Can AI Help Reduce Human Bias? Insights from Police Rearrest Predictions

YONG SUK LEE, University of Notre Dame, USA

Reference Format:
Yong Suk Lee. 2025. Can AI Help Reduce Human Bias? Insights from Police Rearrest Predictions. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 6 pages.

1 Introduction

The integration of machine learning (ML) algorithms into public policy areas such as criminal justice, social welfare, immigration, and homeland security has made concerns about bias more acute and the deliberation on algorithmic fairness increasingly important. Consider algorithmic bias in the context of criminal justice, where a judge must decide whether to incarcerate a defendant awaiting trial, or imagine a homeland security agent using AI software to help determine entry or deportation. A biased algorithm increases the likelihood of errors, such as wrongfully detaining a defendant, denying entry at the border, or deporting a resident.

It is helpful to be precise about where bias emerges in these contexts. First, there is bias in the algorithm's predictions. As documented extensively in the literature, this bias results from biased training data—group differences stemming from implicit or explicit discrimination present in our society. In a hypothetical scenario where the algorithm makes the final decision without human intervention, this data bias would be the ultimate source of bias. However, typically, we prefer to keep humans involved in making final decisions. In those scenarios the human becomes aware of the AI's prediction or suggestion but makes the ultimate decision.

In this context, another source of bias arises if humans differentially accept the algorithm's recommendations based on specific features of the person being evaluated. For instance, an algorithm might suggest that a defendant poses a high flight risk and recommend detention until trial. The magistrate can accept or overrule this recommendation. If the decision to accept or overrule varies due to specific individual characteristics, another source of bias emerges. This paper focuses specifically on this latter form of bias—the bias introduced by humans in accepting or utilizing algorithmic predictions.

If humans making decisions are more biased than algorithms in contexts such as criminal justice, adopting AI algorithms has the potential to mitigate human biases and produce relatively fairer outcomes. In fact, risk assessment tools are increasingly adopted for tasks such as recidivism prediction, risk assessment, and sentencing decisions, and are often presented as neutral, data-driven solutions to entrenched problems of fairness and efficiency. However, their widespread adoption has raised critical questions about the reliability, equity, and societal consequences of algorithmic decision-making. A particularly significant concern is racial bias within the criminal justice system,

Author's Contact Information: Yong Suk Lee, University of Notre Dame, Notre Dame, USA, yong.s.lee@nd.edu.

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

© 2025 Copyright held by the owner/author(s).

where predictive algorithms are widely used to assess crime-related risks, although public awareness of this issue remains limited. Recent studies provide insights into the opportunities and challenges associated with predictive algorithms in criminal justice. For example, Miron et al. (2020) demonstrated that ML models could outperform structured professional tools in predicting juvenile recidivism. However, their findings also highlighted trade-offs, as the algorithms failed to satisfy key fairness metrics, particularly for marginalized groups such as women and foreigners [18]. Similarly, Biswas et al. (2020) examined the COMPAS risk assessment tool, finding that while human predictions were less accurate, they achieved greater fairness in terms of equalized odds [3]. These findings underscore the dual challenge of designing algorithms that enhance predictive performance without exacerbating existing social inequities. More specifically related to race, the results are also much nuanced. While some studies have found that these algorithms are biased against Black defendants [1], [2], others find that they can lower incarceration rates without increasing crime levels and racial bias [13].

This short paper introduces the findings of Lee (2025) that examines the racial implications of police interaction with predictive algorithms, particularly in the context of racial disparities in rearrest predictions in the United States [16]. Although the data come from the US, algorithmic prediction and racial bias are an issue that is directly relevant to the EU. Lee (2025) examines police officers' responses to algorithm-generated rearrest risk assessments. He conducted an experiment where police officers were shown the profiles of young offenders and were asked to predict each offender's rearrest probability within three years, both before and after being shown the algorithm's prediction. The experiment varied the visibility of the offender's race to the officers and also experimented with informing the officers of the model's accuracy.

Lee (2025) finds that when the race of the offender is disclosed, officers tend to adjust their predictions towards the algorithm's assessment. However, the adjustments made by the officers showed significant racial disparities: there was a noticeable gap in initial rearrest predictions between Black and White offenders, even when controlling for the characteristics of the offenders. The police tended to predict higher rearrest rates for Black offenders only when race was visible, but reduced their predictions after seeing the algorithm's assessment. However, not all police officers reduced their predictions after seeing the algorithm's predictions. Only Black police officers made significant downward adjustments following the algorithm's prediction, while White police officers did not significantly alter their assessments.

