THE DIGITAL "TO KILL A MOCKINGBIRD": ARTIFICIAL INTELLIGENCE BIASES IN COURTS

VERA LÚCIA RAPOSO*

ABSTRACT

This Paper addresses the use of artificial intelligence ("AI") in the judicial system, specifically its application in making predictions that influence judicial decisions, and the legal and ethical concerns stemming from AI biases. First, this Paper will explore the various roles that AI can play in this domain, with a particular emphasis on AI in risk assessment and recidivism prediction. This involves analyzing data related to a crime and a defendant to generate predictions. A significant concern in this area revolves around biases. While biases in criminal justice have long been recognized as a critical issue, there has been optimism that AI could mitigate these biases. However, this may not necessarily be the case. There is potential for risk assessment AI to enhance sentencing accuracy and reduce human error and bias.

* Vera Lúcia Raposo earned her law degree from the Faculty of Law at the University of Coimbra, where she also completed her postgraduate studies in medical law and obtained her master's and doctoral degrees in legal and political sciences. She is an Assistant Professor of Law and Technology and Vice-Director for Value Creation and Internationalization at Nova School of Law/Faculty of Law, NOVA University Lisbon. She also served as a supervisor for postgraduate studies at the Centre for Medical Ethics and Law at the University of Hong Kong (China) and as a guest lecturer at the School of Law, National Yang Ming Chiao Tung University in Taiwan. She has taught at University of Macau (China), the University of Coimbra (Portugal), and Agostinho Neto University (Angola). She has also worked as a lawyer at the Vieira de Almeida e Associados law firm in Lisbon, specializing in health law and privacy law. She is a member of the Executive Committee of the WhatNext.Law research centre as a researcher and leads the FutureHealth research line at What-Next.Law, which focuses on the use of new technologies in health, medicine, and the human body. She is an active member of the European Association of Health Law and a Governor of the World Association for Medical Law. In 2024, she became a fellow of the prestigious Hastings Centre.

However, there is apprehension that it could perpetuate or exacerbate existing biases and even undermine fundamental principles of fairness in the justice system. Several factors contribute to this risk, including biased coding and incomplete and inaccurate training and testing datasets. Additionally, the presence of dynamic algorithms and the lack of transparency and explainability make it challenging to identify and address biases effectively. The forthcoming European regulation on artificial intelligence, known as the AI Act ("AIA"), aims to mitigate biases. However, it is acknowledged by experts that biases cannot be completely eradicated,[1] like biases that are inherent in human decisions.

TABLE OF CONTENTS

1. *See, e.g.*, Ruby Isley, *Algorithmic Bias and Its Implications: How to Maintain Ethics through AI Governance*., N.Y.U. AM. PUB. POL'Y REV. (Oct. 30, 2022), https://nyuappr.pubpub.org/pub/61cuny79/release/2.

INTRODUCTION

In *To Kill a Mockingbird*, the novelist Harper Lee illustrates the prevalence of racial injustice and societal prejudices, particularly in the judicial system of the 1930s American South (and American society as a whole).[2] The drama of the story is underpinned by judges letting their racial biases unjustly influence their rulings.[3]

Almost a century later, the same concerns arise, not because of human biases—which certainly have not disappeared—but due to technology-fueled AI biases. This paper covers the legal and ethical concerns related to AI biases when using predictive AI to support judicial decisions.

I. THE ROLE OF ARTIFICIAL INTELLIGENCE IN CRIMINAL JUSTICE

The justice system is currently undergoing a transformation fueled by technology,[4] and this evolution is evident in three key ways. Firstly, technology operates as a support system, providing information, assistance, and guidance to individuals within the justice system.[5] Secondly, technology can replace tasks traditionally performed by humans.[6] Lastly, at a more advanced level, technology may revolutionize how judges operate, bringing about distinct forms of justice through disruptive technology.[7] Concerns arise within these latter two methods

2. HARPER LEE, TO KILL A MOCKINGBIRD (1960).

3. *Id.*

4. For a discussion on the potential impact of AI on the legal system, *see* Andreia Martinho, *Surveying Judges About Artificial Intelligence: Profession, Judicial Adjudication, and Legal Principles*, 2024 A.I. & SOC'Y.

5. *See* Mahesh Rengaswamy, *How Technology Is Modernizing the Court System and Enabling Access to Justice*, THOMSON REUTERS, https://www.thomsonreuters.com/en/careers/careers-blog/how-technology-is-modernizing-court-system-and-enabling-access-to-justice.html (last visited Mar. 10, 2024).

6. *See* DENNIS D. DRAEGER, DEP'T JUST. CAN., JUSTICE TRENDS 2: AUTOMATED JUSTICE GET THE GIST OF THE FUTURE FOR TECHNOLOGY IN JUSTICE 13 (2018), https://www.justice.gc.ca/eng/rp-pr/jr/jt2-tmj2/jt2-tmj2.pdf (discussing technology and the legal field).

7. *See* Tania Michelle Sourdin, *Justice and Technological Innovation*, 25 J. JUD. ADMIN. 96 (2015); Tania Sourdin, *Judge v Robot?: Artificial Intelligence and Judicial Decision-Making*, 41 U. N.S.W. L.J. 1114, 1117 (2018) [hereinafter *Judge v Robot*].

because of technology's impact on the role and function of judges, specifically in their adjudicative responsibilities.[8]

Currently, AI is performing several tasks within law enforcement and the judiciary. For instance, AI is used to structure data because of its capability to identify patterns within textual documents and files.[9] This is especially advantageous in scenarios involving the categorization of vast data volumes or intricate cases filled with extensive information. Courts in some jurisdictions acknowledge this investigative approach[10] as offering swifter and more accurate results compared to manual document research.[11]

In law enforcement, facial recognition technology ("FRT") serves preventive purposes by identifying previously known perpetrators to forestall potential future crimes.[12] Additionally, it can be employed to locate someone, aiding in the identification of individuals wanted for criminal activities.[13] For example, FRT can be applied in scenarios where every person within a specific context (like crossing a street or passing the gates in an airport) is facially scanned.[14]

AI can efficiently write judicial decisions by processing the provided data while also creating a cohesive and logical text.[15] Legal

---

8. *Judge v Robot*, *supra* note 7.

9. A. D. (Dory) Reiling, *Courts and Artificial Intelligence*, 11 INT'L J. FOR CT. ADMIN, Aug. 10, 2020, at 3, https://storage.googleapis.com/jnl-up-j-ijca-files/journals/1/articles/343/submission/proof/343-1-1484-1-10-20200810.pdf.

10. For example, in the United States, eDiscovery is a prime illustration. It uses automated methods to explore electronic data for discovery before the commencement of legal proceedings. Machine learning AI are trained to use the most effective algorithm for extracting pertinent details from copious amounts of information. Involved parties must mutually agree on the search terms and coding, which the judge then evaluates and approves. *Id.* at 3. *But cf.* Keeton Christian, *The Fortification of the Great Firewall and Its Effect on e-Discovery Disputes in U.S. Courts*, 82 U. PITT. L. REV. 173 (2020).

11. Reiling, *supra* note 9, at 4.

12. Vera Lúcia Raposo, *The Use of Facial Recognition Technology by Law Enforcement in Europe: A Non-Orwellian Draft Proposal*, 29 EUR. J. ON CRIM. POL'Y & RSCH. 515, 518 (2023) [hereinafter Raposo, *Use of FRT*].

