AI Enforcement:
Examining the Impact of AI on Judicial Fairness and Public Safety

Yi-Jen (Ian) Ho[1]
ian.ho@tulane.edu

Wael Jabr[2]
wjabr@psu.edu

Yifan Zhang[3]
yzhang60@kennesaw.edu

[1] Freeman School of Business, Tulane University, New Orleans, LA
[2] Smeal College of Business, Pennsylvania State University, University Park, PA
[3] College of Business, Kennesaw State University, Marietta, GA

Abstract

State judicial systems face the challenge of managing overwhelming prisoner populations and record-high incarceration costs. To be efficient, judicial systems are adopting artificial intelligence (AI) to assess offenders' recidivism risks and recommend alternative punishments (instead of incarceration) for low-risk offenders. However, the impacts of such AI initiates on judges' decision-making, offenders' fairness, and public safety remain unknown. We investigate the effects of AI recommendations on judges' sentencing decisions and the subsequent societal impact on public safety. Using a regression discontinuity design and unique data from 56,941 sentencing cases in Virginia, we first note that AI recommendations significantly increase the probability of offering alternative punishments, lower the probability of incarceration, and shorten the length of imprisonment. More importantly, we show that AI can promote or demote judicial fairness. While judges are more lenient toward females than males, AI helps alleviate such a gender-based difference. In addition, judges stay fair when sentencing risky offenders but give more favor to whites than blacks, both of whom receive AI alternative punishment recommendations. We last analyze the quality of judges' decisions regarding offenders' recidivism. The results indicate that judges' leniency towards females and whites and strictness towards males and blacks hurt public safety. We compile the results to provide actionable implications for the public, judges, and policymakers to promote judicial fairness with AI support.

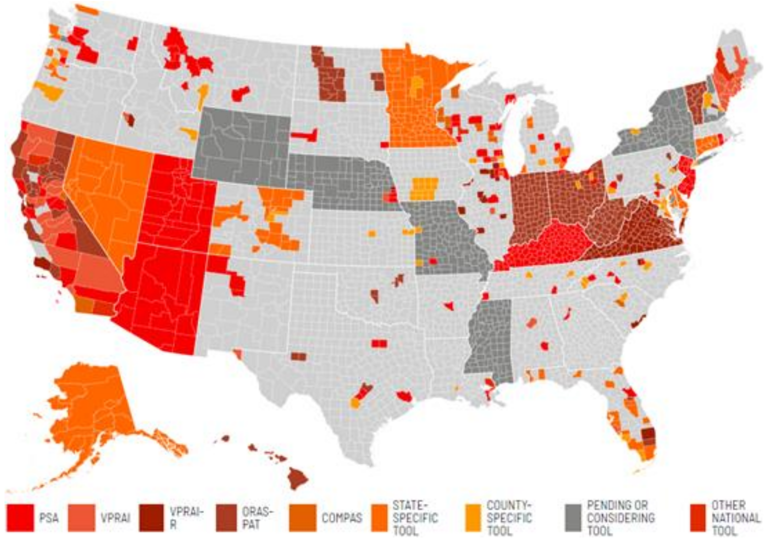
Keywords: *acritical intelligence*, *judicial system*, *sentencing*, *bias*, *regression discontinuity*

1. Introduction

State judicial systems face the ongoing challenge of an overwhelming incarceration population. By 20202, there have been over 1.1 million prisoners in county jails and state and federal prisons. Such a population is expected to grow exponentially in the next few years. Incarceration, the most standard correction format for crimes, costs more than 35 thousand a year to imprison an offender, making the country spend over 180 billion dollars on mass incarceration in 2022. Clearly, an effective solution is urgently wanted to lower the prisoner population and the cost of imprisonment operations. One possible option is imprisoning high-recidivism-risk (*risky*) offenders while sentencing the less-risky ones to alternative punishments (e.g., fines, electronic monitoring, and probations). This option requires a reliable instrument to assess offenders' recidivism risks. Thus, state judicial systems are motivated to adopt artificial intelligence (AI) to support risk assessments. Also, AI as an objective instrument is expected to enhance fairness in judicial systems, wherein judges excise significant autonomy to sentence offenders. The public remains concerned that they could still be biased despite their best efforts to evaluate offenders based on facts (e.g., case severity and criminal history). Given the above reasons, state judicial systems proactively make progress in adopting AI tools at different paces. Figure 1 illustrates different adoption strategies across the country.

Three concerns of judicial AI applications also arise along with the above-desired benefits. First, as a standard reference, AI may homogenize judges' decisions and convert their diversified opinions into a single voice, impairing judicial fairness insured by independent judgments. Second, AI could adversely amplify the existing biases (gender and race) within the system. For instance, a judge can benefit specific demographic segments by complying with AI when AI recommendations are aligned with her bias. Third, AI, by design, may provide misinformation if trained by biased data. Thus, it remains unclear how AI impacts judicial systems regarding judges' decisions, offenders' fairness, and the public's safety. Our society should have rigorous evidence to quantify the related impacts before taking action. Given the significance of sentence decisions, we address the following research questions: (1) *Would AI-backed recommendations affect judges' sentencing decisions* (*the probability of receiving alternative punishment*, *the probability of incarceration, and the length of imprisonment time)?* (2) *If yes, what characteristics of offenders (gender and race) and judges (experience) moderate the AI-affected decisions?* (3) *Could AI help improve long-term public safety with respect to recidivism?* The answers to these grounded questions shall provide clear guidance for policymakers to govern AI applications in the judiciary more effectively.