Algorithms may have the potential to help reduce the racial gap in rearrest predictions, but the findings show that the implications are much more nuanced and that algorithm-assisted human adjustment systems may be sub-optimal in moving towards social justice. The differential responses by Black and White police officers underscores the racial bias present across different demographics, as well as the partial capacity of AI-assisted human judgement systems to mitigate bias in human recidivism predictions.

Such findings contribute to the growing body of literature on algorithmic fairness and decision-making in criminal justice [18], [3], [1], [14], [20], by examining how police officers respond to algorithmic predictions of recidivism, and provides a nuanced understanding of the potential and limitations of algorithm-assisted decision making. Studies have examined the biases inherent in AI algorithms due to training data and how such biases compare to human biases [15], [1], [11], [21], [22]. However, how the different degrees of human interaction with algorithms can affect bias has been less studied [10], [6], [23].

The debate continues on whether decision making that involves complete automation through AI algorithms could be more or less optimal compared to human decision making with AI assistance [8], [22], [19]. Future

AI systems may surpass humans or even AI-augmented human decision-making systems in terms of predictive performance as well as fairness. But even if future systems devoid of human involvement perform better and exhibit less bias than those with humans in the loop, we must consider the potential risks associated with delegating decision-making authority as AI systems rapidly advance [10], [8].

While algorithms are sometimes viewed as tools to mitigate bias, the study by Lee (2025) underscores the need for caution and critical engagement. Algorithms are not neutral; they are shaped by the data, assumptions, and contexts in which they are developed and deployed. Without addressing the structural inequities embedded in historical data and decision-making processes, these tools risk perpetuating or even amplifying existing disparities. Moreover, the effectiveness of algorithmic interventions depends not only on their design but also on how they are interpreted and acted upon by human decision-makers. These insights highlight the importance of considering both technical and sociopolitical dimensions in the development and evaluation of algorithmic systems. In the following section, I describe the experimental design and findings of Lee (2025) in more detail and then conclude with discussions of the implications, limitations, and future research directions.

2 Data and Model Training

The paper involves data construction of the arrest and re-arrest profiles using the US National Longitudinal Survey of Youth 1997 dataset (NLSY97) [4], machine learning analysis on the constructed data, and the online survey experiment of police officers using the NLSY97 arrest profiles and the algorithm's prediction of each profile. The NLSY97 is a nationally representative sample of individuals who were born between 1980 and 1984 when the survey began in 1997. NLSY97 provides extensive information about a range of characteristics and experiences for each individual over time including arrests.

The author trains a classification model that predicts the probability that an individual will be rearrested in the next three years. To train the model, a dataset that is independent of the data used to generate the vignettes presented in the survey is used. This training data contains labeled observations for 800 individuals, and is randomly split into 72% data for training, 18% data for validation and 10% data for testing. The dataset is highly imbalanced across classes, with only 25.4% of observations related to individuals who were rearrested (positive label), and the rest belonging to individuals who did not (negative label). For this reason, the AUC score – a common measure of performance for classification models – is used to tune models on the validation set and evaluate model performance on the test set. After experimenting with several model architectures including logistic regression, random forest and feedforward neural network, the logistic regression model was chosen to generate the probabilities used in the experimental part of the research.

3 The Experiment and Key Findings

The online survey experiment investigates the effect of race information, AI predictions, and accuracy rate treatments on rearrest predictions by police officers. The author recruited police officers through Qualtrics, which vets the veracity of their panel participants. To further verify survey participants as law enforcement officers, respondents were asked to list their job title as well as describe their main activities. Most of the descriptions were in line with police and law enforcement tasks. After dropping a few responses that were done in too short of a time frame, there

were 156 police officers. Each police officer was randomized into three different groups with varying degrees of race information. The three Race Information Treatment Groups were:

1. Race group: This group is exposed to profiles with race information and comprises the main treatment group,
2. No-race group: This group is exposed to offender profiles without race information,
3. Mixed group: This group is exposed to a mixture of both types of profiles (with and without race information).

Officers in the mixed group would have naturally realized that the survey was examining how their predictions change when offender race was revealed. The vignettes were designed so that race was salient (capitalized) and included in the first sentence.