13. *Id.*

14. *Id.*

15. *See* John Campbell, *Ex Machina: Technological Disruption and the Future of Artificial Intelligence in Persuasive Legal Writing*, 5 U. BOLOGNA L. REV. 294, 308-311 (2020).

and judicial work heavily relies on language, documents, and texts as foundational elements.[16] This data holds immense significance in the judicial system, aiding investigators, lawyers, and judges in piecing together the intricacies of a specific case, ultimately contributing to the pursuit of justice. Moreover, the advent of digitization within law firms and court systems has unlocked a vast reservoir of data, encompassing court opinions, statutes, regulations, books, practice guides, law reviews, legal white papers, and news reports.[17] This wealth of information is instrumental in training AI foundation models for judicial purposes. These models, once trained, can serve the court's staff in organizing, searching, and summarizing extensive piles of legal text.

Predictive analytics stand out as a significant function of AI within legal proceedings because some AI garner attention by claiming to predict court decisions.[18] This ability is often referred to as "predictive justice."[19] This concept entails using machine learning algorithms to conduct a probabilistic analysis of specific legal disputes by referencing case law precedents.[20] The term "forecast" better reflects these algorithms because they resemble a weather forecast more than a predetermined fact. Like weather predictions, court proceedings carry inherent unpredictability.[21]

This essay will focus on the use of AI in risk assessment and recidivism prediction on the global scale.[22] It will identify potential biases arising from using AI in various countries and assess whether those biases should lead to the exclusion of predictive AI in this domain.

---

16. *See id.* at 299-301.

17. Reiling, *supra* note 9.

18. Reiling, *supra* note 9, 4–5.

19. Seth Lazar & Jake Stone, *On the Site of Predictive Justice*, 2023 NOÛS 1.

20. *See AI in the Criminal Justice System*, ELEC. PRIV. INFO. CTR. [EPIC], https://epic.org/issues/ai/ai-in-the-criminal-justice-system/ (last visited Apr. 6, 2024). *See generally* Marc Queudot & Marie-Jean Meurs, *Artificial Intelligence and Predictive Justice: Limitations and Perspectives*, *in* RECENT TRENDS AND FUTURE TECHNOLOGY IN APPLIED INTELLIGENCE 889 (Malek Mouhoub et al. eds., 2018). However, this term has sparked debates, as the outcomes of prediction algorithms neither represent justice nor ensure predictability, see Reiling, *supra* note 9, at 4.

21. *Id.* at 4–5.

22. Meaning, using AI to analyze data involving a crime and a defendant to make predictions based on that analysis.

II. RISK ASSESSMENT AND PREDICTION IN CRIMINAL JUSTICE

Risk assessment tools have undergone a transformative journey, progressing from first-generation tools rooted in the clinical judgment and experience of decision-makers to second-generation tools that relied on static risk factors, like criminal history, age, and gender. [23] Third-generation instruments incorporated both risk and needs assessments, considering both static and dynamic (e.g., educational and employment status) factors.[24] Current fourth-generation tools offer individualized plans based on assessments of both static and dynamic factors.[25] A fifth generation of these tools is on the horizon, exploring the integration of machine learning techniques.[26] This next phase aims to predict recidivism in real-time, employing more intricate analyses for a deeper understanding of the complexities involved.[27]

Research into predicting criminal recidivism through statistical analysis has spanned nearly a century, involving contributions from criminal justice, psychology, and law.[28] Actuarial risk assessment models, a product of this extensive research, are now widely employed across various jurisdictions.[29] These assessments pivotally assist judges in making decisions about pre-trial release, sentencing, and probation, which influence the course of individuals' lives.[30] Risk assessments have effectively mitigated costs and reduced decision time.[31] However, the system is far from perfect.

---

23. *History of Risk Assessment*, BUREAU OF JUST. ASSISTANCE [BJA], https://bja.ojp.gov/program/psrac/basics/history-risk-assessment (last visited Jan. 11, 2024).

24. Brandon L. Garrett & John Monahan, *Judging Risk*, 108 CALIF. L. REV. 439, 451 (2020).

25. *Id.*

26. *Id.*

27. *Id.*

28. *See generally* GEORGIA ZARA & DAVID FARRINGTON, CRIMINAL RECIDIVISM: EXPLANATION, PREDICTION AND PREVENTION (2015).

29. STEVEN L. CHANENSON & JORDAN M. HYATT, BUREAU JUST. ASSISTANCE, THE USE OF RISK ASSESSMENT AT SENTENCING: IMPLICATIONS FOR RESEARCH AND POLICY 3 (2016).

30. ZARA & FARRINGTON, *supra* note 28.

31. Should AI be able to more accurately predict recidivism? Governments may save time and money because less resources will be dedicated to re-offenders. For a further discussion *see* Richard Berk, *An Impact Assessment of Machine*

For a start, one must keep in mind that risk assessments offer a probability-based forecast of an individual's potential for reoffending.[32] While these assessments assist practitioners in gauging the likelihood of an individual re-offending, they do not possess the capability to predict a person's behavior with complete certainty.[33]Another particular concern regarding risk assessment and recidivism prediction involves biases.[34] This is not a new concern, as human bias in criminal justice is a long-lasting apprehension.[35] Although many have argued AI could reduce it,[36] the situation is not that simple.

III. BIASES OF HUMAN DECISION MAKERS IN THE JUDICIAL SYSTEM

The outcomes of human adjudicative decision-making are shaped by various elements that impact substantive justice, including factors as trivial as the timing and content of a judge's meals,[37] the time of the decision,[38] the cumulative effect of making decisions throughout the

Learning Risk Forecasts on Parole Board Decisions and Recidivism, 13 J. EXP CRIMINOL 193 (2017).

32. *See* Caroline Wang et al., *In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction*, 39 J. QUANTITATIVE CRIMINOLOGY 519, 538-40 (2022).

33. *Id.*

34. *Id.*

35. Marjorie S. Zatz, *The Convergence of Race, Ethnicity, Gender, and Class in Court Decisionmaking: Looking Toward The 21st Century*, *in* POLICIES, PROCESSES, AND DECISIONS OF THE CRIMINAL JUSTICE SYSTEM 503, 539 (Julie Horney ed. 2000); Yu Du, *Racial Bias Still Exists in Criminal Justice System? A Review of Recent Empirical Research*, 37 TOURO L. REV. 79, (2021).

36. *See* Mirko Bagaric et al., *The Solution to the Pervasive Bias and Discrimination in the Criminal Justice System: Transparent and Fair Artificial Intelligence*, 59 AM. CRIM. L. REV. 95, 98 (2022).

37. *See* Zoe Corbyn, *Hungry Judges Dispense Rough Justice*, NATURE (Apr. 11, 2011), https://www.nature.com/articles/news.2011.227.

38. Thaís Guedes Ramos, *Utilizando o horário e a alimentação para sair do modo padrão e vencer na advocacia* [*Using Schedule and Nutrition to Break the Default Mode and Succeed in Law and Practice*], JUSBRASIL, https://www.jusbrasil.com.br/artigos/utilizando-o-horario-e-a-alimentacao-para-sair-do-modo-padrao-e-vencer-na-advocacia/301389811 (last visited Feb. 4, 2024).

day ("decision fatigue"),[39] personal values,[40] unconscious assumptions,[41] and reliance on intuition.[42] Additionally, factors external to the judge, such as the quality of legal representation[43] or the litigant's available resources,[44] play a significant role in shaping court decisions.