Figure 1. Adoption of Recidivism-Risk Assessment Tools across States[1]



We employ a regression discontinuity design (RDD) to identify the causal relationship between AI recommendations and outcome of interests using 5,1160 sentencing cases of drug, fraud, and larceny in Virginia since 2013 (see more in Section 3.1). The AI instrument considers various factors (including gender, age, and criminal history) to assess an offender's recidivism risk (as an integer risk score) and recommend offering alternative punishments (or not) to the designated judge in a sentencing trial. If the offender's risk score (e.g., 10) is below the threshold (e.g., 15), an AI recommendation for alternative punishments (RAP) is given in this case. As our RDD chooses a narrow, robust bandwidth *right* below and above the threshold (e.g., [12,14] and [15,17]), we can infer the desired casualty by comparing the

[1] The map shows the fragmentation of approaches toward risk assessments nationwide by 2020. Some states require either using or avoiding specific tools while some encourage and allow counties to choose preferred tools. For example, while states (e.g., AZ, NV, and UT) had statewide tools (e.g., PSA, VPRAI, or VPRAI-R), other states (e.g., CA) had county-level tools. See https://pretrialrisk.com/national-landscape/where-are-prai-being-used/.

offenders with and without RAPs (i.e., the treated and untreated samples).[2] In addition, we specify our RDD in a hierarchical Bayes manner of random coefficients. These random intercepts and slopes help better characterize the different natures and significant heterogeneities across judicial districts, if any.

Our results show three sets of interesting casual relationships. First, we discuss average treatment effects. The treated offenders (with RAPs) have a higher chance of receiving alternative punishments, a lower probability of incarceration, and shorter imprisonment. Quantitatively, AI recommendations lower the three decision outcomes by 16.2%, 5.2%, and 28% for drug-related cases, respectively. The cases of fraud and larceny maintain a similar causal pattern. Our empirics imply that AI helps free jail and prison spaces significantly at first glance.

Second, we explore heterogeneous treatment effects (HTEs) by considering the characteristics of stakeholders. In specific, we examine the moderating roles of (1) offenders' genders and races and (2) judges' experiences (i.e., tenure, the number of RAP-related cases, and compliance with RAPs). While sentence decisions must be based on objective facts, this is not always true. We notice that judges are more lenient to women (than men), offering them more alternative punishments, less incarceration, and shorter imprisonment time by 3.9%, 2.1%, and 16.3%, respectively. Such a bias could be unconscious but never compromised. Fortunately, we also find that AI helps judges correct the bias to treat both genders more equally. When considering offenders' races, an opposite pattern occurs. Judges generally sentence black offenders similarly to whites, but the latter benefit from RAPs significantly more than the former. In other words, AI may evoke racial bias since judges selectively apply objective recommendations to a specific demographic segment. This selection bias throws blacks in an unfavorable situation, wherein they suffer from a lower probability of receiving alternative punishments, a higher chase of incarceration, and longer imprisonment time by 5.6%, 4.3%, and 30.7%, respectively. As for the moderation of judges' experience, we find no significant evidence consistently confirming that the AI impact on decision-making does not vary across more and less-experienced judges.

[2] We select the optimal bandwidth based on the mean-square-error (MSE) approach (Imbens and Kalyanaraman, 2012).

Last, we focus on the ultimate impact of AI on recidivism (defined as an offender's consecutive conviction within three years after finishing her previous correction). To highlight the effectiveness of AI, we conduct several confusion-matrix-like analyses, wherein a 2x2 matrix includes the dimensions of AI recommendations and judges' decisions. Each dimension has a value of offering alternative punishments or not, resulting in four scenarios: (1) recommended by AI and offered by judges, (2) recommended but not offered, (3) not recommended but offered, and (4) neither recommended nor offered. Our empirics demonstrate the synergy between AI and judges if the two make consistent decisions (i.e., Scenarios (1) and (4)). However, when judges' decisions deviate from AI recommendations, the recidivism of Scenario (3), wherein judges give alternative punishments whereas AI recommends not, is significantly worse than its counterpart, Scenario (2). Such an undesirable outcome is amplified by offenders' genders and races, making the judges' subjective lenience to females and whites unworthy at all. Overall, our several results confirm the effectiveness of AI in direct sentences and ultimate societal outcomes.

Our research makes a series of contributions to two streams of literature, namely, AI and judicial systems. We fill the gap in the literature intersection of AI and judicial systems by studying the impacts of AI recommendations on sentencing. Deploying the rigorous RDD identification, we are among the first to quantify how AI affects judges' decision-making. Our results show that judges do incorporate AI recommendations to make their decisions. We not only notice that judges are lenient towards females but show how AI helps correct such behavioral bias and makes judges more objective. Yet, just when we feel relieved about judges' fairness for risky offenders across races, judges apply AI recommendations to offenders selectively and put blacks in an unflavored situation. We also document the beauty of human-AI synergy and the danger of human-AI discrepancy. While some studies concern AI biases, we worry much more about judges' subjectivity and autocracy to new technological aids. We firmly believe that these new inspirations and caveats of AI efficacy in the judicial context are grounded and significant to answer the debates of AI in law. Our results bring implications to the public, judges, and policymakers to build a fairer judicial system (see Section 6 for more implications).

The rest of this paper is organized as follows. We provide a brief on the literature in Section 2.

Section 3 introduces our empirical context and data operationalization. We specify the models in Section 4, while Section 5 discusses and interprets the results. Finally, we conclude this research with managerial implications.