There were two phases in the experiment. In the first phase, 10 random profiles from a pool of 45 profiles were presented to each participant. The police were asked to predict the probability of rearrest within 3 years for each profile. Then they were shown the model's rearrest prediction for that profile and asked to make predictions again.

In the second phase, 10 random profiles from a different pool of 45 profiles were presented to the same participant. Again, they were first asked to predict rearrest probability for each profile and then again asked to make predictions after seeing the AI's prediction, but this time while being informed of the accuracy rate of the model's prediction by race. In summary, the police were randomly allocated to three groups – with race information, without race information, and mixed (with and without) race information – and made rearrest predictions on 10 offender profiles and then again after being informed of the algorithm's prediction. Each respondent proceeded to another batch of 10 profiles and performed the same exercise but in the second batch were informed of the accuracy rates of the model's prediction by race. Hence, each respondent examined 20 profiles and made 40 predictions.

The police were asked to make predictions on the probability of rearrest between 0 and 100%. On average, the first police prediction is 37.83% and the second prediction, i.e., after seeing the AI's prediction score is 36.01% indicating a slight but statistically significant decrease (-1.82) from the first to the second prediction. The average algorithm prediction is considerably lower than the police predictions at 23.82%. The police on average lowered their predictions after seeing the algorithm's lower predictions. The adjustment in police predictions is noticeable among participants in the group where race was revealed. All other aspects of the profiles were identical. In the group where profile race was hidden, the police did not downgrade their priors. In the mixed group, where the police would have been aware that some profiles had race information and others did not, the magnitude of downgrade is smaller compared to participants in the group where race was shown for all profiles. The findings indicate that the information provided by the algorithm influenced participant rearrest predictions when race was shown. If participants were adjusting their beliefs based on race, could the adjustments have varied by race, and if so did providing the algorithm's prediction mitigate or exacerbate racial bias in rearrest predictions? This is one of the main questions examined in the paper.

Indeed, in the context of rearrest predictions the author found that police participants were more biased against Black offenders compared to the algorithm. For otherwise observationally equivalent profiles the police predicted that rearrest probability would on average be higher by about 4.7% points for a Black offender compared to a White offender. A significant amount given that average predictions were 36%. In such contexts where bias exists, if police officers are allowed to update their predictions based on the information provided by an AI's prediction, then introducing AI could help mitigate the police's preconceived bias. This indeed was the case. The Black-White gap in rearrest predictions declined with AI's assessment. This finding alone might paint an optimistic view of how AI can help mitigate human bias in contexts like criminal justice.

However, a closer examination into the race of the police officers reveals that the findings are much more nuanced. Though both White and Black officers are similarly biased against Black offenders, the Black police officers are the ones adjusting their beliefs. The White police officers do not adjust their priors after being informed of the AI's predictions. Hence, the potential for AI to mitigate bias in recidivism predictions is only partial. As such, the claim that AI mitigates racial bias requires careful reevaluation. The limited change in predictions by White officers point to significant limitations in the broader applicability of these tools for reducing systemic inequities. These findings highlight the need for caution in asserting that algorithms can broadly reduce racism and discrimination without addressing the structural factors that shape human judgment.

4 Discussion

In criminal justice where human bias against certain minority groups is formed over a long period of time, prior research has pointed out that the prediction of the rearrest, flight, or crime risk by AI algorithms may be less biased than human predictions [5], [13], [12]. Indeed, in the context of rearrest predictions Lee (2025) find that police participants were more biased against Black offenders compared to the algorithm. For otherwise observationally equivalent profiles the police predicted that rearrest probability would on average be higher by about 4.7% points for a Black offender compared to a White offender. In such contexts where bias exists, if police officers are allowed to update their predictions based on the information provided by an AI's prediction, then introducing AI could help mitigate the police's preconceived bias. Lee (2025) indeed find this to be the case. The Black-White gap in rearrest predictions decline with AI's assessment. This finding alone might paint an optimistic view of how AI can help mitigate human bias in contexts like criminal justice. However, a closer examination into the race of the police officers reveals a much more nuanced finding. Though both White and Black officers are similarly biased against Black offenders, the Black police officers are the ones adjusting their beliefs. The White police officers do not adjust their priors after being informed of the AI's predictions. Hence, the potential for AI to mitigate bias in recidivism predictions is only partial. As such, the claim that AI mitigates racial bias requires careful reevaluation. These findings highlight the need for caution in asserting that algorithms can broadly reduce racism and discrimination without addressing the structural factors that shape human judgment.