Human discretion is a crucial component in decision-making.[45] In discretionary decision-making, choices are driven by reasoning and judgment rather than solely adhering to predetermined criteria.[46] Community values, subjective party features, and other relevant circumstances are, even if unconsciously, considered in the exercise of discretion.[47] Contrarily, when the mandate for decision-making stems from legislation or internal procedures, it is typical for the authority to outline specific criteria guiding the decision-making process to make it less discretionary and more logical.[48]

Shifting from discretionary principles to defined criteria through increased automated decision-making to enhance efficiency might seem seductive. While intended to simplify the law and make it more definite, this change raises concerns about fairness and arbitrariness because of a

---

39. Luis C. Torres & Joshua H. Williams, *Tired Judges? An Examination of the Effect of Decision Fatigue in Bail Proceedings*, 49 CRIM. JUST. & BEHAV. 1233 (2022).

40. Rachel J. Cahill-O'Callaghan, *The Influence of Personal Values on Legal Judgments*, 49 J.L. & SOC'Y 596, (2013).

41. Jeffrey J. Rachlinski et al., *Does Unconscious Racial Bias Affect Trial Judges*, 84 NOTRE DAME L. REV. 1195, 1221 (2008).

42. Chris Guthrie et al., *Blinking on the Bench: How Judges Decide Cases*, 93 CORNELL L. REV. 1, 6–13 (2007).

43. *See* Richard A. Posner & Albert H. Yoon, *What Judges Think of the Quality of Legal Representation*, 63 STAN. L. REV. 317, 320 (2011).

44. Albert Yoon, *The Importance of Litigant Wealth*, 59 DEPAUL L. REV. 649, 652 (2010).

45. *Judge v Robot*, *supra* note 7, at 1128–30.

46. CRIME & CORRUPTION COMM'N (QUEENSLAND), DISCRETIONARY DECISION-MAKING POWERS: IDENTIFYING POTENTIAL CORRUPTION RISKS 1–2 (Dec. 2020), https://www.ccc.qld.gov.au/sites/default/files/Docs/Publications/CCC/Prevention-in-Focus-Discretionary-decision-making-identifying-potential-corruption-risks-updated-December-2020.pdf.

47. *Judge v Robot*, *supra* note 7, at 1128.

48. *See generally* DECISION MAKING IN CRIMINAL JUSTICE: TOWARD THE RATIONAL EXERCISE OF DISCRETION (Michael R. Gottfredson & Don M. Gottfredson eds., 1987).

potential lack of individualized justice, discretion, and a lack of nuance in the law, which could lead to unfair or arbitrary decisions.[49]

Judges are not purely rational actors who employ logical reasoning based on various legal sources to make decisions. It is widely acknowledged that this model only captures a portion of the decision-making process, overlooking non-doctrinal factors that significantly influence case outcomes.[50] Therefore, there is a place for discretion in judicial decision-making.[51]

This element of discretion is a quality that computer programs, operating on logical principles, may find challenging to accommodate.[52] Programmed algorithms process data to predetermine results, thus highlighting the rigid AI systems that conflict with the nuanced nature of discretionary decisions.[53]

A. Human Biases in the Judiciary

There are two types of biases affecting human judgment. First, there are "social biases," wherein one instinctively forms opinions or makes quick judgments based on an individual's social group.[54] Examples include instantly favoring someone with a similar accent or assuming someone from a different ethnic background might be untruthful.[55] The second category comprises "cognitive biases," representing systematic tendencies in one's thought processes that can lead to errors.[56] "Confirmation bias" is a good example illustrating how people are

49. Melissa Perry & Alexander Smith, *iDecide: The Legal Implications of Automated Decision-Making*, FED. CT. AUSTL. (Sep. 17, 2014), https://www.fedcourt.gov.au/digital-law-library/judges-speeches/justice-perry/perry-j-20140915.

50. *See* Cahill-O'Callaghan, *supra* note 40.

51. John N. Drobak & Douglass C. North, *Understanding Judicial Decision-making: The Importance of Constraints on Non-rational Deliberations*, 26 WASH. U. J. L. & POL'Y 131, 132 (2008).

52. *Judge v Robot*, *supra* note 7, at 1128.

53. *Id.*

54. *Cognitive Biases, Social Biases, and the Law*, AUSTL. L. REFORM COMM'N (June 16, 2021), https://www.alrc.gov.au/inquiry/review-of-judicial-impartiality/spotlight-on/cognitive-biases/.

55. *See id.*

56. *Id.*

prone to seek information confirming their beliefs while neglecting potential contradictions.[57]

Some years ago, an empirical study carried out in the United States ("U.S.") involving 239 serving federal and state judges—including 100 federal district judges that represent all Circuits—explored the potential manifestations of the implicit biases in their decisions.[58] The study revealed that judges exhibited strong to moderate negative implicit stereotypes against Asian Americans and the Jewish community.[59] In contrast, they held favorable implicit stereotypes towards Whites and Christians. Whereas Whites and Christians were associated with positive moral traits such as "trustworthy," "honest," and "giving," Asians and Jews were perceived with negative stereotypes such as "greedy," "dishonest," and "controlling."[60] Notably, the study indicated that federal district court judges tended to impose marginally longer prison terms on Jewish defendants compared to identical Christian defendants,[61] and their implicit biases were identified as an underlying factor to this unfair disparity in sentencing.[62]

IV. CASES OF AI SYSTEMS IN THE JUDICIAL SYSTEM

A primary difficulty in using AI to predict criminal behavior is embedded into the very foundation of AI predictions. AI prediction inherently and implicitly assumes individuals convicted of "similar" crimes share enough commonalities for past recidivism rates to predict the future recidivism risk of others in the same group.[63] However, this assumption is not without complexities. The assumption presupposes

57. See Abdul Malek, *Criminal Courts' Artificial Intelligence: The Way It Reinforces Bias and Discrimination*, 2 AI & ETHICS 233, 237 (2022), https://doi.org/10.1007/s43681-022-00137-9.

58. Justin D. Levinson et al., *Judging Implicit Bias: A National Empirical Study of Judicial Stereotypes,* 69 FLA. L. REV. 63, 69 (2017).

59. *Id.* at 63.

60. *Id.*

61. *Id.*

62. *Id.*

63. Ariel G. Stone et al., *Trajectories of Change in Acute Dynamic Risk Ratings and Associated Risk for Recidivism in Paroled New Zealanders: A Joint Latent Class Modelling Approach*, J. QUANTITATIVE CRIMINOLOGY, Jan. 16, 2023, https://link.springer.com/article/10.1007/s10940-022-09566-5.

uniformity among individuals who have committed "similar" crimes, oversimplifying the diverse factors that contribute to criminal behavior. Human behavior is complicated and influenced by various factors such as socioeconomic background, personal history, and individual circumstances,[64] which may not be fully captured by a narrow categorization of offenses. The pitfalls of the assumption become evident when considering the dynamic and multifaceted nature of a crime. Individuals within the group who committed "similar" crimes can have distinct life experiences, motivations, and levels of culpability. Relying solely on historical data to predict future recidivism may oversimplify the complexities of human behavior, which leads to biased outcomes and reinforces existing societal disparities.[65]

The performance of AI in courts or any other domain heavily relies on data to yield an appropriate outcome. High-quality data is essential for optimal AI outcomes since inaccurate and/or incomplete data can lead to legally flawed decisions.[66]

Further, accurate and complete datasets alone are insufficient. Natural language processing and text recognition play a crucial role by enabling an external evaluation of a legal professional's behavior.[67] For more effective AI utilization, legal information (e.g., court decisions and criminal precedents) should be made machine-processable, encompassing textual readability, document structures, identification codes, and metadata, as all these materials would enrich the training

64. *See* Aletha C. Huston & Alison C. Bentley, *Human development in Societal Context*, 61 ANN. REV. PSYCH. 411 (2010).