2. Literature Background

It is an intensive debate whether the collaboration between humans and AI outperforms one alone (Luong et al. 2020a; Luong et al. 2020b). Some argue that humans prefer AI recommendations over humans (Bai et al. 2022), whereas others show that experts tend not to comply with what AI recommends (Caro and de Tejada Cuenca 2023). The latter happens when decision-makers incorporate private information (Xu et al. 2020). Following or deviating from AI recommendations leads to better or worse decisions, depending on the quality of recommendations and private information (Kesavan and Kushwaha 2020). Such uncertainty is further amplified in practicing law and judiciary. *The rule of law* disciplines judicial systems to maintain the value of equal treatment and fairness in the design and application of the law via diversified opinions (Gray 1999). Skeptics state that AI initiatives can threaten these fundamental values and ruin fairness across interest groups. In contrast, advocates believe that AI provides judges with objective clues to make bias-free decisions.

Despite the debate, AI has been increasingly utilized in judicial systems across different functions, wherein sentencing support is the most significant. Traditionally, judges make sentencing assessments in the pre-AI era using various information, such as hearings and impressionistic assessments. However, this information can be misleading and subject to subjective interpretations. AI and algorithms have thus been developed to support judges' decision-making. An iconic use case depicts that AI provides judges with offenders' predicted recidivism-risk scores in sentencing trials.

To characterize the nature of AI in law, the literature is interested in comparing humans and AI. Dressel and Farid (2018) devalue AI, but Lin et al. (2020) applaud AI. Recent studies focus on judges' AI adoption, as Garrett and Monahan (2020) note several reasons hindering AI adoption. The community is concerned about machine learning bias. Sargent and Weber (2021) note that using biased sentencing cases as the training data leads to biased predictive models and recommendations. Though prior studies explore various topics of judicial AI applications, the literature still has no clue about the relationship between AI and judges' decision-making and societal outcomes. This study attempts to contribute to the literature by identifying the causal impacts of AI on sentencing via rigorous empirical models.

3. Empirical Setting

3.1. Institutional Background

Virginia is one of the pioneer states in the country for adopting AI instruments in judicial systems. The AI instrument garners positive reception among judges. In the survey of Monahan, Metz, and Garrett (2018), 80% of judge participants agree that the instrument is helpful in reflecting offenders' recidivism risks and helping them make better sentencing. Half of the participants also state that they *almost always* refer to AI recommendations before finalizing decisions, and one-third of the judges *usually* consider them. To better understand our institutional background, we brief the development history below.

Virginia, in 1995, abolished parole and implemented the Truth-in-Sentencing (TIS) Act, which aimed to ensure that violent offenders serve a significant portion of their imposed sentences before being eligible for release. Due to the high fiscal cost of incarcerating violent offenders under the TIS Act, the Virginia Criminal Sentencing Commission (VCSC) was designated to build an AI-based risk-assessment instrument that identified non-violent offenders with the lowest recidivism risk to public safety. The first-generation risk assessment instrument was rolled out statewide in 2003 after the pilot implementation in six judicial districts between 1996 and 1999.

The first-generation instrument considers a comprehensive set of each offender's characteristics and criminal records to assess her recidivism risk.[3] The AI instrument assigned weights to calculate an overall recidivism risk score. In 2013, VCSC revamped the instrument using more advanced technologies. Compared with the previous generation (a one-for-all model), the new instrument selects different factors and varies corresponding weights to predict offenders' risk scores across offense types (i.e., drug, larceny,

[3] The factors include a focal offender's demographics (gender, age, employment, and marital status) and criminal history (offense type of the conviction, presence of additional offenses, recent arrests or periods of confinement, prior felony convictions or adjudications, and previous adult incarcerations.

and fraud). More importantly, AI now takes a more proactive role in supporting judges' decision-making by offering sentence recommendations. By choosing an optimal risk-score threshold for each offense type, AI provides judges with alternative-punishment recommendations for offenders whose scores are below thresholds. Of course, such a smarter instrument has made high-quality recommendations and substantial impacts on sentencing for over ten years, the above survey notes.

3.2. Data and Model-Free Evidence

We collect offense data from the VCSC, which consists of all felony convictions in Virginia between July 2013 and June 2022. Our dataset records many details of an offense case, including but not limited to the types of conviction charges, risk scores (based on the second-generation instrument), and final sentences. Offenders' races are crawled from the Virginia Judiciary Online Case Information System.[4] Our samples (56,941 non-violent cases) include 65.4% white and 34.6% black offenders. Table 1 details the average probabilities of offering alternative punishments ($ProbAltPuni$) and incarceration ($ProbIncar$) and the median length of imprisonment ($LengIncar$) by offense types, AI recommendations, and offenders' races.

