to assume that prosecutors maximize convictions/sentences (e.g., Gordon and Huber 2002; 2009; Grossman and Katz 1983). We consider a prosecutor who cares about maximizing sentences *but also* (as is common) derives substantial career benefits from winning trials; for example, via the lucrative private practice opportunities available to proven young litigators (e.g., Boylan and Long 2005; Sauer 1998). This trial-oriented prosecutor may offer a plea deal to a criminal defendant, who may accept or reject the offer. If the defendant accepts, the game ends; if the defendant rejects the offer, the prosecutor pays a cost to try the defendant in court. To isolate the role played by the prosecutor's incentives in determining who pleads guilty, we assume that the defendant is rational and risk-neutral. We show that even so, if the prosecutor cares sufficiently about winning trials, a similar sorting problem emerges where the likely-guilty go to trial, while the likely-innocent plead guilty.

The reason is that both the probability of winning at trial and the likelihood of actual guilt increase with the amount of evidence a prosecutor has against a defendant. Consequently, while strategic prosecutors could theoretically sort the innocent from the guilty by offering plea bargains that only the guilty would accept, prosecutors who greatly value trial wins are often incentivized to take the likely-guilty to trial, while pressuring the likely-innocent to plead out. This effect persists even if prosecutors also suffer a cost from wrongly punishing the innocent (or failing to punish the guilty), and may be exacerbated if defendants vary in risk aversion as in our first model.

These results suggest two theoretically distinct impediments to the standard understanding of plea bargaining as a mechanism to sort the guilty from the innocent. First, our decision-theoretic analysis implies that there may exist a set of "confounders"—in the form of defendant characteristics like risk aversion or overconfidence—that interfere with the basic, positive relationship between guilt and the likelihood of taking a (fixed) guilty plea.[2] Second, our analysis of trial-oriented prosecutors' optimal plea bargaining strategies suggests that even if defendants are rational and risk-neutral, the incentives of other actors in the criminal justice system may generate precisely the same perverse sorting effect. Here, prosecutors' strategic choices do not confound the relationship between guilt and pleading guilty; instead, by selecting which pleas to offer to which defendants, they fully determine it.

Our article builds on, and expands, existing work across a range of related substantive areas. First, our results are relevant to the large scholarship on crime and criminal behavior which argues that criminals are likely to be more risk-acceptant than the general population, and points out the important implications of these risk preferences for policing and sentencing strategies (e.g., Becker 1968; Block and Gerety 1995; Block and Lind 1975; Ehrlich 1973; Engel and Nagin 2015; Grogger 1991; Mata et al. 2018; Polinsky and Shavell 1999). We replicate this finding, and demonstrate its importance in a novel way by deriving its implications for the institution of plea bargaining.[3]

Second, we contribute to an ongoing scholarly and popular debate about the costs and benefits of plea bargaining (e.g., Baker and Mezzetti 2001; Blume and Helm 2014; Grossman and Katz 1983; Hollander-Blumoff 1997; King et al. 2005; Landes 1971). We show that even when there are no collateral consequences that might induce guilty pleas among the innocent, plea bargains may not result in the punishment of the guilty

[1] Though see Mungan (2017) for an argument that the empirical observation of risk-preferring behavior by criminals may be an artifact of the repeated interactions between criminals and law enforcement.

[2] That is, if one could condition on these confounders (and the plea offer), it would indeed be the case that guilty individuals are more likely to plead guilty. However, when asking whether plea bargaining successfully sorts the innocent from the guilty among the pool of actual defendants, the unconditional relationship is what matters.

[3] Grossman and Katz (1983), in their analysis of plea bargaining as a socially optimal sorting mechanism, do consider that defendants may vary in both guilt and risk aversion. In their model, heterogeneity in risk aversion adds noise, making it harder to screen out the guilty, but it is never the case that a *higher* proportion of the innocent than the guilty plead guilty. Their analysis differs from ours in this critical respect because they do not consider risk aversion's role in the initial decision to commit a crime.

and the freeing of the innocent, but instead, in the reverse. This result may occur either if citizens' individual characteristics affect both their aptitude for crime and their preferences over pleas, or if prosecutors are career-motivated and successful trials generate career benefits.

Third, our article contributes to the political science literature on strategic prosecutions (e.g., Gordon and Huber 2002; Shotts and Wiseman 2010) and strategic court actors more generally (e.g., Beim, Clark, and Patty 2017; Beim, Hirsch, and Kastellec 2016; Clark 2011; Gordon and Yntiso 2022; Hübert 2019; Lax and Cameron 2007). Within this large field, work on prosecutors tends implicitly to focus on (elected or appointed) bureau chiefs whose tenure depends on their demonstrated competence and/or congruence with the preferences of a principal (e.g., Gordon and Huber 2002; 2009; Shotts and Wiseman 2010). Often, scholars argue, these prosecutors demonstrate competence and congruence by maximizing convictions and/or sentences (e.g., Gordon and Huber 2002; Grossman and Katz 1983). Our article builds on this literature by incorporating work in law and economics that emphasizes the importance to prosecutors of trial wins—as well as conviction rates—as a means of obtaining both retention or promotion in the public sector (Bandyopadhyay and McCannon 2014) and the option of private sector employment (Boylan and Long 2005; Sauer 1998). We show that prosecutors' need for trial wins can affect their prosecution strategies in highly consequential ways.

Finally, our results are relevant to ongoing policy discussions about overloaded courts, and to our understanding of the consequences of racial and economic disadvantage in the court system. With regard to the former, plea bargaining is often touted as a solution to case overload: a way (perhaps the only way) for the criminal justice system to resolve cases quickly and cheaply without sacrificing accuracy. Yet if overload leads to mistakes by prosecutors, judges, and defense attorneys, thereby increasing the rate of trial errors, it may be precisely when courts are overloaded that perverse sorting is at its worst. With regard to the latter, systematic differences across racial or economic groups in the probability of wrongful arrest or conviction might generate systematic variation in the severity of perverse sorting problems. For example, if innocent individuals are systematically more likely to be arrested and charged if they are poor or belong to an ethnic minority, then the distribution of risk aversion among these individuals would be higher, and the distribution of evidence lower, generating an especially high level of perverse sorting. Similar results would obtain if the accuracy of court outcomes is systematically lower for poor or minority-member defendants, perhaps due to bias or less access to counsel. We return to these issues in our discussion.

THE MODELS

The next two sections present two separate models of plea bargaining, although they can be nested in the same wider model and share many features. In the wider model, (1) a citizen decides whether or not to commit a crime, and in turn may be arrested and charged, (2) if arrested and charged, the prosecutor makes a plea offer, and (3) the citizen chooses whether to accept the plea offer or go to trial.

The key difference between the two models lies in which steps are endogenized. In the first, decision-theoretic model, we take the plea deal (step 2) as fixed and focus solely on the citizen's decisions (whether to commit a crime and whether to plead guilty/go to trial). In the second, we take the pool of arrested citizens (step 1) as fixed, and endogenize the prosecutor's choice of a plea deal, along with the citizen's decision to accept the plea. This separation allows us to highlight, in a clear fashion, two separate reasons perverse sorting may occur: variation in the characteristics of individual citizens, and prosecutors' career incentives. At the end of the second model, we discuss how the mechanisms might interact, with a formalization of some such interactions in the Supplementary Material.

DEFENDANT CHARACTERISTICS AND PLEA BARGAINING

There is a citizen, characterized by some level of risk aversion α, who must decide both whether to commit a crime and, if arrested, whether to plead guilty or go to trial. At the beginning of the game, the citizen enjoys a baseline level of consumption y_0. Committing a crime generates a benefit $b \geq 0$, which represents the monetary gain associated with the crime (or its equivalent). As the decision to commit a crime determines whether she is guilty, we write this choice $G \in \{0,1\}$.

After deciding whether to commit the crime, the citizen is arrested and charged with probability p_G. We assume that this probability is at least weakly higher if the citizen is actually guilty, but that innocent citizens are arrested and charged with positive probability: $0 < p_0 \leq p_1 \leq 1$.

If she is charged, the citizen decides whether to plead guilty ($P = 1$) and accept penalty $x_P > 0$, or proceed to trial ($P = 0$) and risk the imposition of penalty $x_K > x_P$ if convicted. Going to trial is risky because with some probability it results in a higher sentence than the plea deal, and with complementary probability it results in no sentence. The probability of conviction at trial is $\pi_G \in (0,1)$, and is weakly higher if the citizen is indeed guilty, $\pi_0 \leq \pi_1$. Following Becker (1968), we interpret the penalties x_P and x_K as monetary losses, either literal fines or the dollar-equivalent value of time spent in jail.

Combining, the citizen's total consumption is

$$y = y_0 + Gb - x,$$

where $x = 0$ if she is not arrested or is acquitted, $x = x_P$ if she accepts a plea deal, and $x = x_K$ if she is convicted at trial.