65. *See* Lindsay Weinberg, *Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches*, 74 J. A.I. RSCH. 75 (2022).

66. *See* Partrick Mikalef & Manjul Gupta, *Artificial Intelligence Capability: Conceptualization, Measurement Calibration, and Empirical Study on Its Impact on Organizational Creativity and Firm Performance*, 58 INFO. & MGMT. 1, 4 (2021); OSONDE OSOBA & WILLIAM WELSER IV, AN INTELLIGENCE IN OUR IMAGE: THE RISKS OF BIAS AND ERRORS IN ARTIFICIAL INTELLIGENCE 3–4, 22 (2017), https://www.rand.org/content/dam/rand/pubs/research_reports/RR1700/RR1744/RAND_RR1744.pdf.

67. Reiling, *supra* note 9, at 8 (explaining how accurate and complete datasets are not sufficient for determining legal decisions due to the implications with natural language processing).

datasets.[68] Adding clear definitions and structured terminology to legal information will further elevate AI's capabilities.[69]

Another critical feature is AI's ability to elucidate its outcomes, requiring explanations of its process and substantive reasoning.[70] Although AI can technically provide explanations akin to that of humans, research indicates that human-generated explanations are clearer in some instances.[71]

An overreliance on technology poses another risk because individuals can become dependent on automated decision-making systems, placing unwavering trust in statistical data rather than making independent judgments.[72] This dependence can lead to overlooking system errors.[73]

A. AI Biases

Bias is the ultimate obstacle hindering the effective performance of AI in courts. While algorithms used for risk assessment in the criminal justice system can potentially make sentencing more accurate and reduce human error, they could conversely deepen existing biases.[74]

---

68. *Id.* at 7; *see* Don Farrands, *Artificial Intelligence and Litigation – Future Possibilities*, 9 J. Civ. Litigation & Practice 7 (2020).

69. Reiling, *supra* note 9, at 8; Mikalef & Gupta, *supra* note 66, at 4 ("Adding to this issue, skewed data during labeling and training can potentially result in biased AI applications.").

70. *See, e.g.*, Alejandro Barredo Arrieta et al., *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI*, 58 INFO. FUSION 82 (2020).

71. For more details, see Xinru Wang & Ming Yin, *Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making*, 2021 INT'L CONF. ON INTELLIGENT USER INTERFACES 318, https://dl.acm.org/doi/pdf/10.1145/3397481.3450650. In their paper, Wang and Yin claim the fact that humans usually use contrastive explanations—explanations that focus on the difference between two scenarios: one where the event happens and one where it does not—help others to better understand the information communicated and they recommend AI explanations to follow this same model. *Id.*

72. Matthew Grissinger, *Understanding Human Over-Reliance on Technology*, 44 PHARMACY & THERAPEUTICS 320, 320-21 (2019).

73. *Id.*

74. Danielle Kehl et al., *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing. Responsive Communities Initiative,*

Additionally, they might challenge the core principles of fairness that are crucial to the justice system.[75]

There are several stories illustrating instances where the system misrepresented the recidivism risks associated with various offenders, showing a notable disparity in the misclassification of risk levels based on race.[76] In many instances, Black prisoners were unduly given higher risk scores than their non-Black counterparts, even if the latter committed more severe crimes.[77] In another study, Black defendants had a two times higher risk of being mislabeled as potential violent recidivists relative to their White counterparts; additionally, White individuals with a history of recidivism were incorrectly classified as low risk 63.2% more frequently than their Black counterparts.[78] These findings suggest a systematic racial bias entrenched in estimating risk.

In conclusion, the use of AI in the criminal justice system presents a paradox of potential and peril. While it holds the promise of enhancing accuracy and reducing human biases in sentencing,[79] the reality, marred by systematic biases in risk assessment algorithms, poses a stark contradiction.[80] The documented disparities, particularly in the misclassification of risk levels that disproportionately affect ethnic individuals, underscore the urgent need for reform. These biases not only compromise the fairness and integrity of the justice system but also betray its foundational principles.[81]

BERKMAN KLEIN CTR. FOR INTERNET & SOC'Y, Aug. 2017, at 3, https://dash.harvard.edu/handle/1/33746041.

75. *Id.*

76. *See id.* at 29; *see, e.g.*, Julia Angwin et al., *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

77. OSOBA & WELSER IV, s*upra* note 66, at 13.

78. Jeff Larson et al., *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

79. John Villasenor & Virginia Foggo, *Artificial Intelligence, Due Process, and Criminal Sentencing*, 2020 MICH. STATE L. REV. 295, 298 (2020).

80. *See* Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218 (2019).

81. Expert Comm. on Hum. Rts. Dimensions of Automated Data Processing & Different Forms of A.I., *A Study of the Implications of Advanced Digital Technologies*, *(Including AI Systems) for the Concept of Responsibility Within a Human*

B. Causes of AI Biases

i. Code Biases

Code bias denotes consistent and systemic inaccuracies within a computer system that cause unfair outcomes, favoring one arbitrary group of users at the expense of others.[82] Code bias can emerge when programming reflects the developer's unintentional or implicit biases during the development of software and algorithms.[83] These biases can be related to race, gender, ethnicity, socioeconomic status, or other factors, and they may perpetuate existing disparities in society by unfairly creating outcomes based on them.[84]

Developers crafting AI code often envision the typical user as someone resembling themselves or a similar demographic.[85] As a result, individuals who deviate from this "presumed average user" are more likely to suffer from negative consequences from the AI functioning.[86]

When AI performs its role in the judiciary—including risk assessment and recidivism prediction—an additional element can lead to

---

Rights Framework, at 32–33, DGI(2019)05, https://rm.coe.int/responsability-and-ai-en/168097d9c5.

82. Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 4–5 (2014).

83. *See* Gabbrielle M. Johnson, *Algorithmic Bias: On the Implicit Biases of Social Technology*, 198 SYNTHESE 9941 (2021).

84. *See* Danieli Evans Peterman, *Socioeconomic Status Discrimination*, 104 VA. L. REV. 1283, 1302–1304 (2018).

85. Simon Pienaar, *Programming and People: Unwritten Bias in Written Code*, TTRO BLOG, https://www.ttro.com/blog/technology/programming-and-people-unwritten-bias-in-written-code/ (last visited Jan. 2, 2024).