Table 1. Summary Statistics of the Outcomes of Interest

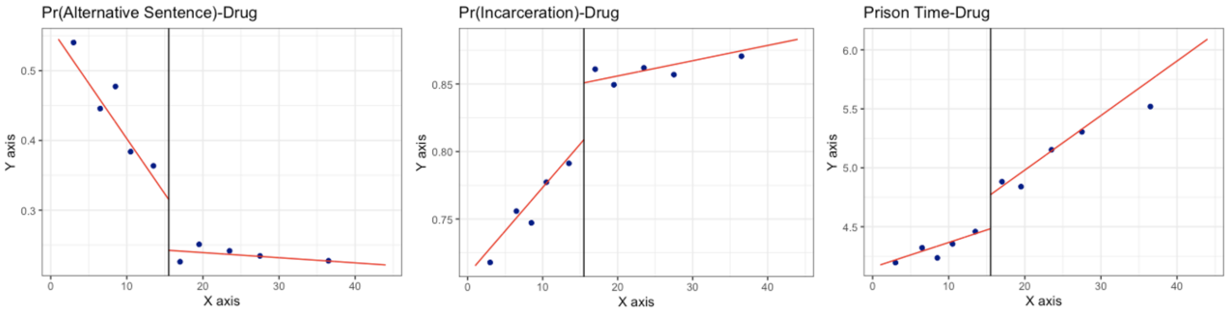
Offense	Drug		Fraud		Larceny	
AI Recommendation	Yes	No	Yes	No	Yes	No
White (65.4%)						
$ProbAltPuni$	43.4%	24.0%	62.8%	43.1%	24.9%	22.3%
$ProbIncar$	74.9%	85.7%	78.8%	89.2%	80.9%	88.8%
$LengIncar$ (months)	3.00	6.00	7.00	12.00	2.96	6.00
Black (34.6%)						
$ProbAltPuni$	40.5%	23.0%	67.8%	45.7%	23.4%	22.5%
$ProbIncar$	80.1%	86.6%	76.7%	91.6%	83.7%	89.3%
$LengIncar$ (months)	4.00	8.00	7.00	11.00	2.96	6.00

Overall, the impacts of AI recommendations vary across offense types due to the availability of alternative punishments and the nature of offenses. For example, while it is common to send drug addicts to rehab centers and fine frauds, nearly no alternative punishment option is available for larceny offenders. On average, 36.6% of drug offenders, 55.3% of fraud offenders, and 22.7% of larceny offenders receive

[4] See more information at https://eapps.courts.state.va.us/ocis/search.

alternative punishments. Figure 2 uses drug offenses to visualize how AI recommendations (the two sides of the line) affect judges' decisions (y-axis) over offenders' risk scores (x-axis). The distinctive "jumps" at the RAP thresholds suggest that AI plays a significant role in judges' decision-making.

Figure 2. Visual Check of the Treatment Shock



4. Model

The objectives of the empirical analyses are to identify and quantify (1) the impacts of the AI instrument on judges' sentencing decisions, (2) the moderation of offenders' characteristics on these AI-driven impacts, and (3) the consequential effects on recidivism. We start our model discussion with the direct impacts of AI on three decisions: the probability of offering alternative punishments ($ProbAltPuni$), the probability of offenders' incarceration ($ProbIncar$), and the length of imprisonment ($LengthIncar$). Our identification is based on the risk thresholds of offering recommendations for alternative punishments. Specifically, offenders with below-threshold risk scores receive RAPs, while those with above-threshold scores do not. Such a contextual setting enables a regression discontinuity design to identify the causality of interest. Our RDD compares the offenders with and without RAPs using a narrow, ensuring they are naturally similar. Let Y_i denote the outcomes of interest for Offender i, and we specify our RDD for the average treatment effects (ATEs) as:

$$Y_i = \alpha_0 + \boldsymbol{\alpha_1} Recommendation_i + \alpha_2 Risk_i + \alpha_3 Recommendation_i * Risk_i + \boldsymbol{AX}_i + \varepsilon_i, \tag{1}$$

where $Recommendation_i$ is the treatment indicator, wherein Offender i receives the recommendation for alternative punishments if the indicator takes a value of 1. $Risk_i$ is i's risk score calculated by AI, and the interaction captures the relationship between the treatment and running variable. $\boldsymbol{X}_i$ are the controls that include the focal offender's characteristics (e.g., risk score, gender, and race), the designated judge's experience, the details of the case, guideline-recommended sentence, and time-fixed effects. $\boldsymbol{\alpha_1}$ is the coefficient of interest, so significant $\boldsymbol{\alpha_1}$ suggests that AI RAPs do affect judges' decision-making.[5]

Next, we explore whether offenders' genders and races moderate the ATEs. These analyses are critical for detecting the existing unfairness, if any, and the role of AI in it. Will AI correct or amplify the existing unfairness? To answer this question, we extend our RDD by interacting the treatment indicator with Offenders i's gender and race as:

$$\begin{aligned} Y_i = \beta_0 + \beta_1 Recommendation_i + \beta_2 Gender_i + \beta_3 Race_i + \beta_4 Risk_i + \\ \boldsymbol{\beta_5} Recommendation_i * Gender_i + \boldsymbol{\beta_6} Recommendation_i * Race_i + \\ \beta_7 Recommendation_i * Risk_i + \boldsymbol{BX}_i + \varepsilon_i, \end{aligned} \tag{2}$$

where the offender is a male if $Gender_i$ = 1. $Race_i$ takes a value of 1 to indicate that the offender is black. $\boldsymbol{\beta_4}$ and $\boldsymbol{\beta_5}$ are the coefficients of interest, quantifying the additional AI effects on males and blacks on top of the baseline effects on females and whites. Similarly, we apply the same model specifications to study the moderating effects of a judge's experience regarding her tenure length and the number of RAP-related cases handled. It is worth noting that the courts in the state are organized into 31 judicial districts, which could vary in economic disparity, political ideology, and resource availability. Thus, we adopt a Bayesian hierarchical approach to capture the heterogeneity among judicial districts.