The citizen's utility, $u(y;\alpha)$, is a strictly increasing function of her consumption y, and is also characterized

by her risk aversion α. We capture risk aversion in a standard fashion, first defining the Arrow–Pratt measure of risk aversion as $\mathcal{A}(y;\alpha) = \frac{-u''(y;\alpha)}{u'(y;\alpha)}$. Assume that higher values of α mean that the citizen is more risk-averse, in the sense that if $\alpha_1 < \alpha_2$, then $\mathcal{A}(y,\alpha_1) < \mathcal{A}(y,\alpha_2)$ for all y. This holds for standard utility functions such as $u(y;\alpha) = \alpha^{-1}(1-e^{-\alpha y})$ and $u(y;\alpha) = \frac{y^{1-\alpha}}{1-\alpha}$. Some results will depend on how the Arrow–Pratt measure of risk aversion changes as a function of y; say that the utility exhibits increasing, constant, or decreasing risk aversion when it is increasing, constant, or decreasing in y.

To summarize, the moves are as follows:

1. The citizen chooses whether to commit the crime, $G \in \{0,1\}$.
2. Nature determines whether the citizen is arrested and charged (probability p_G) or not ($1-p_G$).
3. If the citizen is not arrested, the game ends. If the citizen is arrested, she chooses whether to plead guilty $P = 1$ or go to trial $P = 0$.
4. If the citizen goes to trial, Nature determines whether she is convicted (probability π_G) or acquitted ($1-\pi_G$).

Optimal Behavior

We solve by backwards induction, beginning with the citizen's decision to take a plea.

Plea Decision

Suppose the citizen is guilty. If caught, she accepts a plea bargain if

$$u(y_0 + b - x_P;\alpha) \geq \pi_1 u(y_0 + b - x_K;\alpha) + (1-\pi_1)u(y_0 + b;\alpha).$$

To see how risk aversion affects this choice, define the *certainty equivalent* level of consumption, c, associated with going to trial. Formally, this is the c that solves

$$u(c;\alpha) = \pi_1 u(y_0 + b - x_K;\alpha) + (1-\pi_1)u(y_0 + b;\alpha), \quad (1)$$

i.e., the consumption level at which a guilty citizen with risk aversion α is indifferent between obtaining c for sure, and facing the lottery of a trial. Write the solution to this equation as $c_1(\alpha)$.[4] We can then rewrite the decision to accept a plea as $y_0 + b - x_P \geq c_1(\alpha)$ or

$$x_P \leq y_0 + b - c_1(\alpha) \equiv \bar{x}_{P,1}. \quad (2)$$

Unsurprisingly, the guilty citizen accepts a deal only if it involves a sufficiently small punishment. Central for our purposes is how risk aversion affects the maximal accepted plea deal. By a standard result (e.g., Mas-Colell, Whinston, and Green 1995, Proposition 6.C.2), the certainty equivalent $c_1(\alpha)$ is decreasing in α. Since $c_1(\alpha)$ enters negatively into the right-hand side of Inequality 2, the greater the citizen's risk aversion α, the more punitive the plea deal at which she is indifferent between accepting the plea and going to trial. Put differently, harsher plea deals are accepted as risk aversion increases.

A similar analysis reveals that an innocent citizen accepts plea bargain x_P instead of going to trial if

$$u(y_0 - x_P;\alpha) \geq \pi_0 u(y_0 - x_K;\alpha) + (1-\pi_0)u(y_0;\alpha). \quad (3)$$

Writing the certainty equivalent for the lottery of going to trial when innocent as $c_0(\alpha)$, the innocent citizen accepts a plea deal if

$$x_P \leq y_0 - c_0(\alpha) \equiv \bar{x}_{P,0}. \quad (4)$$

There are two differences which determine whether the innocent or guilty are more apt to accept plea bargains, keeping risk aversion fixed. First, the probability of being convicted is at least weakly lower for an innocent citizen ($\pi_0 \leq \pi_1$). Second, those who are guilty acquired an extra benefit, b, from crime, which increases their utility from any outcome and could potentially affect their tolerance for risk. This yields the following result.

Proposition 1 (Risk Aversion and Guilt Increase Plea Acceptance).

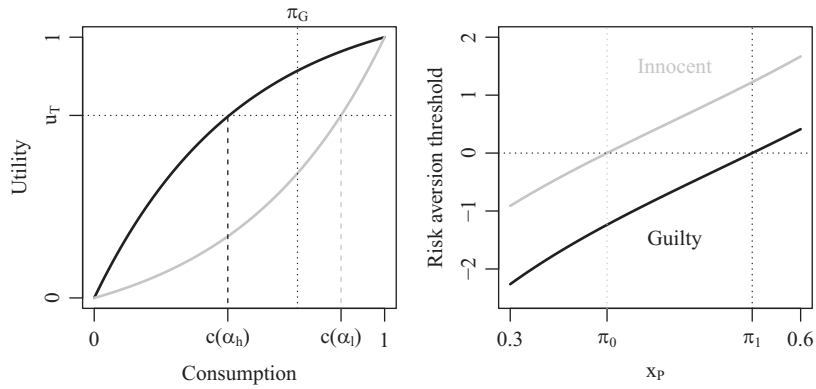
(i) Citizens accept a wider range of plea bargains when they have higher risk aversion ($\frac{\partial \bar{x}_{P,G}}{\partial \alpha} > 0$) or are more likely to be convicted ($\frac{\partial \bar{x}_{P,G}}{\partial \pi_G} > 0$).
(ii) If u has constant or increasing absolute risk aversion in y and $\pi_0 < \pi_1$, then the guilty always accept a wider range of plea deals.
(iii) For any u, if π_0 is sufficiently small (π_1 sufficiently large) the guilty always accept a wider range of plea deals.

Proof. All proofs are in the Supplementary Material.

Figure 1 provides a graphical illustration of Proposition 1. The left panel illustrates how the certainty equivalent of trial changes under low risk aversion (gray line) and high risk aversion (black line) for fixed guilt status G. Consider a lottery which results in consumption 0 or 1; for comparison both utility functions are equal to 0 and 1 at these points, respectively. In particular, suppose that the citizen obtains 1 if she wins at trial and 0 if she loses, and that she wins with probability π_G. So, for either utility function, the expected utility from a trial is $u_T = \pi_G$. The black curve is a utility function with high risk aversion, and the point $c(\alpha_h)$ is the certainty equivalent to the lottery of trial for a citizen with this utility function. The gray curve is a utility function with low risk aversion, and certainty equivalent $c(\alpha_l)$. The black curve crosses the dotted horizontal line at a lower level of consumption, which means the certainty equivalent of trial is lower for

[4] Since u is strictly increasing in c, $u(c;\alpha)$ is less than the trial utility for $c \leq y_0 + b - x_K$ and $u(c;\alpha)$ is greater than the trial utility for $c \geq y_0 + b$; such a solution exists and is unique.

FIGURE 1. Illustration of Risk Aversion and Plea Acceptance with Constant Absolute Risk Aversion: $u(y,\alpha) = \alpha^{-1}(1-e^{-y\alpha})$



Note: Left panel: Certainty equivalent with positive risk aversion ($\alpha = 2$; black) and risk acceptance ($\alpha = -2$; gray). Right panel: Threshold in risk aversion to go to trial as a function of x_P with $y_0 = 1$, $b = 0.1$, $x_K = 1$, for innocent individuals (gray, with $\pi_0 = 0.4$) and guilty individuals (black, with $\pi_1 = 0.55$).

the citizen with high risk aversion. Combined with Inequalities 2 and 4, this means that the threshold plea the citizen would accept rather than go to trial (keeping guilt status fixed) is increasing in her risk aversion α.

The right panel illustrates how guilt status affects the critical value of risk aversion at which the citizen is indifferent between the lottery of trial and plea deal x_P. Higher values of α indicate higher risk aversion, so a citizen offered plea deal x_P accepts if her risk aversion is above the relevant curve. First, notice that since the threshold for the innocent is always higher in this figure, for fixed risk aversion and a fixed plea offer, innocent citizens are always more likely to go to trial. To illustrate further, recall that in this example, the citizen anticipates payoff π_G from trial. At $x_P = \pi_0$, an innocent and risk-neutral ($\alpha = 0$) citizen is indifferent between pleading guilty and trial (and all risk-averse innocent citizens prefer to plead guilty)—but a risk-neutral, guilty citizen would strictly prefer to plead guilty. At $x_P = \pi_1 > \pi_0$, a *guilty* and risk-neutral citizen is indifferent between pleading guilty and trial, while her innocent counterpart strictly prefers to go to trial.

So far, we have shown that for fixed risk aversion and a fixed plea offer, the guilty are generally more likely to plead guilty. This is consistent with the standard view. However, we have also demonstrated that for fixed guilt status, higher risk aversion increases the likelihood of accepting a plea offer. We now consider how a citizen's risk aversion might affect her willingness to commit crimes.

Crime Decision

It is optimal for a citizen to commit a crime if

$$p_1\hat{U}_1(\alpha) + (1-p_1)u(y_0+b;\alpha) \geq p_0\hat{U}_0(\alpha) + (1-p_0)u(y_0;\alpha),$$

where $\hat{U}_G(\alpha)$ is the (expected) utility associated with being charged with a crime with guilt status G and risk acceptance α, given the plea decisions derived above.