The study is, of course, not without limitations. The sample of police that participated in the survey experiment is by no means representative of all law enforcement officers in the US. As such, there is no guarantee that the results from the Qualtrics sample would replicate in the population. However, this study is an improvement over other well-known studies of recidivism predictions that conduct online survey experiments using participants recruited from Amazon Turk [9], [17]. As with any online survey experiments, whether real world behavior would have been similar to online responses is an open question. However, as long as intent is correctly captured through online survey experiments, the findings on intent would serve as precursors to behavior [7]. Future research would benefit from conducting experiments directly with the stakeholders. Finally, The findings are specific to the context of rearrest predictions, and can differ in other contexts, for example, hiring or mortgage lending, and other cultural contexts.

In conclusion, while this study provides important insights into the potential and limitations of algorithm-assisted decision-making, it also emphasizes the need for caution and critical engagement in deploying these tools. By addressing the identified limitations and involving a broader range of stakeholders, future work can contribute to developing more equitable and effective systems for the criminal justice context.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. propublica (2016). *URL: https://www. propublica. org/article/machine-bias-risk-asses sments-in-criminal-sentencing* (2016).

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.

[3] Arpita Biswas, Marta Kolczynska, Saana Rantanen, and Polina Rozenshtein. 2020. The Role of In-Group Bias and Balanced Data: A Comparison of Human and Machine Recidivism Risk Predictions. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies* (Ecuador) *(COMPASS '20)*. Association for Computing Machinery, New York, NY, USA, 97–104. https://doi.org/10.1145/3378393.3402507

[4] Bureau of Labor Statistics. 2019. National Longitudinal Survey of Youth 1997. https://www.bls.gov/nls/nlsy97.htm. Accessed: date-of-access.

[5] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[6] Lorrie F Cranor. 2008. A framework for reasoning about the human in the loop. (2008).

[7] Mariano-Florentino Cuéllar, Benjamin Larsen, Yong Suk Lee, and Michael Webb. 2022. Does Information About AI Regulation Change Manager Evaluation of Ethical Concerns and Intent to Adopt AI? *The Journal of Law, Economics, and Organization* (2022), ewac004.

[8] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[9] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[10] Andreas Fügener, Jörn Grahl, Alok Gupta, and Wolfgang Ketter. 2021. Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *Management Information Systems Quarterly (MISQ)-Vol* 45 (2021).

[11] Karen Hao and Jonathan Stray. 2019. Can you make AI fairer than a judge? Play our courtroom algorithm game. *MIT Technology Review* (2019).

[12] Jon Kleinberg. 2018. Inherent trade-offs in algorithmic fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*. 40–40.

[13] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[14] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic fairness. In *Aea papers and proceedings*, Vol. 108. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 22–27.

[15] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. 2018. Discrimination in the Age of Algorithms. *Journal of Legal Analysis* 10 (2018), 113–174.

[16] Yong Suk Lee. 2025. Racial Implications of Police-Algorithm Interactions: Evidence from Rearrest Predictions. *CESifo Working Paper No. 11877* (2025).

[17] Zhiyuan "Jerry" Lin, Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. The limits of human predictions of recidivism. *Science advances* 6, 7 (2020), eaaz0652.

[18] Tolan S. Gómez E. et al. Miron, M. 2021. Evaluating causes of algorithmic bias in juvenile criminal recidivism. *Artif Intell Law* (2021), 111–147.

[19] Iyad Rahwan. 2018. Society-in-the-loop: programming the algorithmic social contract. *Ethics and information technology* 20, 1 (2018), 5–14.

[20] Arul George Scaria, Vidya Subramanian, Nevin K. George, and Nandana Sengupta. 2024. Algorithms and Recidivism: A Multi-disciplinary Systematic Review. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (Oct. 2024), 1292–1305. https://doi.org/10.1609/aies.v7i1.31724

[21] Cass R Sunstein. 2019. Algorithms, correcting biases. *Social Research: An International Quarterly* 86, 2 (2019), 499–511.

[22] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* 135 (2022), 364–381.

[23] Fabio Massimo Zanzotto. 2019. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research* 64 (2019), 243–252.