Last, we analyze the societal impact of AI by probing how its recommendations affect offenders' recidivism. Following the VCSC, we define recidivism as a consecutive felony conviction within three years after correction. To understand the relationships between recidivism and the parameters of interest, we have the following linear probably model:

$$\begin{aligned} Recidivism_i = \gamma_0 + \gamma_1 Recommendation_i + \gamma_2 AltPuni_i + \\ \boldsymbol{\gamma_3} Recommendation_i * AltPuni_i + \boldsymbol{\Gamma X}_i + \varepsilon_i, \end{aligned} \tag{3}$$

where $AltPuni_i$ indicates whether Offender i receives alternative punishments at the end. Once obtaining the estimates of Equation (3), we predict the probability of a focal offender's recidivism and contrast the impacts of AI under different scenarios across all offenders (e.g., AI recommends alternative punishments,

[5] A linear (probability) model is specified to estimate the impacts of AI on the outcomes. Our results stay robust while different specifications (e.g., probit and logistic models) are applied.

whereas judges do not offer them).

5. Results

5.1. Direct Impacts on Sentencing Decisions

We first discuss the average treatment effects of AI recommendations. Table 2 summarizes our regression discontinuity design estimates for three sentencing decisions (alternative punishments, incarceration, and imprisonment time) across three types of offenses (drug, fraud, and larceny). To mitigate the influence of extreme values on the results, we apply an inverse hyperbolic sine (arcsinh) transformation to the variable incarceration length (Burbidge et al. 1988). Our results note that AI recommendations influence judges' sentencing decisions significantly. For drug offenders, receiving a recommendation increases the probability of receiving alternative punishments by 16.2% and decreases the probability of incarceration by 5.2%, on average. When we focus on imprisonment time, offenders receiving RAPs get a shortened length of incarceration by 28.0%. Such a reduction is economically significant, equivalent to 18 days shorter than standard sentences. Similarly, for fraud offenders, receiving a recommendation lowers the odds of incarceration by 10.7% and reduces the imprisonment time by 65.3%. When we check larceny offenders, RAPs promote the probability of receiving alternative punishments by 6.1%,

Table 2. Average Treatment Effects of AI Recommendations

	$ProbAltPuni$	$ProbIncar$	$LengthIncar$
Drug (N=6,099)			
$Recommendation$	0.162 (0.028) ***	–0.052 (0.024) *	–0.284 (0.142) *
Fraud (N=2417)			
$Recommendation$	0.071 (0.047)	–0.107 (0.042) *	–0.653 (0.270) *
Larceny (N=8,895)			
$Recommendation$	0.061 (0.023) **	0.001 (0.021)	0.031 (0.053)

Note. ***, **, and * indicates 0.001, 0.01, and 0.05 levels of significance, respectively. The full results are upon request.

5.2. Heterogeneous Effects of Gender and Race

We further show how the direct impacts of AI are moderated by offenders' genders and races in Table 3.[6] We have a pair of good and bad news regarding the moderation of gender and race, respectively. First, the significant coefficient of $Gender_i$ suggests that judges are more lenient to female offenders. Among riskier offenders, females have a higher probability of receiving alternative punishments, lower odds of incarceration, and shorter imprisonment time than males by 4.7%, 4.2%, and 28.7%, respectively. These empirics find that judges are vulnerable to behavioral compassion and subjectivity and make unfair sentencing. Fortunately, our good news is that the significant marginal effects of AI on males ($Recommendations_i * Gender_i$) offset males' disadvantage in incarceration and imprisonment time meaningfully. It is encouraging to witness that AI recommendations awaken judges' consciousness and make them stay objective, especially when females are more likely to receive RAPs.[7]

Table 3. Heterogeneous Treatment Effects of AI Recommendations (Drug)

N=6,099	$ProbAltPuni$	$ProbIncar$	$LengthIncar$
$Recommendation$	0.167 (0.032) ***	–0.035 (0.028)	–0.191 (0.142)
$Gender$ (Male = 1)	–0.047 (0.016) **	0.042 (0.014) **	0.287 (0.086) ***
$Recomm * Gender$	0.024 (0.027)	–0.055 (0.024) *	–0.337 (0.144) *
$Race$ (Black = 1)	0.021 (0.013) +	–0.010 (0.012)	–0.093 (0.071)
$Recomm * Race$	–0.056 (0.023) *	0.043 (0.020) *	0.307 (0.120) *

Note. ***, **, *, and + indicate 0.001, 0.01, 0.05, and 0.1 levels of significance, respectively. The full results are upon request.

Judges do an excellent job of maintaining sentencing fairness regardless of offenders' skin color. The insignificant coefficient of $Race_i$ (the offender is black if $Race_i$= 1) supports this argument. Yet, we receive bad news that judges unfavorably discount AI recommendations for alternative punishments when sentencing black offenders. Conditional on the same probability of receiving RAPs across races, judges sentence blacks to fewer alternative punishments, more incarceration, and longer imprisonment than

[6] We stay focused on the results of drug offenses in the following discussion due to the page limit. The results of fraud and larceny cases are upon request. In general, a similar pattern is maintained. The result variations across offense types could come from either the different nature of offense types or the small-sample issue.

[7] The AI instrument by desgin specifically penalizes male offenders when assessing risk socres.

whites by 5.6%, 4.3%, and 30.7%, respectively.[8] The results reveal that objective AI can be adversely used as a discriminating tool, as Albright. (2022) show. We call on the public to send a caveat to judges.

We also explore the moderation of treatment effects by judges' experience in Table A.3. Our analyses find no evidence implying that judges do not vary in their usage of AI due to their experience or familiarity with AI recommendations. In addition, to capture the heterogeneity among judicial districts, we employ a Bayesian hierarchical model to obtain empirical regularities. Figure A.1 in Appendix C visualizes the estimated treatment effects and the corresponding 95% credible interval (i.e., the Bayesian equivalent of a confidence interval). We find significant differences in the treatment effects at the judicial district level.