The effect of risk aversion on this choice is complicated by the fact that the citizen's expected utility at this stage is always a lottery, regardless of her crime choice, since she may be arrested even if she does not commit a crime. In the extreme, suppose the citizen is arrested with near certainty if she does commit a crime and has an intermediate chance of arrest if she does not commit a crime. If so, the decision *not* to commit a crime is "riskier" in a technical sense: the commission of a crime results in certain capture (and potentially a certain plea deal) while abstaining leads to uncertainty.

We focus on the case where the probability of an innocent person being charged with a crime in any given time period is fairly small. (There may still be a nontrivial share of innocent individuals who are charged, as it is natural to assume that the number of citizens choosing to commit a crime in any given time period is also very small.[5]) As the probability of arrest when innocent, p_0, approaches 0, the citizen commits a crime if

$$u(y_0;\alpha) \leq p_1\hat{U}_1(\alpha) + (1-p_1)u(y_0+b;\alpha).$$

If the citizen (having committed a crime) would accept a plea bargain if caught, we can rewrite this inequality as

$$u(y_0;\alpha) \leq p_1u(y_0+b-x_P,\alpha) + (1-p_1)u(y_0+b;\alpha).$$

[5] For example, suppose the probability that a citizen commits a crime in some time period is $1/1,000$, the probability of being picked up when innocent is also $1/1,000$, and the probability of being picked up when guilty is $1/2$. Then the share of those picked up who are innocent will be $\frac{(999/1,000)(1/1,000)}{(999/1,000)(1/1,000)+(1/1,000)(1/2)} \approx 2/3$.

If the citizen would instead prefer to go to trial if caught, the inequality becomes

$$u(y_0;\alpha) \leq p_1\pi_1 u(y_0 + b - x_K, \alpha) + (1 - p_1\pi_1)u(y_0 + b;\alpha).$$

Let $C_P(b;\alpha)$ be the certainty equivalent of the lottery of committing a crime with benefit b when pleading guilty if caught, and let $C_T(b,\alpha)$ be the certainty equivalent of going to trial if caught. Notice that both are strictly increasing in b and strictly decreasing in α. Combining, we can define the certainty equivalent of committing a crime and making the optimal plea choice as

$$C(b,\alpha) = \max\{C_P(b;\alpha), C_T(b;\alpha)\}.$$

Since both C_P and C_T are strictly increasing in b and strictly decreasing in α, C has the same properties. This yields the following result.

Proposition 2 (Crime Benefit Cutpoint). As $p_0 \to 0$, for any risk aversion α, there exists a critical $\hat{b}(\alpha)$ such that an individual commits a crime if and only if $b > \hat{b}(\alpha)$, with $\hat{b}$ strictly increasing in α.

The key takeaway from this result is the greater the citizen's risk aversion, the larger the benefit must be to induce her to engage in crime. By continuity, this result holds as long as the chance of being charged when innocent is sufficiently low. This implies that, consistent with research in psychology, criminology, and law and economics, individuals who engage in crime are, on average, more risk-acceptant than those who do not.

Who Pleads Guilty?

Having solved the citizen's sequential decision problem, we now consider what her derived optimal choices imply for the efficacy of plea bargaining as a sorting mechanism. Recall that the standard view is that the guilty are more likely to accept a plea deal than the innocent. So far, we have shown that, for fixed risk aversion, this is indeed typically the case. However, we have also shown that innocent citizens are on average more risk-averse than the guilty, and that the more risk-averse a citizen, the more likely she is to accept a plea deal. As a result, depending on the relative strength of these different effects, the innocent may be more likely to plead guilty.

To formalize this possibility, assume that the risk aversion parameter α and the benefit from committing a crime b are independent random variables drawn from continuous distributions. Call the marginal cumulative density functions F_α and F_b.

The three outcomes we study are then random variables at the outset of the decision problem: the decision to commit a crime ($G \in \{0,1\}$), whether a citizen is arrested and charged (call this $A \in \{0,1\}$), and whether a citizen who has been charged accepts a plea ($P \in \{0,1\}$). In terms of these random variables, we are interested in whether it can be the case that

$$Pr(P = 1|G = 0, A = 1) > Pr(P = 1|G = 1, A = 1), \quad (5)$$

i.e., that conditional on being charged, the probability of accepting a plea is higher for the innocent.

From the analysis above, we can represent the probability of pleading guilty conditional on being guilty and being charged as

$$\begin{aligned} &Pr(P = 1|G = 1, A = 1) \\ &= \frac{Pr(\alpha > \hat{\alpha}_1(b), b > \hat{b}(\alpha))p_1}{Pr(\alpha > \hat{\alpha}_1(b), b > \hat{b}(\alpha))p_1 + Pr(\alpha < \hat{\alpha}_1(b), b > \hat{b}(\alpha))p_1} \\ &= \frac{Pr(\alpha > \hat{\alpha}_1(b), b > \hat{b}(\alpha))}{Pr(\alpha > \hat{\alpha}_1(b), b > \hat{b}(\alpha)) + Pr(\alpha < \hat{\alpha}_1(b), b > \hat{b}(\alpha))}, \end{aligned}$$

where $\hat{\alpha}_1(b)$ is the critical value of risk aversion that makes an individual who received benefit b from crime indifferent between accepting a plea and not. Similarly, we can write the probability of pleading guilty conditional on being innocent and charged as

$$\begin{aligned} &Pr(P = 1|G = 0, A = 1) \\ &= \frac{Pr(\alpha > \hat{\alpha}_0, b < \hat{b}(\alpha))}{Pr(\alpha > \hat{\alpha}_0, b < \hat{b}(\alpha)) + Pr(\alpha < \hat{\alpha}_0, b < \hat{b}(\alpha))}, \end{aligned}$$

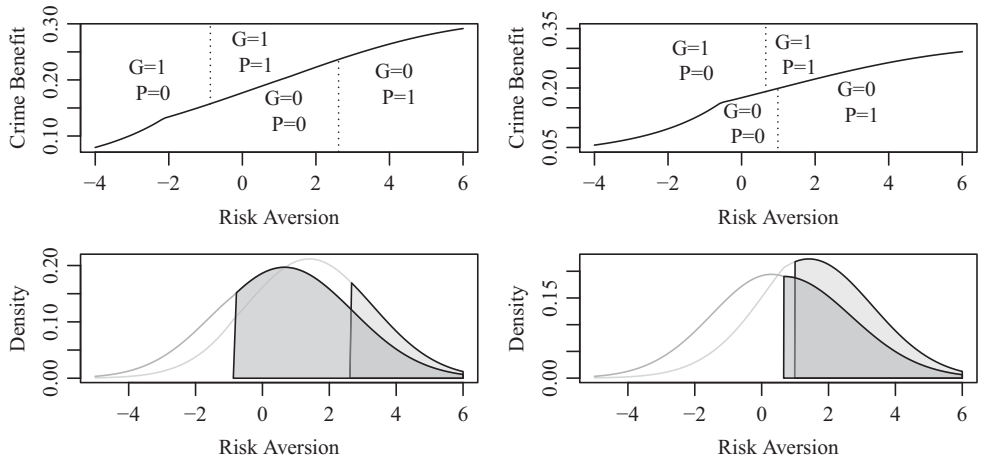
where $\hat{\alpha}_0$ is the critical value of risk aversion that makes an innocent individual indifferent between pleading guilty and going to trial. Note that $\hat{\alpha}_0$ is not a function of the benefit from crime (b) since the innocent do not receive this benefit.

Recall that the willingness of either type of citizen to plead guilty depends in part on her likelihood of conviction at trial, and consider how each type's critical value of risk aversion changes with this likelihood. First, notice that as the probability of conviction at trial when innocent approaches 0 ($\pi_0 \to 0$), the innocent always go to trial, regardless of risk preferences ($\hat{\alpha}_0 \to \infty$), while as the probability of conviction at trial when guilty approaches 1 ($\pi_1 \to 1$), the guilty always plead guilty, no matter their risk preferences ($\hat{\alpha}_1(b) \to -\infty$). This implies that when trials are highly accurate, the standard logic is correct, and the guilty are more apt to take pleas.

Now consider the case where the guilty and innocent face the same chance of being convicted at trial, i.e., $\pi_0 = \pi_1$. For simplicity, assume constant absolute risk aversion. In this case, the threshold in risk aversion at which a plea is preferred is the same for both types: $\alpha_0 = \alpha_1$. However, once we endogenize the crime choice, the innocent are more likely to have a level of risk aversion above this threshold, and hence are more likely to accept a plea. Formally, the CDF of risk aversion conditional on being guilty lies below the CDF conditional on being innocent (when interior): $F_\alpha(\alpha|G = 1) < F_\alpha(\alpha|G = 0)$. As a result, in this part of the parameter space, there is perverse sorting where the innocent are more likely to plead guilty.