86. *Id.* This led to reconceptualizing algorithmic decision-making software and its outputs as cultural artifacts. The term "cultural artifact" may evoke thoughts of documentaries and archaeology, but in this context, it encompasses any tangible creation by humans that provides insights into their culture or society. This reframing emphasizes the need for inclusivity and awareness of diverse perspectives in the development and deployment of technology. *Compare id.*, *with* Arianna Falbo & Travis LaCroix, *Est-ce que Vous Compute? Code-Switching, Cultural Identity, and AI*, 8 Feminist Philosophy Quarterly 1 (2022) (explaining how algorithmic decision-making should embrace inclusivity and awareness of diverse perspectives in the development and deployment of technology instead of merely yielding cultural artifacts).

code biases: the inevitable transformation of legal concepts into codes, commands, and functions.[87] Legal texts are inherently nuanced and reliant on context, which adds another complexity for computer programmers and information technology ("IT") professionals because they usually lack legal qualifications or experience. Despite this, these professionals are responsible for translating legislation and case law into computer codes and commands to facilitate the system's autonomous decision-making.[88] The legal sources, which are already intricate on their own, further operate within the vague framework of statutory presumptions and discretionary judgments. Furthermore, these codes need to be constantly updated since laws are frequently amended, cases get overturned, and complex transitional provisions are made.[89]

ii. Data Biases

The effectiveness of an AI hinges on the quality of the data it learns from.[90] The effectiveness of AI training hinges on having access to vast, precise datasets even though the methods for training AI differ.[91] Flawed training data lies at the core of the main issues encountered in AI, a concept colloquially referred to as "garbage in, garbage out."[92] When automated learning uses biased data, it inevitably yields biased results.[93]

---

87. *See* Francesco Contini, *Artificial Intelligence and the Transformation of Humans, Law and Technology Interactions in Judicial Proceedings*, 2 L. TECH. & HUM. 4, 6-7 (2020).

88. Although this work is not done solely by IT people but also involves the assistance of legal experts, it is important to note that the coding itself is primarily executed by the former.

89. *See* Perry & Smith, s*upra* note 49, at 32.

90. Steven Euijong Whang et al., *Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective*, 32 VLDB J. 791, 791-92 (2023).

91. *Id.* at 792 ("[A] common complaint in the industry is that research institutions spend 90% of their machine learning efforts on algorithms and 10% on data preparation, although based on the amounts of time spent, the numbers should be 10% and 90% the other way.").

92. The concept is referring to L. Todd Rose & Kurt W. Fischer, *Garbage In, Garbage Out: Having Useful Data Is Everything*, 9 MEASUREMENT: INTERDISCIPLINARY RSCH. & PERSP. 222 (2011).

93. *See generally* Eirini Ntoutsi et al., *Bias in Data-Driven Artificial Intelligence Systems—An Introductory Survey*, 10 WIRES DATA MINING & KNOWLEDGE DISCOVERY 1 (2020).

The Correctional Offender Management Profiling for Alternative Sanctions ("COMPAS") prominently illustrates how biased data exacerbates AI biases within the legal system.[94] COMPAS is an AI tool employed by criminal judges in certain U.S. states.[95] It is used to assess the recidivism risk of convicted defendants, influencing decisions related to pre-trial detention, sentencing, or early release.[96] Proponents often argue that tools like COMPAS contribute to increasing objectivity in evaluating recidivism and reducing the number of individuals held in detention.[97] Despite its intent to identify non-threatening defendants, by relying on historical data, COMPAS consistently overestimates recidivism among African American defendants compared to Caucasian Americans.[98] COMPAS relies on criminal records and a 137-question questionnaire, posing inquiries such as, "Is someone who is hungry allowed to steal? Strongly disagree, disagree, etc."[99] In the words of Beth Karp, "It is dangerously reductive to assume that any person's value system is determined by their race, ethnicity, or financial status, or that particular values or behaviors are exclusive to certain 'cultures.' . . . It is deeply offensive (and, of course, incorrect) to presume that the extent of agreement or disagreement with that statement is a matter of 'cultural

94. *See* Anne L. Washington, *How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate*, 17 COLO. TECH. L.J. 131, 135 (2018); *see also* Willem Gravett, *Sentenced by an Algorithm—Bias and Lack of Accuracy in Risk-Assessment Software in the United States Criminal Justice System*, 34 S. AFR J. CRIM. JUST. [SACJ] 31 (2021); Villasenor & Foggo, *supra* note 79, 334–340 (2020).

95. *See* Washington, *supra* note 94, at 133.

96. *See id.* at 143.

97. *See* Sam Corbett-Davies et al., *A Computer Program Used for Bail and Sentencing Decisions Was Labelled Biased Against Blacks. It's Actually Not That Clear*, WASH. POST (Oct. 17, 2016, 5:00 AM), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.

98. JULIA DRESSEL & HANY FARID, THE DANGERS OF RISK PREDICTION IN THE CRIMINAL JUSTICE SYSTEM 7-8 (2021); Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 SCI. ADVANCES 1, 2-3 (2018), https://www.science.org/doi/epdf/10.1126/sciadv.aao5580.

99. Julia Angwin, *Sample-COMPAS-Risk-Assessment-COMPAS-"CORE"*, PROPUBLICA, https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE (last visited Mar. 27, 2024).

values.'"[100] ProPublica's investigation revealed Black defendants were more likely to be improperly assessed for higher recidivism risk, while White defendants were not.[101]

The lack of representation within the dataset can also fuel biases. Machine learning algorithms use statistical estimation methods, and the metrics gauging estimation errors tend to fluctuate inversely with data sample sizes.[102] Consequently, these methods are more prone to errors when training with classes that have lower representation compared to others.[103] For instance, if ethnic minorities are over-policed, and thus overrepresented in crime data, AI models might unjustly predict higher recidivism rates for individuals from these groups.[104] Likewise, predictive models can also exhibit gender biases. Since most criminal datasets contain a higher proportion of male offenders than female offenders,[105] AI systems might be less accurate for females, which can potentially lead to harsher or inappropriate sentencing recommendations. This lack of respresentation can affect the accuracy of predictions for female defendants, either by overestimating or underestimating their risk of recidivism.[106]

---

100. Beth Karp, *What Even Is a Criminal Attitude?-And Other Problems with Attitude and Associational Factors in Criminal Risk Assessment*, 75 STAN. L. REV. 1431, 1504 (2023).

101. *See id.* at 1439; Washington, *supra* note 94.

102. *See* Muhammad Asim et al., *Invertible Generative Models for Inverse Problems: Mitigating Representation Error and Dataset Bias*, 119 PROC. MACH. LEARNING RSCH. [PMLR] 399 (2020), https://proceedings.mlr.press/v119/asim20a/asim20a.pdf.

103. A prime example of this issue is evident in Yahoo's automated image tagging system, which made biased image-labelling choices due to demographic inhomogeneity in its training data. *See* OSOBA & WELSER IV, s*upra* note 66, at 19–20; *see also* Moritz Hardt, *How Big Data Is Unfair: Understanding Sources of Unfairness in Data Driven Decision Making*, MEDIUM (Sep. 26, 2014), https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de.

104. *See* Matthew Browning & Bruce Arrigo, *Stop and Risk: Policing, Data, and the Digital Age of Discrimination*, 46 AM. J. CRIM. JUST. 298, 299-301 (2021).

105. Jennifer Skeem et al., *Gender, Risk Assessment, and Sanctioning: The Cost of Treating Women like Men*, 40 L. & HUM. BEHAV. 580, 581 (2016).