In short, our findings highlight the significance of offenders' genders and races and geographic differences in moderating AI's impacts. Indeed, AI as a neutral device either enhances or hinders fairness at the same time, depending on how judges apply it. We must understand the societal moderation of the existing gender and racial biases on AI usage before making final judgments in judicial systems.

5.3. Ultimate Outcome of Recidivism

To accurately assess the societal impact of the AI-supported recommendation, we investigate its effect on recidivism. We exclude the cases in which offenders' correction periods are beyond December 31, 2017, while considering the time gaps between offenders' arrests and convictions and the pandemic interruption. This procedure results in a final sample of 25,770 cases, including 12,804 drug cases, 3,113 fraud cases, and 9,853 larceny cases. Table 4 compiles predictive offenders' recidivism into the scenarios of a 2x2 matrix using the estimates of Equation (3). The four scenarios are (1) recommended by AI and offered by judges, (2) recommended but not offered, (3) not recommended but offered, and (4) neither recommended nor offered. We find synergy (releasing less-risky offenders and imprisoning risky ones) if AI and judges make consistent decisions in Scenarios (1) and (4). However, the discrepancy between AI and judges in the other scenarios leads to undesirable outcomes, especially when judges release the offenders predicted

[8] We find no significant difference in the proabiltiy of receving RAPs across blacks and whites.

to have higher recidivism risks in Scenario (3). It is also worth noting that judges do use their expertise to identify additional riskier offenders from the recommended pool when we compare Scenarios (1) and (2).

Table 4. Overall Predicted Recidivism

		Judges	
		Offered	Not Offered
AI	Recomm	14.11%	17.32%
	Not Recomm	25.71%	25.52%

Table 5 decomposes the overall recidivism by offenders' genders and races in Panels (a) and (b), respectively. In Panel (a), the results across males and females maintain the above pattern in Table 4. Let us focus on the matrix of females, the highest recidivism rate (25.17%) does not take place in Scenario (4) but (2), wherein judges insist on releasing the offenders who are supposed to be incarcerated. Such an astonishing result echoes judges' improper lenience towards females, discussed earlier. The outcome could be even worse since females are more like to receive RAPs than males. As for offenders' races, we document judges' unfair treatment of blacks by comparing Scenario (3) across two matrices in Panel (b). Judges punish the blacks who are recommended alternative punishments much more severer than their white counterparts, even though the two have the same probability of receiving RAPs. Only 13.78% of blacks commit consecutive offenses in the post-correction three years, compared with 18.10% of whites.

Table 5. Predicted Recidivism Heterogeneity

(a) by Gender

Female		Judges	
		Offered	Not Offered
AI	Recomm	14.57%	18.47%
	Not Recomm	25.17%	22.12%

Male		Judges	
		Offered	Not Offered
AI	Recomm	16.09%	19.01%
	Not Recomm	28.12%	28.56%

(b) by Race

Black		Judges	
		Offered	Not Offered
AI	Recomm	12.82%	13.78%
	Not Recomm	24.14%	25.93%

White		Judges	
		Offered	Not Offered
AI	Recomm	13.82%	18.10%
	Not Recomm	23.69%	25.11%

6. Conclusion

AI, as we witness, is increasingly important in every aspect of judicial systems, ranging from automation to decision-making support. Still, the impact of AI on short-term judges' decisions and long-term public safety remains unclear. This study has identified the rigorous causal link between AI recommendations and sentencing outcomes. Our empirics have shown that AI lowers not only the overwhelming prisoner populations and incarceration costs but also recidivism by making decent recommendations. We have also noticed that AI can either enhance or hinder pursued fairness across different minority groups. It is worth noting that the discrepancy between judges and AI often leads to the most unfavorable outcomes. We thus like to provide the following implications for the public, judges, and policymakers.

First, the public should keep open-minded to discussing the potential benefits and drawbacks of judicial AI applications. Unfortunately, judgmental calls are made before the understanding of ongoing AI initiatives. By proactively seeking hard evidence like ours, citizens will be more objective to advocate or skepticize these AI projects. After all, the effectiveness of AI will be maximized only if the public has a consensus and is willing to try and accept it.

Second, judges need to be aware of their own behavioral tendencies and biases. They have treated specific gender and race groups (e.g., male vs. female and white vs. black) differently for a long time for a long time. Though AI can provide decent recommendations based on facts, a fair decision-making process is still the judges' call. Indeed, we are pleased that AI helps correct judges' leniency toward females (or strictness toward males) but worried that judges use AI to benefit whites (or disadvantaged blacks). The two facts jointly suggest that judges may suffer from unconscious biases, whereas AI recommendations help them refocus on objective judgments. As a result, we encourage judges to minimize intrinsic bias by pausing and reconsidering when their decisions deviate from AI recommendations. In addition, judges can proactively provide policymakers with helpful feedback to improve the AI system since they are the *users*. For example, judges can detail what drives them to go in the opposite direction from AI or share their private information. This information can be effectively incorporated into AI via a virtuous feedback loop, nourishing the trust between judges and AI.

Last, policymakers ought to pay extra attention to system adoption strategies. Just as firms adopt enterprise systems, friendly training programs and communication channels should be provided. Without the two, the effectiveness of AI may never come out. Indeed, we find no impact of AI recommendations in Pennsylvania, wherein a similar risk-assessment system was rolled out in 2020. Due to the outbreak of COVID-19, training programs and communication channels did not get judges' buy-in. More importantly, policymakers should evaluate AI accuracy and societal impacts often. Virginia has not evaluated its AI instrument for more than ten years. Given the dynamics of society, we sincerely ask policymakers to take a proactive role in building a modern, fair judicial system with the support of AI.