Our main result in this section is that this perverse sorting can occur as long as the parameters are not "too far" from this special case. This follows immediately from the preceding argument and the fact that the probability of pleading guilty is continuous in the relevant probability of

FIGURE 2. (Top Panels) Regions of Citizen Behavior and (Bottom Panels) Distribution of Risk Aversion among the Guilty (Dark Gray) and Innocent (Light Gray)



Note: Levels of crime benefit and risk aversion (with $u(y,\alpha) = \alpha^{-1}(1-e^{-y\alpha})$) determine whether a crime is committed and whether a plea would be accepted. In the left panels, there is a large difference between the probability of conviction when guilty ($\pi_1 = 0.6$) and when innocent ($\pi_1 = 0.2$). In the right panels, there is a small difference ($\pi_1 = 0.52$, $\pi_0 = 0.48$).

being convicted (since α is a continuous random variable and the threshold α_G is continuous in π_G).

Figure 2 provides some intuition. The top row presents a partition of the potential realizations of (α, b) into four regions that determine whether a citizen commits a crime and whether each type of citizen pleads guilty if charged. In the left panel, the probability of conviction is much higher for the guilty than for the innocent; in the right panel, the probabilities are similar across types. In both panels, the thick line represents the precise benefit to crime $\hat{b}(\alpha)$ above which a citizen with risk aversion α strictly prefers to commit crime ($G = 1$) and below which she strictly prefers not to ($G = 0$). This critical benefit level $\hat{b}(\alpha)$ is increasing in her risk aversion, since more risk-averse citizens require a higher benefit from crime to risk a criminal charge. The dashed vertical lines represent the level of risk aversion below which citizens go to trial when guilty ($\hat{\alpha}_1(b)$, above the diagonal) and innocent ($\hat{\alpha}_0$, below the diagonal), and above which they plead guilty. Notice that in the left panel, where the probability of conviction is much higher for the guilty, a large proportion of the guilty accept plea deals and a large proportion of the innocent go to trial. Visually, in most of the parameter space, the plea choices "match" the guilt status: $G = P = 1$ or $G = P = 0$. In the right panel, however, where the probability of conviction is similar across types, the "mismatched" regions of guilty citizens going to trial and innocent citizens taking pleas become larger, driven by their differences in risk aversion.

The panels in the bottom row of Figure 2 show this more precisely, by plotting the distribution of risk aversion conditional on being guilty (dark gray) and innocent (light gray). The shaded areas correspond to those with high enough risk aversion to accept a plea deal. In both panels, the distribution of risk aversion is shifted to the right for the innocent. However, in the left panel, there is again a large difference across types in the probability of conviction. As a result, here the risk aversion threshold for a guilty plea is much higher for the innocent, and so the guilty are more likely to plead guilty. In the right panel, where there is only a small difference in the probability of conviction, the risk aversion thresholds are similar across types, and so the innocent are more likely to plead guilty.

The preceding discussion sets aside the possibility that obtaining the benefit to crime b affects risk aversion by changing the citizen's baseline wealth. It seems most plausible in our case to suppose that risk aversion does not change too much with b (i.e., that the citizen's utility function features relatively constant absolute risk aversion). The perverse sorting effect is strengthened if the citizen's utility function has decreasing absolute risk aversion (i.e., obtaining the benefit from crime makes the citizen *even less* risk-averse) and weakened if obtaining b makes the citizen relatively more risk-averse.

Combining, we have the following result.

Proposition 3 (When Are Innocent More Likely to Plead Guilty). Suppose $Pr(P = 1|G, A = 1) \in (0,1)$ for both $G \in \{0,1\}$. Then:

(i) For any fixed π_0, there exists a $\hat{\pi}_1$ such that the innocent are strictly more likely to accept a plea bargain if $\pi_1 \leq \hat{\pi}_1$.
(ii) For any fixed π_1, there exists a $\hat{\pi}_0$ such that the innocent are strictly more likely to accept a plea bargain if $\pi_0 \geq \hat{\pi}_0$.
(iii) If u has constant or decreasing absolute risk aversion, and if the probabilities of conviction π_0 and π_1 are sufficiently close, then the innocent are strictly more likely to accept plea deals.

Proposition 3 says that when trials do a fairly poor job at separating the guilty from the innocent, there is

perverse sorting where the innocent are more likely to accept plea bargains than the guilty. In extreme cases, this implies that if the trial process is sufficiently noisy, *those who are acquitted at trial may be more likely to be guilty than those who accepted plea bargains.*

Such perverse sorting may be common in real-world justice systems, because these systems often feature noisy trial processes with non-negligible rates of error, both in convicting the innocent and in acquitting the guilty. In countries where a high burden of proof—such as proof beyond a reasonable doubt—is required for conviction at trial, the probability of conviction when guilty may be bounded well away from 1. Similarly, in states where the quality of legal representation varies widely, or adjudicators are prejudiced or do not have the time or ability to carefully examine the evidence,[6] the probability of conviction when innocent may be bounded well away from 0. These two features are not mutually exclusive: it is possible for both types of error to coexist in the same system, leading to high rates of both wrongful conviction and wrongful acquittal.

Other Confounders

Risk aversion is not the only individual characteristic that may confound the relationship between guilt and the decision to plead guilty. Various cognitive biases may play a similar role, among them overconfidence—arguably the most widely documented and robust psychological bias (e.g., Moore and Healy 2008). An overconfident citizen could overestimate both the probability that she will get away with committing a crime in the first place *and* the probability that, if caught, she would be acquitted at trial, either because she cannot accurately judge her own abilities or because she engages in a high level of wishful thinking. If so, overconfidence could similarly generate an outcome where the innocent are more likely to plead guilty than are the guilty.

More concretely, let $\tilde{\pi}_1$ represent a guilty citizen's perceived probability of being convicted at trial, and write the choice to accept a plea when guilty as

$$u(y_0 + b - x_P; \alpha) \geq \tilde{\pi}_1 u(y_0 + b - x_K; \alpha) + (1 - \tilde{\pi}_1) u(y_0 + b; \alpha).$$

The more overconfident the citizen, the lower her $\tilde{\pi}_1$, making her more likely to go to trial. The choice to accept a plea when innocent could be represented in a similar fashion, with $\tilde{\pi}_0$ representing the innocent citizen's perceived probability of conviction at trial.

As before, at the crime decision stage, focus on the case where p_0 is small, i.e., where innocent people are arrested with low probability. Since p_0 is small, we can write the choice to commit a crime as

$$u(y_0; \alpha) \leq \tilde{p}_1 \hat{U}_1(\alpha) + (1 - \tilde{p}_1) u(y_0 + b; \alpha),$$

where $\tilde{p}_1$ is the perceived probability of being caught when committing a crime. Since the more overconfident the citizen, the lower she perceives $\tilde{p}_1$ to be, the more overconfident she is the more likely she is to commit a crime.

Combining, overconfidence can increase an individual's willingness to commit crimes and to go to trial, just as risk-acceptance does. Thus, by logic similar to our analysis centered on risk aversion, if trial outcomes are sufficiently noisy, the innocent may be more likely to accept plea bargains than the guilty.

STRATEGIC PROSECUTORS AND PLEA BARGAINING

Our previous analysis takes the sentence associated with a plea bargain as fixed, to focus on the relationship between the decision to plead guilty and various personal characteristics of criminal defendants. In reality, however, plea offers are strategic choices made by prosecutors. The literature on strategic prosecutors generally assumes that prosecutors maximize convictions and/or sentences (e.g., Gordon and Huber 2002; 2009). Such prosecutors would always try to offer defendants the certainty equivalent of the expected outcome at trial, to avoid the possibility of acquittal. However, as a number of scholars have pointed out (e.g., Bandyopadhyay and McCannon 2014; Boylan and Long 2005; Sauer 1998), most prosecutors do not merely value maximizing sentences or conviction rates: they also value trials. More specifically, many prosecutors derive substantial career benefits from *winning* trials, both within government practice and in terms of the opportunities they have to move to private practice. This suggests that prosecutors may have an incentive to try cases, but only those they are most likely to win.

To understand how these incentives may affect the types of plea offers made to innocent and guilty defendants, we consider a situation in which there is some pool of defendants, and study the interaction between a prosecutor and one of these defendants where the actors may agree upon a guilty plea or go to trial. To shut down the dynamics discussed in the previous section, we assume all defendants are risk-neutral. We let $q \in (0,1)$ be the proportion of defendants who are guilty, and assume that guilt is known to the defendant, but not the prosecutor.

At the beginning of the game, both players observe the strength of the evidence available to the prosecutor, $e \in [0,1]$. Suppose that the evidence e is drawn from a conditional distribution $F_e(e|G)$ for $G \in \{0,1\}$, such that $f_e(e|G)$ satisfies the strict monotone likelihood ratio condition.[7] Let $q(e)$ be the posterior probability (from the prosecutor's perspective) that a defendant is guilty given e. By the monotone likelihood ratio condition, $q(e)$ is strictly increasing in e and continuous.

[6] See, for example, Stuntz (2011, 57–8), arguing that inaccurate "[n]oninvestigation is the norm in American criminal litigation."

[7] That is, $f_e(e|G=1)/f_e(e|G=0)$ is a strictly monotone function of e, increasing in e, on $[0,1]$.