106. *Id.*

C. Factors Challenging the Identification (and Subsequent Correction) of AI Biases

Identifying and correcting biase in AI present a challenge due to various factors. These include biases inherent in the data used for training, complexities in algorithms that make it hard to pinpoint sources of bias, and the possibility of developers inadvertently introducing their own implicit biases or lack of awareness.[107] Additionally, the dynamic nature of bias, the opacity of AI and resource constraints further complicate the process. Overcoming these challenges requires interdisciplinary collaboration, diversity in development teams, and standardized tools for bias detection and correction, all while navigating legal and ethical considerations.[108]

i. Dynamic Algorithms

AI algorithms can be static, dynamic, or something in between. A static algorithm remains unchanged once it is finalized and put into service.[109] The system manufacturer designs the static algorithm not to evolve independently once the design is completed.[110] This characteristic makes the algorithm known to the manufacturer, potentially allowing analysis for a due process claim.[111] While a manufacturer may attempt to restrict access by invoking trade secrets, this presents a legal hurdle rather than a technological one.

In contrast, a dynamic algorithm continuously evolves from the new data it collects.[112] This evolution can occur rapidly; for example, an AI-based, pre-sentencing risk assessment system continuously monitors nationwide news feeds and arrest records.[113] If the AI system

---

107. Nizan Geslevich Packin & Yafit Lev-Aretz, *Learning Algorithms and Discrimination, in* RESEARCH HANDBOOK ON THE LAW OF ARTIFICIAL INTELLIGENCE 88 (Woodrow Barfield & Ugo Pagallo eds. 2018).

108. *See* Varsha P.S., *How Can We Manage Biases in Artificial Intelligence Systems – A Systematic Literature Review*, 3 INT'L J. INFO. MGMT. DATA INSIGHTS 1 (2023).

109. Villasenor & Foggo, *supra* note 79, at 312–13.

110. *Id.*

111. *See id.* at 312.

112. *See id.* at 312–13.

113. *Id.* at 313.

identifies previously unknown statistical correlations that could impact the recidivism rate for a specific crime, it adapts its algorithm for computing risk scores for future cases.[114] This adjustment can happen swiftly: even a single new recidivism incident can immediately prompt a recalculation of statistical correlations and cause changes in numerical parameters affecting risk score computation.[115]

The potential for autonomous evolution introduces a unique problem; no one, not even the system manufacturer, possesses a snapshot of the dynamic algorithm precisely as it existed when calculating a specific risk assessment score because it is always changing.[116] The information the algorithm used to calculate an individual's score may no longer be available when a request for information is made weeks or months after the score was computed.[117]

ii. Lack of Transparency and Explainability

The lack of information poses significant concerns in light of predictive AI bias.[118] The terms "legal black boxes" and "technical black boxes" underscore the difficulty of ensuring transparency since the algorithm itself is hard to comprehend.[119] These so-called "legal black boxes" of algorithmic code are safeguarded by contracts and kept confidential from the public even though they contain relevant information about the AI system, which causes a lack of transparency.[120] This information includes the code; the datasets used to train the system; performance metrics; identification and mitigation of biases; the presence of human oversight; and adherence to legal requirements.[121] In contrast,

114. *Id.*

115. *Id.*

116. *Id.*

117. *Id.*

118. *See* Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence*, 119 COLUM. L. REV. 1829, 1833 (2019); *see also* Malek, *supra* note 57, 238–240.

119. Maja Brkan & Gregory Bonnet, *Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: of Black Boxes, White Boxes and Fata Morganas*, 11 EUR. J. RISK REGUL. 18, 18–19 (2020).

120. *See id.* at 43–46.

121. Rita Matulionyte et al., *Should AI-Enabled Medical Devices Be Explainable?*, 30 INT'L J.L. & INFO. TECH. 151, 157 (2022).

technical black boxes arise when the algorithm's process is either undisclosed to developers or too intricate for human comprehension due to cognitive limitations, which causes a lack of explainability.[122]

Since important information about a legal black box AI system is kept confidential, its biased patterns and discriminatory behavior are hard to identify or even rectify. Numerous vital algorithms influencing public life are often classified as proprietary or trade secrets, which contributes to this scenario.[123] Maintaining this shroud of secrecy generally does not foster informed public discourse. Likewise, without transparency into evaluations and performance assessments, users will not have the insight into the metrics that determine the system's effectiveness.[124] For this reason, it is urgent to impose some economic sanctions on AI developers who do not disclose relevant information for legal black boxes. Since AI developers can provide certain information, they must be encouraged or incentivized to weigh whether the pros and cons (lack of competitive advantage, above all) will be conducive to improving the current status quo.

---

122. Transparency usually describes the deficiency of information in all aspects of an AI system. In this sense, the absence of explainability can be seen as one facet of the broader transparency issue. However, they should be considered distinct concepts. *See generally* Vera Lúcia Raposo, *How Is 'Unexplainable' and Non-Transparent AI Affecting Healthcare Delivery?*, OSLO L. REV. (forthcoming 2024) [hereinafter Raposo, *Unexplainable and Non-Transparent*]; Nagadivya Balasubramaniam et al., *Transparency and Explainability of AI Systems: From Ethical Guidelines to Requirements*, 159 INFO. & SOFTWARE TECH., Mar. 2023, https://doi.org/10.1016/j.infsof.2023.107197. Explainability, while forming a specific aspect of transparency, does not completely encapsulate the transparency spectrum. Furthermore, it is important to recognize that compliance with other transparency dimensions is usually technically feasible, and non-compliance is often rooted in legal or business considerations. In contrast, explainability encounters technical challenges due to the art's current state that prevents comprehensive explanations. *See* Sajid Ali et al., *Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence*, 99 INFO. FUSION, Nov. 2023, https://doi.org/10.1016/j.inffus.2023.101805; *see also* Raposo, *Unexplainable and Non-Transparent*, *supra* note 122.

123. *See* Katarina Foss Solbrekk, *Searchlights Across the Black Box: Trade Secrecy Versus Access to Information*, 50 COMPUT. L. & SEC. REV., Sep. 2023. Raposo, *Unexplainable and Non-Transparent, supra* note 122.

124. Balasubramaniam et al., *supra* note 122, at 8. Raposo, *Unexplainable and Non-Transparent*, *supra* note 122.

Dealing with technical black boxes is more intricate because legal compliance does not depend on human will. Society, including the developers themselves, does not understand how these AI systems reached their outcomes. White-box models (i.e., explainable models) allow developers and users to scrutinize the decision-making process by identifying patterns and detecting discriminatory tendencies.[125] In contrast, the lack of explainability creates a black-box scenario where biases go unnoticed and thus unaddressed.[126] Elucidating outcomes is crucial to safeguard democratic rights and legal frameworks, and the lack of transparency is especially problematic in judicial decision-making.

iii. Deficient Reward Mechanisms

Reward learning is a process through which a machine learns to associate certain actions or states with rewards.[127] Reward learning is fundamental to reinforcement learning, which is a type of machine learning where an agent learns to make decisions by interacting with an environment to maximize cumulative rewards.[128] Rewards serve as indicators guiding AI agents on which actions are favorable or unfavorable within a particular context.[129] For instance, in generative AI, these signals measure the system's effectiveness in meeting predefined goals. They attribute a reward score to each generated output, with higher scores indicating closer alignment with the desired objective.[130]

However, this very reward learning mechanism can also be a source of bias.[131] Both reward and reinforcement learning play critical roles in contemporary AI systems as an AI learns what is the *correct*

---

125. Octavio Loyola-Gonzalez, *Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View*, 7 IEEE ACCESS 154096, 154096 (2019).