Appendix A. Selected Reference

Albright A (2022). No Money Bail, No Problems? Trade-offs in a Pretrial Automatic Release Program. *Working paper*, Harvard University, Boston, MA.

Bai B, Dai H, Zhang DJ, Zhang F, Hu H (2022) The Impacts of Algorithmic Work Assignment on Fairness Perceptions and Productivity: Evidence from Field Experiments. *Manufacturing and Service Operations Management*, **24**(6), pp. 3060-3078.

Burbidge JB, Magee L, Robb AL (1988). Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association*, **83**(401), pp. 123-127.

Caro F, de Tejada Cuenca, AS (2023) Believing in Analytics: Managers' Adherence to Price Recommendations from a Dss. *Manufacturing and Service Operations Management*, forthcoming.

Dressel J, Farid H (2018) The Accuracy, Fairness, and Limits of Predicting Recidivism. *Science Advances* **4**(1), pp. eaao5580.

Garrett BL, Monahan J (2020) Judging Risk. *California Law Review* 108, pp. 439.

Gray CB (1999) *The Philosophy of Law: An Encyclopedia* (Vol 1). Taylor & Francis.

Imbens G, Kalyanaraman K (2012) Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, **79**(3), pp. 933-959.

Kesavan S, Kushwaha T (2020) Field Experiment on the Profit Implications of Merchants' Discretionary Power to Override Data-Driven Decision-Making Tools. *Management Science*, **66**(11), pp. 5182-5190.

Lin ZJ, Jung J, Goel S, Skeem J (2020) The Limits of Human Predictions of Recidivism. *Science Advances* **6**(7), pp. eaaz0652.

Luong A, Kumar N, Lang KR (2020a) Algorithmic Decision-Making: Examining the Interplay of People, Technology, and Organizational Practices through an Economic Experiment. Working paper, University of Warwick, Warwick, UK.

Luong A, Kumar N, Lang KR (2020b) Human–Machine Collaboration and Algorithmic Decision-Making in Organizations: Examining the Impact of Algorithm Prediction Bias on Decision Bias and Perceived Fairness. Working paper, University of Warwick, Warwick, UK.

Sargent J, Weber M (2021) Identifying Biases in Legal Data: An Algorithmic Fairness Perspective. *Working paper*, University of Michigan, Ann Arbor, MI.

Xu R, Cui P, Kuang K, Li B, Zhou L, Shen Z, Cui W (2020) Algorithmic Decision Making with Conditional Fairness. *Proceedings of 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2125-2135.