Before a trial takes place, the prosecutor offers a plea deal $x_P \geq 0$, which the defendant can accept ($P = 1$) or reject ($P = 0$). If the defendant does not accept the offer, the players go to trial. If the defendant is convicted, sentence x_K is imposed. Suppose that the probability of conviction with guilt status G and evidence e is $\pi_G(e)$, with π_G continuous and strictly increasing in e for either guilt status. In addition, assume $\pi_0(e) \leq \pi_1(e)$ for all e; i.e., fixing the prosecutor's evidence, the innocent are weakly less likely to be convicted, perhaps because the prosecutor does not observe exonerating evidence that may come out at the trial. (However, notice that assuming the same rate of conviction *conditional on the evidence*, i.e., $\pi_0(e) = \pi_1(e)$, would not imply that the guilty and innocent are convicted at the same general rate. Because we assume that the evidence tends to be stronger when the defendant is guilty, it would still be the case that the unconditional probability of conviction is higher for guilty than innocent defendants.) Finally, suppose that the evidence can be extremely strong or weak, in the sense that $q(0) = \pi_G(0) = 0$ and $q(1) = \pi_G(1) = 1$.

In summary, the game proceeds as follows:

1. Nature chooses whether the defendant is guilty (with probability q) or innocent.
2. Both players observe the strength of the evidence e.
3. The prosecutor offers a plea deal x_P.
4. The defendant accepts ($P = 1$) or rejects ($P = 0$) the offer.
5. If the defendant accepts, the game ends. If she rejects, a trial is held. With probability $\pi_G(e)$, she is convicted and sentence x_K is imposed; with probability $1-\pi_G(e)$ she is acquitted.

We assume the defendant's payoff is simply $-x$, where x represents the sentence imposed either via plea or after trial.[8] The linearity of this payoff captures the risk-neutrality of defendants in this section. Given this and the assumptions above, the maximal acceptable plea deal for a defendant with guilt status G facing evidence quality e is

$$\bar{x}_P(G, e) = \pi_G(e)x_K,$$

i.e., the plea deal equivalent to the sentence that would be imposed upon conviction at trial (x_K), weighted by the probability of conviction ($\pi_G(e)$). Since the probability of conviction at trial is weakly higher for a guilty defendant, the maximal acceptable plea deal for a guilty defendant is always weakly less favorable for a fixed level of evidence e. Thus, setting aside the prosecutor's strategic choice of a plea and fixing the evidence, the innocent would be at least weakly less likely than the guilty to take any given plea offer.

We assume that the prosecutor's payoff is linearly increasing in the sentence x imposed on the defendant, regardless of guilt. Assume further that the prosecutor obtains a benefit $w \geq 0$ from winning a trial, but pays a cost $\kappa > 0$ to go to trial. Another way to think about this formalization is that we have normalized the prosecutor's utility from obtaining a conviction via plea bargain (apart from the sentence) to zero. Then, the relative value of a conviction following a trial win to one via guilty plea is $w-\kappa$, while the relative value of a trial loss to a conviction via guilty plea is $-\kappa$. Formally, the prosecutor's utility is

$$u = x + (1-P)(wC-\kappa),$$

where $C = 1$ if the defendant is convicted at the trial, and 0 otherwise.

Potential Offers

We now derive the plea bargain the prosecutor optimally offers to the defendant. Fix the level of evidence e. For any e such that the likelihood of conviction is higher for a guilty defendant (i.e., such that $\pi_0(e) < \pi_1(e)$), there are three possible prosecutor strategies. First, the prosecutor may make a *trial-inducing* offer, i.e., an offer so harsh that both types of defendants reject it and go to trial. Second, the prosecutor may make a *plea-inducing* offer, i.e., an offer so lenient that both types of defendants accept it. Finally, the prosecutor may make a *screening* offer, i.e., an offer which the guilty accept, but the innocent reject.[9]

Begin with the screening strategy, in which the prosecutor makes an offer that only the guilty accept. Because the prosecutor's utility is increasing in sentence length, he makes the maximum offer acceptable to the guilty. This is the offer at which a guilty defendant is indifferent between taking a plea and his expected payoff at trial: $\pi_1(e)x_K$. As long as $\pi_1(e) > \pi_0(e)$, this offer is not acceptable to an innocent defendant with evidence level e, because the innocent defendant expects the sentence $\pi_0 x_K < \pi_1 x_K$ at trial. When making the screening offer, the prosecutor's utility is $\pi_1(e)x_K$ if the defendant is guilty and accepts, and $\pi_0(e)(w + x_K)-\kappa$ if the defendant is innocent and goes to trial. His total expected payoff is then

$$u_S(e) = q(e)\pi_1(e)x_K + (1-q(e))(\pi_0(e)(w + x_K)-\kappa). \quad (6)$$

Now consider the trial-inducing strategy. For a given level of evidence e, the trial-inducing offer is any x_P such that $x_P > \pi_1(e)x_K$. Because both types of defendant now anticipate a lower expected sentence at trial than that attached to the plea deal, both types reject this offer and the prosecutor's expected payoff is

[8] To make this more closely resemble the previous model, we could rewrite the utility is $y-x$, where y is baseline consumption. This would not affect the equilibrium choices.

[9] We assume the trial adjudicator does not make inferences about guilt based on the offer and acceptance choice; empirically speaking, plea bargaining is a private process, so these are usually unknown to the adjudicator. As elaborated in the "Discussion" section, such inferences, if possible, would make screening less likely.

$$u_T(e) = [q(e)\pi_1(e) + (1-q(e))\pi_0(e)](x_K + w) - \kappa, \tag{7}$$

where $q(e)\pi_1(e) + (1-q(e))\pi_0(e)$ is the overall probability of winning at trial, averaging across guilt status.

Finally, consider the plea-inducing strategy. Here, the prosecutor must offer a sentence that both a guilty and an innocent defendant would accept. Accordingly, the maximum plea deal he can offer is $x_P = \pi_0(e)x_K$, i.e., the equivalent to the innocent defendant's expected payoff at trial. His expected payoff from this strategy is

$$u_P(e) = \pi_0(e)x_K. \tag{8}$$

To determine the prosecutor's optimal offer, we pairwise compare his payoffs from each of the three strategies above. In particular, we ask how the relative values of each offer changes as the evidence gets stronger. In the main text, we restrict attention to comparisons where each relative value has at most one crossing.

Assumption 1. Each utility difference, $u_T(e) - u_S(e)$, $u_S(e) - u_P(e)$, and $u_T(e) - u_P(e)$, crosses zero at most once.

In other words, we assume that $u_T(e) - u_P(e)$, $u_T(e) - u_S(e)$, and $u_S(e) - u_P(e)$ switch signs at most once over the range of $e \in [0, 1]$. This means that a prosecutor's preference ranking of any two strategies cannot fluctuate back and forth as the evidence increases.[10] The Supplementary Material contains a more detailed discussion of when this assumption holds and how the results change when it does not.

Special Case 1: $\pi_0(e) = \pi_1(e)$

We start by analyzing the special case where, conditional on the strength of the evidence available at the time of plea bargaining, the probability of conviction at trial is the same for both types, and so we can write it $\pi(e)$. Recall that this does not mean that the innocent and guilty are equally likely to be convicted at trial in general, since the guilty tend to have more evidence against them.

Since the guilty and innocent have the same utility from going to trial, the prosecutor cannot screen the guilty: he must either make an offer all defendants accept, $x_P = \pi(e)x_K$, or take all to trial. He prefers the former if

$$\pi(e)x_K \geq \pi(e)(w + x_K) - \kappa, \text{ or}$$

$$\pi(e) \leq \kappa / w. \tag{9}$$

If $w < \kappa$, then this condition holds for all e. Intuitively, if going to trial is costly for the prosecutor at any value of $e \in [0, 1]$, he always strikes a deal to avoid paying this cost.

If $w > \kappa$, then Inequality 9 is not met at $\pi(e) = 0$, is met at $\pi(e) = 1$, and because $\pi(e)$ is monotonically increasing in e, there is a critical value of evidence $e^T \in (0, 1)$ at which Inequality 9 is met with equality. When $e < e^T$ the prosecutor makes an offer which all defendants accept, and when $e > e^T$ the prosecutor makes an offer which all reject. However, because we have assumed that the evidence tends to be stronger when the defendant is guilty, the probability that $e > e^T$ is also higher when the defendant is guilty. As a result, in this part of the parameter space, there is always perverse sorting.

Proposition 4. Suppose $\pi_1(e) = \pi_0(e)$ for all e. Then:

(i) If $w < \kappa$, the prosecutor offers $\pi(e)x_K$ to all defendants for all values of e, and all accept: both types of defendant plead guilty at the same rate.
(ii) If $w > \kappa$, there exists an $e^T \in (0, 1)$ such that when $e < e^T$ the prosecutor makes an offer accepted by all defendants, and when $e > e^T$ the prosecutor makes an offer rejected by all defendants. The innocent are strictly more likely to plead guilty.

As this special case highlights, plea offers only induce *correct* sorting when the prosecutor makes the screening offer, which in turn requires that the probability of conviction, conditional on the available evidence, is higher for the guilty than for the innocent ($\pi_1(e) > \pi_0(e)$).