126. *Id.*

127. Thomas Krendl Gilbert et al., *Reward Reports for Reinforcement Learning*, 2023 AAAI/ACM CONF. ON A.I., ETHICS, & SOC'Y [AIES '23], at 84.

128. *Id.*

129. *Id.*

130. *See Reward Modeling for Generative AI*, INNODATA, https://innodata.com/reward-modeling-for-generative-ai/ (last visited Mar. 27, 2024).

131. Peter Dayan & Bernard W. Balleine, *Reward, Motivation, and Reinforcement Learning*, 36 NEURON 285, 294 (2002). The foundation of reward functions in machine learning and AI theory can be traced back to behaviorist psychology. *Id.*

behavior.[132] Throughout an AI system's learning process, the reward function effectively measures how much positive or negative reinforcement is attributed to its actions and decisions;[133] this is similar to the terminology used in human psychology.[134] Subsequently, learning algorithms adjust the AI's parameters and behavior to maximize the overall reward.[135] Consequently, the shaping of AI behavior often boils down to crafting motivating reward functions.

However, a behaviorist learning approach is not foolproof and can be manipulated.[136] When a reward system is poorly outlined and given to an AI system, the reward system can lead to unintended side effects or behaviors, often referred to as "reward hacking."[137] If the reward system does not accurately capture the true objectives, values, or intentions of the designers, the AI may exploit loopholes or find shortcuts to achieve high rewards that are not aligned with the desired outcomes.[138]

V. MITIGATING AI BIASES

There are mitigation techniques that can be applied to overcome bias and thus contribute to the creation of trustworthy AI.[139]

A very popular anti-bias recommendation is to ensure training on a diverse dataset.[140] It is crucial to use a training dataset that mirrors a diverse and representative population to prevent the AI from

---

132. *See, cf. id.* at 285 (discussing how both reward and reinforcement techniques shape how animals respond to environmental events).

133. *Id.*

134. *Id. See* Norman M. White, *Reward: What Is It? How Can It Be Inferred from Behavior?*, *in* NEUROBIOLOGY OF SENSATION AND REWARD 45 (Jay A. Gottfried, ed., 2011).

135. Dayan & Balleine, *supra* note 132, at 285 (explaining how the predictive rewards from reinforcement learning maximizes rewards and minimizes punishments).

136. OSOBA AND WELSER IV, *supra* note 66, at 20.

137. *Id.*; Patrick Bradley, *Risk Management Standards and the Active Management of Malicious Intent in Artificial Superintelligence*, 35 A.I. & SOC'Y 319, 325 (2020).

138. *See* OSOBA & WELSER IV, *supra* note 66, at 20.

139. Michael Mayowa Farayola et al., *Ethics and Trustworthiness of AI for Predicting the Risk of Recidivism: A Systematic Literature Review*, 14 INFO., July 2023, at 1, 2, 5.

140. Emilio Ferrara, *Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies*, 6 SCIENCE, Dec. 2023, at 1–2.

emphasizing or de-emphasizing any particular group of people. Using a non-diverse dataset could lead to disproportional outcomes by introducing group bias into the model.

Another mechanism to mitigate biases involves balancing the representation of different groups.[141] For instance, oversampling the minority class in a dataset can help achieve a more balanced distribution as it ensures that the predictive model has sufficient representation from minority groups, allowing it to learn patterns and relationships specific to these demographics.[142] By balancing the class distribution, the model becomes less likely to exhibit bias or disparities in its predictions based on groups often subject to stereotypes and prejudices.

Another way to mitigate bias is to exclude sensitive variables from the training dataset. Sensitive variables are, but not limited to, race, gender, and age.[143] Consider a scenario where predictive AI is used in bail or pre-trial risk assessment hearings within the criminal justice system. The goal of the predictive AI model is to assess the risk of a defendant committing a crime or failing to appear in court if released on bail before their trial. Instead of including variables such as race, gender, or socioeconomic status in the predictive AI model, developers should focus on factors directly related to the defendant's risk of reoffending or failing to appear in court. Non-sensitive variables may include past criminal history, employment status, community involvement, or previous instances of bail violations.[144] However, the challenge is that some of these factors could also belong to a sensitivity list. For instance, the specific community the defendant lives in might be closely related to their specific ethnic or religious group. Therefore, at that point, can we accurately distinguish between a sensitive variable and a non-sensitive variable?

---

141. Sunzida Siddique et al., *Survey on Machine Learning Biases and Mitigation Techniques*, 4 DIGITAL 1, 58 (2024).

142. *See generally* Roweida Mohammed et al., *Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results*, 2020 INT'L CONF. ON INFO. & COMMC'N SYS. [ICICS] 243.

143. Wei-Hung Weng et al., *An Intentional Approach to Managing Bias in General Purpose Embedding Models*, 6 LANCET DIGIT. HEALTH 126, (2024).

144. *Id.*

Incorporating human oversight is another mechanism.[145] This involves the active participation of humans in the design, deployment, and monitoring of AI systems to identify and rectify biases.[146] As demonstrated in the previous example, a statistician could be influential in overseeing the results of a predictive AI system. Continuous testing ensures AI remains aligned with its intended purpose.[147] Designing an AI system should facilitate easy and robust adjustments, and continuous auditing becomes imperative.[148] The successful performance of this task requires that individuals, including judges, have a comprehensive understanding of AI functionality and moral character. Thus, implementing ethical principles within institutions and court processes—specifically decision-making authorities and compliance monitors—is essential.

Frequent AI audits are a crucial aspect of AI use. Auditability entails the preservation of all information utilized in conducting a risk assessment, making it potentially accessible "in the event of a due process challenge."[149] For audits to be possible, it is necessary to document events so that information is accessible to the auditors. Given the rapid evolution potential of an AI algorithm, failure to deliberately record the algorithm's state each time it is used may pose a reconstruction challenge. The volatility of an AI's algorithm emphasizes the importance of systematically documenting the algorithm's state to ensure comprehensive auditability.[150]

---

145. Kyriakos Kyriakou & Jahna Otterbacher, *In Humans, We Trust*, DISCOVER A.I., Dec. 2023, at 1, 3.

146. Johann Laux, *Institutionalised Distrust and Human Oversight of Artificial Intelligence: Towards a Democratic Design of AI Governance Under the European Union AI Act*, A.I. & SOC'Y, Oct. 2023.

147. *See* Ben Shneiderman, *Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems*, 10 ACM TRANSCTIONS ON INTERACTIVE INTEL. SYS., Oct. 2020, https://dl.acm.org/doi/10.1145/3419764.

148. *See generally* Natalia Díaz-Rodríguez et al., *Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation*, 99 INFO. FUSION 1 (2023).

149. Villasenor & Foggo, *supra* note 79, at 339. For a discussion on how Europe should implement Risk Assessment Technologies, see Georgios Bouchagiar, *Is Europe Prepared for Risk Assessment Technologies in Criminal Justice? Lessons from the US Experience*, 15 NEW J. EUR. CRIM. L. 72 (2024).