Appendix B. Full Results

Figure 2. Visual Check of the Treatment Shock

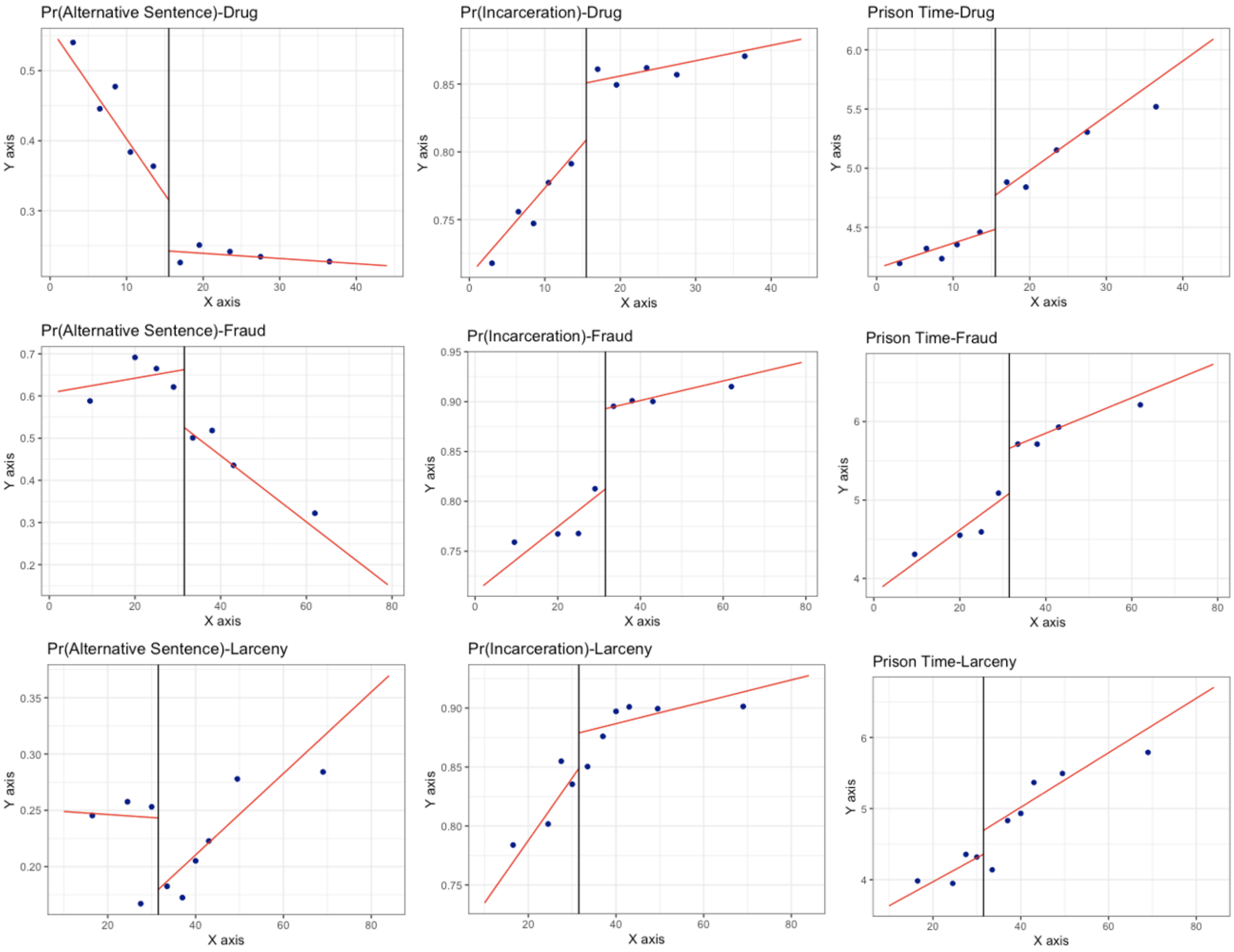


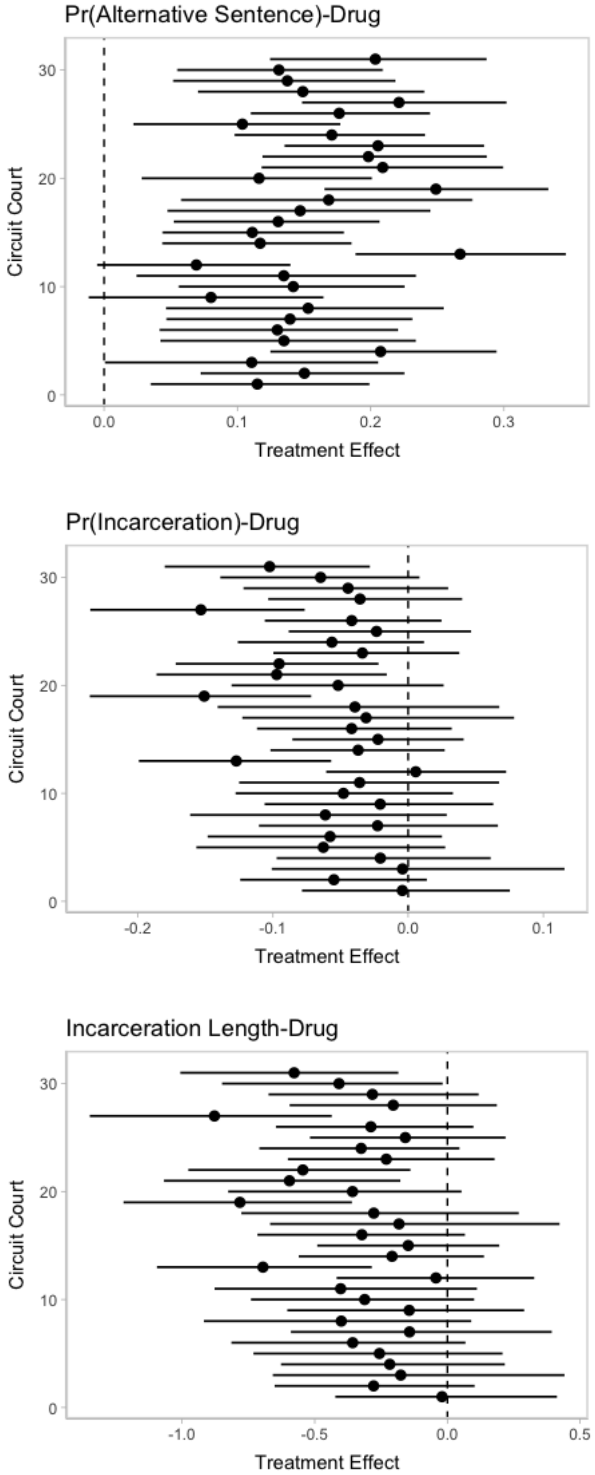
Table 3. Heterogeneous Treatment Effects of AI Recommendations (Drug)

N=6,099	*ProbAltPuni*	*ProbIncar*	*LengthIncar*
Intercept	0.103 (0.017) ***	0.824 (0.016) ***	3.927 (0.096) ***
Recommendation	0.167 (0.032) ***	–0.035 (0.028)	–0.191 (0.142)
Risk	** 0.014 (0.004) **	–0.006 (0.004) *	–0.046 (0.024) *
Recomm ∗ *Risk*	0.008 (0.012)	0.002 (0.011)	0.033 (0.063)
Gender (Male = 1)	–0.047 (0.016) **	0.042 (0.014) **	0.287 (0.086) ***
Recomm ∗ *Gender*	0.024 (0.027)	–0.055 (0.024) *	–0.337 (0.144) *
Race (Black = 1)	0.021 (0.013) +	–0.010 (0.012)	–0.093 (0.071)
Recomm ∗ *Race*	–0.056 (0.023) *	0.043 (0.020) *	0.307 (0.120) *
RecommPrisonTime	–0.532 (0.026) ***	0.054 (0.019) **	1.334 (0.129) ***
FirstTime	0.648 (0.028) ***	–0.650 (0.027) ***	–3.170 (0.130) ***

Note. ***, **, *, and + indicate 0.001, 0.01, 0.05, and 0.1 levels of significance, respectively. The full results are upon request.

Appendix C. Additional Results

Table A.1. District Heterogeneity (Drug)



Pr(Alternative Sentence)-Drug

Pr(Incarceration)-Drug

Incarceration Length-Drug