Special Case 2: $\pi_1(e) > \pi_0(e)$, $w < \kappa$

Now consider the case where $\pi_1(e) > \pi_0(e)$, but the cost of a trial exceeds the benefit from a win, $w < \kappa$. In this case, the prosecutor never takes all defendants to trial. To see why, compare the prosecutor's payoff from the screening offer (Equation 6) to his payoff from the trial-inducing offer (Equation 7). The payoff from trial is larger if

$$u_T(e) - u_S(e) = w\pi_1(e) - \kappa \geq 0, \tag{10}$$

which never holds if $w < \kappa$. In other words, if the benefit from winning trials is below the cost, the prosecutor's only two options are to make the screening offer or to make an offer which all accept. Moreover, comparing the screening offer payoff to the payoff the prosecutor obtains if all plead guilty (Equation 8), he prefers to screen if

$$u_S(e) - u_P(e) = \frac{q(e)}{1-q(e)} x_K(\pi_1(e) - \pi_0(e)) + \pi_0(e)w - \kappa \geq 0. \tag{11}$$

[10] A stronger, but more intuitive, assumption is that these three differences are monotone in e. As they are all generally increasing in e, this assumption is satisfied if, as $e \to 1$, the probability of conviction for an innocent person does not increase too much faster than the probability of conviction for a guilty person ($\frac{\partial \pi_0(e)}{\partial e}$ is not too much larger than $\frac{\partial \pi_1(e)}{\partial e}$).

Notice that since $w < \kappa$, this condition holds neither at $e = 0$ nor at $e = 1$, although it is increasing in the evidence e. Therefore, by the single-crossing condition, we assumed above, there is no level of evidence e at which the prosecutor makes the screening offer.[11]

Proposition 5. If $\pi_1(e) > \pi_0(e)$ and $w < \kappa$, the prosecutor always makes an offer that all defendants accept. There is no sorting.

Main Case: $\pi_1(e) > \pi_0(e)$, $w > \kappa$

The remaining case is that in which the probability of conviction is higher for the guilty even after conditioning on the evidence $\pi_1(e) > \pi_0(e)$, and the benefit to a trial win exceeds the cost of trial, $w > \kappa$. In this case, as $e \to 1$, the fraction of defendants who would be acquitted at trial, whether innocent or guilty, approaches zero, and the prosecutor's optimal strategy is to take all to trial and obtain $x_K + w > \kappa$. Likewise, as $e \to 0$, the prosecutor's utility from both the screening offer and the trial offer approaches $-\kappa$, and he prefers to plead out all defendants and receive payoff $\pi_0(e)x_K \geq 0$. The relevant question then concerns the existence and size of an intermediate range of e where the prosecutor uses the screening offer.

To answer this question, recall the pairwise payoff comparisons we made above. First, observe that the value of the screening offer relative to pleading out all defendants, $u_S(e) - u_P(e)$, is increasing in e, so there exists a critical evidence level e^{SP} such that the prosecutor prefers screening to pleading out all defendants only if $e > e^{SP}$. Similarly, since the relative value of taking all defendants to trial compared to screening out the guilty ($u_T(e) - u_S(e)$) is increasing in e, there is a critical value e^{TS} such that the prosecutor prefers the trial-inducing offer to the screening offer only if $e > e^{TS}$. Finally, a comparison of the payoffs from taking all to trial and inducing all to plead guilty yields that the prosecutor prefers the former if

$$u_T(e) - u_P(e) = q(e)(\pi_1(e) - \pi_0(e))(x_K + w) + \pi_0(e)w - \kappa \geq 0, \tag{12}$$

which again does not hold at $e = 0$ but strictly holds at $e = 1$, implying that there is a critical value e^{TP} such that the prosecutor prefers the trial-inducing offer to the plea-inducing offer only if $e > e^{TP}$.

The prosecutor's equilibrium behavior depends on the ordering of these three evidentiary thresholds. There are two sub-cases. First, if $e^{TS} \leq e^{SP}$, the prosecutor already prefers inducing trial to screening at a level of evidence where he still prefers inducing pleas to screening. This means that screening is never optimal, and the prosecutor's actions depend only on whether he derives greater utility from inducing trial or inducing a plea. Here, the prosecutor tries all types of defendant when the evidence is sufficiently strong, $e > e^{TP}$, and pleads out all defendants otherwise. Because defendants with more evidence against them are more likely to be guilty, this results in perverse sorting.

By contrast, if $e^{SP} < e^{TS}$, then there is an intermediate range of evidence, $e \in (e^{SP}, e^{TS})$, where the prosecutor prefers the screening offer to both the plea-inducing and the trial-inducing offer. In this sub-case, the prosecutor tries all defendants for $e > e^{TS}$, screens the guilty for $e \in (e^{SP}, e^{TS}]$, and pleads all defendants out for $e < e^{SP}$. Here, as shown in the proof of the following formal result, screening is possible for intermediate $e \in (e^{SP}, e^{TS})$ only when $q(e^{TS}) > w/(w + x_K)$, i.e., when the value of trial wins is not too high.

Even if screening happens, we have not determined whether in this sub-case, the distribution of the evidence is such that on average the guilty are more likely to plead guilty: this depends on the shape of the evidentiary distribution and the size of the interval (e^{SP}, e^{TS}) where screening is optimal. Recall from above that at $\pi_1(e) = \pi_0(e)$, the screening offer and the offer which all accept become equal, at $\pi(e)x_K$. By continuity, this implies that when the distance between $\pi_1(e)$ and $\pi_0(e)$ is sufficiently small, so is the distance between e^{SP} and e^{TS}, and the guilty are still (overall) strictly more likely to plead guilty.[12]

Summarizing this case:

Proposition 6. When the benefit to trial wins exceeds the cost, $w > \kappa$, and $\pi_1(e) > \pi_0(e)$:

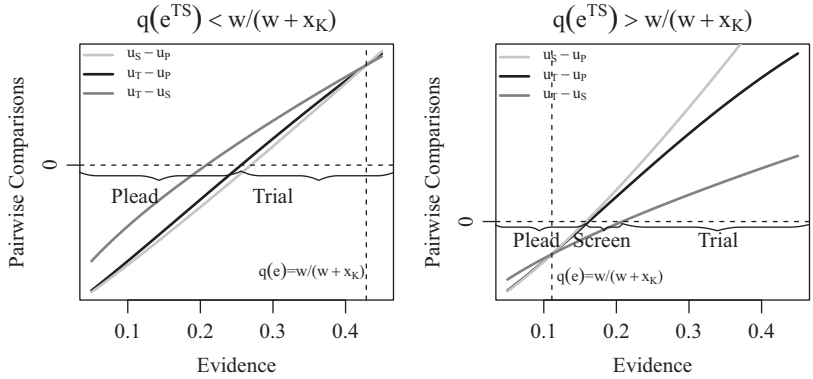
i. When $q(e^{TS}) < w/(w + \kappa)$, the prosecutor makes an offer all accept for $e < e^{TP}$, and tries all defendants for $e > e^{TP}$. The innocent are strictly more likely to plead guilty.
ii. When $q(e^{TS}) > w/(w + \kappa)$, the prosecutor makes an offer all accept for $e < e^{SP}$, makes the screening offer for $e \in [e^{SP}, e^{TS})$, and takes all defendants to trial for $e > e^{TS}$. If the distance between $\pi_1(e)$ and $\pi_0(e)$ is sufficiently small (in a sense formalized in the Supplementary Material), then innocent are more likely to plead guilty.

Figure 3 illustrates Proposition 6. The left panel displays equilibrium behavior in the case in which $q(e^{TS}) < w/(w + \kappa)$. Here, at the critical evidence level (e^{TS}) at which the prosecutor is indifferent between the

[11] In the Supplementary Material, we dispense with this assumption and discuss the conditions under which screening might then occur. Intuitively, the benefit to the prosecutor of imposing a larger sentence, $x_K(\pi_1(e) - \pi_0(e))$ must be large enough to outweigh the cost of sometimes going to trial. This could occur at intermediate e if the difference between $\pi_1(e)$ and $\pi_0(e)$ is large.

[12] In the Supplementary Material, we discuss how our results generalize to a version of the model without the single-crossing assumption made here. Briefly, the three pairwise payoff comparisons derived above cross only once, at $q(e) = w/(w + x_K)$. This means the *order* in which they are satisfied always depends solely upon whether or not $q(e^{TS}) > w/(w + x_K)$, implying that the basic results in Proposition 6 should continue to hold, despite the difficulty of precisely characterizing prosecutor behavior for some interior values of e.

FIGURE 3. Equilibrium Behavior



Note: The left-hand panel presents equilibrium behavior when $q(e^{TS}) < w/(w + x_K)$. The right-hand panel presents equilibrium behavior when $q(e^{TS}) > w/(w + x_K)$. The dashed vertical line in each graph represents the point at which $q(e) = w/(w + x_K)$. Parameter values: $w = 1.5$, $\kappa = 0.5$, $q(e) = e$, $\pi_1(e) = e^{\alpha_1}$ for $\alpha_1 = 0.7$, $\pi_0(e) = e^{\alpha_0}$ for $\alpha_0 = 1$. Left panel: $x_K = 2$. Right panel: $x_K = 12$.

trial-inducing offer and the screening offer (i.e., at which the dark gray curve $u_T - u_S$ crosses 0), he strictly prefers the plea-inducing offer to the screening offer (the light gray curve, $u_S - u_P$, is negative). He therefore makes either the plea-inducing offer (for $e < e^{TP}$) or the trial-inducing offer (for $e > e^{TP}$) and there is perverse sorting in equilibrium.

The right panel displays equilibrium behavior in the case in which $q(e^{TS}) > w/(w + \kappa)$. Here, at the critical level e^{TS}, the prosecutor strictly prefers both trial and screening to the plea-inducing offer (at the point where the dark gray curve, $u_T - u_S$, hits zero, the light gray curve, $u_S - u_P$, is above zero) . He therefore makes the screening offer when $e \in (e^{SP}, e^{TS})$. Observe that in this case, the range of evidence where it is optimal for the prosecutor to make the screening offer is quite small; therefore, for plausible distributions of evidence, there would still be perverse sorting.

The most striking implication of the results detailed in Proposition 6 is that perverse sorting may occur under most plausible empirical conditions. First, perverse sorting is likelier the higher the benefit to prosecutors from trial wins, and we have already argued that at least for the assistant prosecutors who conduct the vast majority of litigation in any prosecutor's office, trial wins are not only "worth the cost" of trial, they are critical both to promotion within the office and (perhaps even more importantly) to obtaining a much more lucrative position at a private law firm.

Second, perverse sorting is likelier when, conditional on the evidence, the innocent and guilty face similar probabilities of conviction. Empirically, this is likely the case. In countries where the defendant has the right to know the evidence against him, the prosecutor's evidence of guilt is known to both parties before trial. In countries like the United States, the prosecutor's evidence of innocence is *also* known (by law) to both parties before trial. This suggests that only when significant "exculpatory surprises" are likely at trial will the expected rate of conviction for guilty and innocent defendants be substantially different—and defendants have numerous incentives not to keep exculpatory information secret in order to reveal it at trial. For example, revealing exonerating evidence pretrial is useful both in pushing the total evidence well below the threshold at which the prosecutor prefers trial and in substantially decreasing the expected plea offer.[13]

One possible argument against these findings is that prosecutors are motivated by the desire to do justice. In the Supplementary Material, we briefly consider the robustness of our perverse sorting result to a situation in which the prosecutor suffers a disutility from inaccurate outcomes. Specifically, we assume that convicting the innocent generates a negative payoff of $-\psi$ for $\psi \in (0, 1)$ for the prosecutor, while acquitting the guilty generates a negative payoff of $-(1-\psi)$. An accurate outcome generates a payoff of 0.

We show that while the accuracy motive increases the attractiveness of the screening offer and therefore broadens the range of the parameter space in which correct sorting is possible, perverse sorting remains an equilibrium outcome under circumstances similar to those elaborated in Proposition 6. If the probabilities of conviction for guilty and innocent are the same conditional on the evidence, accuracy concerns can even exacerbate perverse sorting if the prosecutor is more worried about wrongful acquittals than wrongful convictions. However, given accuracy concerns, when the conditional probabilities of conviction differ across the innocent and the guilty, higher benefits to trial wins are in general needed for perverse sorting to occur, and when it does occur, its *degree* may be lessened, in the sense that concern for accuracy may decrease the evidentiary cutoff above which the prosecutor takes cases to trial.

[13] It is, of course, possible that some exculpatory evidence—for example, the personal credibility of a witness—might be known to the defendant, without being communicable to the prosecutor.

A final question is how prosecutor and defendant behavior might change if defendants varied in risk aversion. In the Supplementary Material, we briefly consider this problem, focusing on the case where the conditional probability of conviction at trial is the same for guilty and innocent defendants. We begin by assuming that the prosecutor observes risk aversion. We show that because more risk-averse defendants accept longer plea sentences to avoid trial, and the prosecutor values large sentences, risk aversion increases the appeal to the prosecutor of a plea bargain relative to a trial. Thus, the threshold level of evidence above which the prosecutor prefers trial is *increasing* in defendant risk aversion: if the defendant is sufficiently risk-averse, the plea sentence she accepts is so large that the prosecutor may prefer a guilty plea for almost any value of the evidence. If the probability of being guilty is lower for more risk-averse defendants, this implies that risk aversion exacerbates perverse sorting.

If risk aversion is unobserved, it is more difficult to characterize the optimal plea and determine how it changes with the strength of the evidence. This is because, if both risk aversion and evidence vary systematically with guilt, the prosecutor uses the level of evidence against the defendant to update his beliefs about both her guilt and her likely risk aversion. However, we show that many of the features that lead to perverse sorting are also present in any equilibrium to this version of the model.

DISCUSSION

We have shown that plea bargaining may generate perverse sorting either if court outcomes are noisy and those who commit crimes differ from those who do not in ways that affect their predisposition to accept plea bargains, or simply if prosecutors value winning trials. We now consider these results in a broader context, first discussing several generalizations to the technical analysis and then discussing insights the model can provide into questions about discrimination and institutional design in criminal justice systems.

First, both our models treat the probability of conviction at trial as exogenous to defendants' plea decisions. In other words, a defendant's refusal to plead guilty does not affect the probability that she is guilty, in the eyes of the convicting body at trial. This is for two reasons. One: the convicting and sentencing bodies at trial are almost never privy to the details of pretrial plea bargaining between prosecutors and defendants. Indeed, a common criticism of the plea bargaining process is its opacity, even to the judges who accept the pleas. And two: neither a jury nor—because the vast majority of cases are resolved via plea bargaining—a judge is likely to know the distribution of evidence across the entire pool of cases handled by the prosecutor.

However, it is natural to ask how our results might change if convicting bodies did make inferences about guilt based on defendants' decisions to plead guilty. In this case, our model suggests that the major consequence would be to further undermine the utility of plea bargaining as a sorting mechanism. This is because if rejection of a plea offer is perceived by the convicting body to signal innocence (guilt), both types of defendant would have a higher (lower) incentive to reject. For example, in our model of strategic prosecution, separation of types at a given level of evidence e implies that the prosecutor is using the screening strategy. If a jury or judge believes that the prosecutor is using the screening strategy, they would conclude that defendants who go to trial are probably innocent, which would create an incentive for guilty defendants to pool with the innocent on plea refusal in order to signal innocence.

A related question is how the results might differ if both the choice to commit a crime and the plea offer were endogenized. As we discuss in the Supplementary Material, endogenizing the crime choice when prosecutors can observe defendant risk aversion would require specifying the joint distribution of guilt, risk aversion, and evidence (because prosecutors would exploit defendant risk aversion to increase the sentences attached to plea deals). However, in this case, strategic prosecution may further deter the risk-averse from crime, because the risk-averse derive a greater disutility from trial (and would face a harsher "certainty equivalent" plea offer). If prosecutors do *not* observe defendant risk aversion, then they cannot condition offers on it—only on the evidence they observe. As a result, in any equilibrium, *realized* risk aversion does not affect the distribution of plea offers, suggesting that here, the more risk averse would again be less apt to commit crime, and more likely to accept plea deals. However, fully characterizing the equilibrium is challenging because the prosecutor now selects offers based on his inferences about defendant risk aversion given the observed evidence—which may affect the initial mapping between citizen risk aversion and the crime decision. In addition, whatever the observability of risk aversion, fully endogenizing this choice would also require making additional assumptions about the defendant's knowledge of prosecutor costs and benefits.

A more substantive question naturally arising from our model involves the consequences of heterogeneity in the tenor or outcomes of individual interactions with the justice system. Specifically, the citizen-defendants we study are differentiated in the first model only by risk aversion and benefit to crime, and in the second model only by exogenous guilt status. Our results are thus an explication of the relationship between endogenous or exogenous guilt status and plea bargaining behavior, fixing all other characteristics that might affect a citizen's interactions with the criminal justice system. However, a large scholarship suggests that characteristics such as race, ethnicity, and class, may significantly and systematically alter the nature of these interactions. For example, scholars have found that economically disadvantaged individuals and members of ethnic or racial minorities face higher probabilities of being arrested (p_G) and convicted at trial (π_G) (e.g., Bjerk and Helland 2020; Galanter 1974; Gelman, Fagan, and Kiss 2007) and that these disparities may

in part be due to different patterns of *wrongful* arrest (e.g., Gelman, Fagan, and Kiss 2007) or conviction (Bjerk and Helland 2020) across groups. Different social groups have also been found to receive different ranges of plea offers (x_P) (e.g., Kutateladze, Andiloro, and Johnson 2016) and may have different distributions of benefit to committing crime b (e.g., Stuntz 2011, 49).

This type of systematic difference would generate variation, across groups, in the likelihood and severity of the perverse sorting problem. For example, if racial minorities are more likely to be arrested and charged when innocent, then the distribution of evidence among arrested and charged racial minorities would be much weaker, and perverse sorting would occur for a much larger fraction of the population. Similar results would obtain for members of social groups for whom court outcomes are noisier, or (fixing conviction sentence) to whom more severe plea bargains are offered.

Finally, our results invite broader inquiry into the extent to which a system of criminal adjudication that relies heavily on plea bargaining is suboptimal for criminal defendants, and for society in general. The answers are unclear. On the one hand, because defendants are not obligated to accept pleas, it may be the case that the option to plea bargain benefits defendants on average: if prosecutors do not precisely tailor pleas to each defendant, but instead make the same offer to similar defendants, all who accept will be weakly better off, and because of imprecision in the plea offer, some will be strictly better off. On the other hand, if defendants are not aware of, or do not fully account for, the negative consequences that attend *all* criminal convictions (social disapprobation, lost job opportunities, etc), they may tend to accept plea deals that are too harsh. Here, the opportunity to plead guilty would make defendants strictly *worse* off.

The implications of dispensing with plea bargaining altogether are also unclear. Since plea bargains are cheaper than trials, requiring only trials might result in many fewer cases being brought. While the cases brought would be those in which the evidence was strongest, complicated cases would not be pursued and many guilty people would go unpunished. Moreover, even within the pool of the very likely guilty, the cheapest cases would be prioritized (see, e.g., Stuntz 2011, 54: "When budget constraints drive the decisions that fill prison beds, the criminals who pay the highest price for their crimes will be those who are most cheaply caught and convicted"). If, as is plausible, the ease of obtaining a conviction increases not only with the quality of the evidence but also with the poverty of the defendant, this could create a situation in which the guilty poor are tried and convicted, and all others are not prosecuted.

SUPPLEMENTARY MATERIAL

The supplementary material for this article can be found at https://doi.org/10.1017/S0003055424000765.

ACKNOWLEDGMENTS

Many thanks to Sam England, Carlo Horz, Ryan Hübert, Keith Schnakenberg, Ian Turner, audience members at APSA 2023, Washington University in St. Louis, and PEPL 2024, and three anonymous reviewers for their helpful comments.

CONFLICT OF INTEREST

The authors declare no ethical issues or conflicts of interest in this research.

ETHICAL STANDARDS

The authors affirm that this research did not involve human participants.

REFERENCES

Baker, Scott, and Claudio Mezzetti. 2001. "Prosecutorial Resources, Plea Bargaining, and the Decision to Go to Trial." *Journal of Law, Economics, and Organization* 17 (1): 149–67.
Bandyopadhyay, Siddhartha, and Bryan C. McCannon. 2014. "The Effect of the Election of Prosecutors on Criminal Trials." *Public Choice* 161 (1–2): 141–56.
Becker, Gary S. 1968. "Crime and Punishment: An Economic Approach." *Journal of Political Economy* 76 (2): 169–217.
Beim, Deborah, Alexander V. Hirsch, and Jonathan P. Kastellec. 2016. "Signaling and Counter-Signaling in the Judicial Hierarchy: An Empirical Analysis of En Banc Review." *American Journal of Political Science* 60 (2): 490–508.
Beim, Deborah, Tom S. Clark, and John W. Patty. 2017. "Why Do Courts Delay?" *Journal of Law and Courts* 5 (2): 199–241.
Bjerk, David, and Eric Helland. 2020. "What can DNA Exonerations Tell Us about Racial Differences in Wrongful-Conviction Rates?" *Journal of Law and Economics* 63 (2): 341–66.
Block, Michael K., and Robert C. Lind. 1975. "An Economic Analysis of Crimes Punishable by Imprisonment." *Journal of Legal Studies* 4 (2): 479–92.
Block, Michael K., and Vernon E. Gerety. 1995. "Some Experimental Evidence on Differences between Student and Prisoner Reactions to Monetary Penalties and Risk." *Journal of Legal Studies* 24 (1): 123–38.
Blume, John H., and Rebecca K. Helm. 2014. "The Unexonerated: Factually Innocent Defendants Who Plead Guilty." *Cornell Law Review* 100: 157.
Boylan, Richard T., and Cheryl X. Long. 2005. "Salaries, Plea Rates, and the Career Objectives of Federal Prosecutors." *Journal of Law and Economics* 48 (2): 627–51.
Clark, Tom S. 2011. *The Limits of Judicial Independence*. New York: Cambridge University Press.
Devers, Lindsey. 2011. "Plea and Charge Bargaining." Bureau of Justice Assistance, Department of Justice. https://bja.ojp.gov/sites/g/files/xyckuh186/files/media/document/PleaBargainingResearchSummary.pdf.
Ehrlich, Isaac. 1973. "Participation in Illegitimate Activities: A Theoretical and Empirical Investigation." *Journal of Political Economy* 81 (3): 521–65.
Engel, Christoph, and Daniel Nagin. 2015. "Who is Afraid of the Stick? Experimentally Testing the Deterrent Effect of Sanction Certainty." *Review of Behavioral Economics* 2 (4): 405–34.
Galanter, Marc. 1974. "Why the Haves Come Out Ahead: Speculations on the Limits of Legal Change." *Law & Society Review* 9 (1): 95–160.
Gelman, Andrew, Jeffrey Fagan, and Alex Kiss. 2007. "An Analysis of the New York City Police Department's "Stop-and-Frisk"

Policy in the Context of Claims of Racial Bias." *Journal of the American Statistical Association* 102 (479): 813–23.
Gordon, Sanford C., and Gregory A. Huber. 2002. "Citizen Oversight and the Electoral Incentives of Criminal Prosecutors." *American Journal of Political Science* 46 (2): 334–51.
Gordon, Sanford C., and Gregory A. Huber. 2009. "The Political Economy of Prosecution." *Annual Review of Law and Social Science* 5: 135–56.
Gordon, Sanford C., and Sidak Yntiso. 2022. "Incentive Effects of Recall Elections: Evidence from Criminal Sentencing in California Courts." *Journal of Politics* 84 (4): 1947–62.
Grogger, Jeffrey. 1991. "Certainty vs. Severity of Punishment." *Economic Inquiry* 29 (2): 297–309.
Grossman, Gene M., and Michael L. Katz. 1983. "Plea Bargaining and Social Welfare." *American Economic Review* 73 (4): 749–57.
Hollander-Blumoff, Rebecca. 1997. "Getting to Guilty: Plea Bargaining as Negotiation." *Harvard Negotiation Law Review* 2: 115–48.
Hübert, Ryan. 2019. "Getting Their Way: Bias and Deference to Trial Courts." *American Journal of Political Science* 63 (3): 706–18.
King, Nancy J., David A. Soule, Sara Steen, and Robert R. Weidner. 2005. "Panel One: Prosecutorial Discretion and Its Challenges." *Columbia Law Review* 105 (4): 959–1009.
Kutateladze, Besiki Luka, Nancy R. Andiloro, and Brian D. Johnson. 2016. "Opening Pandora's Box: How Does Defendant Race Influence Plea Bargaining?" *Justice Quarterly* 33 (3): 398–426.
Landes, William M. 1971. "An Economic Analysis of the Courts." *Journal of Law and Economics* 14 (1): 61–107.
Langer, Maximo. 2004. "From Legal Transplants to Legal Translations: The Globalization of Plea Bargaining and the Americanization Thesis in Criminal Procedure." *Harvard International Law Journal* 45: 1–64.
Lax, Jeffrey R., and Charles M. Cameron. 2007. "Bargaining and Opinion Assignment on the US Supreme Court." *Journal of Law, Economics, & Organization* 23 (2): 276–302.
Mas-Colell, Andreu, Michael Dennis Whinston, and Jerry R. Green. 1995. *Microeconomic Theory*, Vol. 1. New York: Oxford University Press.
Mata, Rui, Renato Frey, David Richter, Jürgen Schupp, and Ralph Hertwig. 2018. "Risk Preference: A View from Psychology." *Journal of Economic Perspectives* 32 (2): 155–72.
Moore, Don A., and Paul J. Healy. 2008. "The Trouble with Overconfidence." *Psychological Review* 115 (2): 502.
Mungan, Murat C. 2017. "The Certainty versus the Severity of Punishment, Repeat Offenders, and Stigmatization." *Economics Letters* 150: 126–29.
Polinsky, A. Mitchell, and Steven Shavell. 1999. "On the Disutility and Discounting of Imprisonment and the Theory of Deterrence." *Journal of Legal Studies* 28 (1): 1–16.
Sauer, Robert M. 1998. "Job Mobility and the Market for Lawyers." *Journal of Political Economy* 106 (1): 147–71.
Shotts, Kenneth W., and Alan E. Wiseman. 2010. "The Politics of Investigations and Regulatory Enforcement by Independent Agents and Cabinet Appointees." *Journal of Politics* 72 (1): 209–26.
Stuntz, William. 2011. *The Collapse of American Criminal Justice*. Cambridge, MA: Harvard University Press.
Turner, Jenia. 2010. *Plea Bargaining across Borders: Criminal Procedure*. New York: Aspen Publishers.