150. Villasenor & Foggo, *supra* note 79, at 339–40.

In addition, it is important to consistently evaluate the model's performance on different subgroups to identify and address potential biases.[151] Imagine a case involving predictive AI in bail decisions within a court system where an AI's common task is to disaggregate the data by demographic groups, such as race, gender, age, and socioeconomic status, to compare outcomes for each subgroup. For instance, bias can be discerned by comparing the rates of reoffending and failure to appear in court among defendants of different racial or ethnic backgrounds. Essentially, the AI could run statistical reports to demonstrate its judicious performance and uncover variables the AI regularly considers when making its assessments.

VI. THE LEGAL REGIME OUTLINED IN THE AI ACT

In 2012, the Artificial Intelligence Act ("AIA") was introduced to tackle potential risks associated with AI applications, particularly those affecting health, safety, and fundamental rights in Europe.[152] This legislative initiative set forth an innovative regulatory framework applicable to all 27 member states of the European Union, becoming the first comprehensive AI regulation introduced in the world.[153]

A. Classification of Predictive AI in Law Enforcement Scenarios

According to the original proposal from the European Commission,[154] AI systems used in law enforcement or justice administration

151. Ferrara, *supra* note 140, at 4.

152. Reena Bajowala et al., *Impact of EU Artificial Intelligence Act on US Entities*, BLOOMBERG L. (Aug. 2023), https://www.bloomberglaw.com/document/X3JV8BFK000000.

153. However, not immune to criticism. *See, e.g.*, Vera Lúcia Raposo, *Ex Machina: Preliminary Critical Assessment of the European Draft Act on Artificial Intelligence*, 30 INT'L J.L. & INFO. TECH. 88, 88-89 (2022) [hereinafter Raposo, *Ex Machina*]; Vera Lúcia Raposo, *The European Draft Regulation on Artificial Intelligence: Houston, We Have a Problem*, *in* PROGRESS IN ARTIFICIAL INTELLIGENCE: 21ST EPIA CONFERENCE ON ARTIFICIAL INTELLIGENCE 66, 66 (Goreti Marreiros et al., eds. 2022) [hereinafter Raposo, *Houston*]. Some of the issues pointed out in these studies have meanwhile been solved in the final version of the AIA, while some others still persist.

154. *Commission Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial*

were classified as high-risk AI systems by the AIA.[155] This means that they are considered to pose a significant risk of harm to the health, safety, or fundamental rights of natural persons. This classification of AI survived the AIA's European Council approval process.[156] However, the version approved by the European Parliament was significantly different.[157] The category of forbidden AI grew to include AI systems used to evaluate the risk associated with individuals or groups and the likelihood of someone committing an offense or reoffending. This includes predicting the recurrence of potential legal violations through profiling.[158] The profiling process considers various factors such as personality traits, characteristics, location, and past criminal history.[159]

As of the March 13, 2024, iteration of the AIA, Article 5, relating to prohibited AI systems, included this same prohibition.[160] The norm states that

---

Intelligence Act) and Amending Certain Union Legislative Acts, at 3, COM (2021) 206 final (Apr. 21, 2021).

155. Kate Crawford & Jason Schultz, *AI Systems as State Actors*, 119 COLUM. L. REV. 1941, 1954 (2019). However, barring specific exceptions, the use of real-time biometric identification systems in public spaces by law enforcement was generally prohibited under the AIA. *See* Vera Lúcia Raposo, *'Look at the Camera and Say Cheese': The Existing European Legal Framework for Facial Recognition Technology in Criminal Investigations*, 33 INFO. & COMM. TECH. L. 1 (2024) [hereinafter Raposo, *Look at the Camera*]; Raposo, *Use of FRT*, *supra* note 12, at 517–18.

156. *See generally* Permanent Reps. Comm., Letter dated Nov. 25, 2022 from the Permanent Representatives Committee to the Council of the Eur. Union, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts – General Approach, 2021/0106(COD) (Nov. 25, 2022), https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf.

157. *See generally* 2024 O.J. (C 506), https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:C_202400506&qid=1711596324451.

158. *Artificial Intelligence Act Article 5*, ACCESIBLE L., https://artificialintelligenceact.com/title-ii/article-5/ (last visited Mar. 27, 2024).

159. Wayne Petherick & Nathan Brooks, *Reframing Criminal Profiling: A Guide for Integrated Practice*, 28 PSYCHIATRY, PSYCH. & L. 694, 698–702 (2021).

160. Resolution on the Proposal for a Regulation of the European Parliament and of the Concil on Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)), EUR. PARL. DOC. P9_TA(2024)0138 ch. II, art. 5(1)(d) (2024).

> [T]he placing on the market, the putting into service for this specific purpose, or the use of an AI system for making risk assessments of natural persons in order to assess or predict the likelihood of a natural person committing a criminal offence, based solely on the profiling of a natural person or on assessing their personality traits and characteristics; this prohibition shall not apply to AI systems used to support the human assessment of the involvement of a person in a criminal activity, which is already based on objective and verifiable facts directly linked to a criminal activity[.][161]

When a predictive AI system falls under the above prohibition's exception, it is classified as a high-risk AI system.[162] The AIA classifies AI that meets the exception as "high risk" because those systems can be traced back to the AIA's Annex III, which outlines a list of AI systems considered high-risk per AIA's Article 6(2) unless they do "not pose a significant risk of harm, to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making."[163] Annex III, Article 6(e) of the referred Annex III refers to:

> AI systems intended to be used by ***or on behalf of law enforcement authorities or by Union institutions, bodies, offices or agencies in support of*** law enforcement authorities for the profiling of natural persons as referred to in Article 3(4) of Directive (EU) 2016/680 in the course of the detection, investigation or prosecution of criminal offences.[164]

High-risk AI systems are mandated to meet specific criteria, including developers using representative, unbiased, and accurate datasets for training; implementing human oversight; maintaining records for compliance checks; and providing relevant information to users.[165] Distinct

---

161. *Id.*

162. Gijs van Dijck, *Predicting Recidivism Risk Meets AI Act*, 28 EUR. J. CRIM. POL'Y & RSCH. 407, 410 (2022).

163. Eur. Parl. Doc. P9_TA(2024)0138, *supra* note 160, ch. III, art. 6(2). The norm subsequently establishes the conditions for this exception to work. Looking at those requirements, it is very unlikely that predictive AI for recidivism prediction escapes to this classification.

164. EUR. PARL. DOC. P9_TA(2024)0138, *supra* note 160, annex III, art. 6(e).

165. *See generally id.* at ch. III.

requirements apply to various stakeholders, such as providers, importers, distributors, and users of AI systems.[166] These requirements involve adherence to the Regulation's specifications and the application of CE marking of conformity[167] to indicate conformity with the Regulation.

B. Bias Mitigation Measures in the AI Act

The AIA requires high-risk AI systems to undergo conformity assessments before deployment, ensuring they meet established standards for bias mitigation.[168] These assessments are a preventative measure, ensuring only AI systems that have effectively addressed potential biases are introduced to the market. An AI system will obtain a positive result insofar as the below requirements are met. This robust framework is aimed at enhancing the ethical use of AI technologies, with a strong emphasis on reducing bias.[169]

For starters, the AIA mandates strict data governance and the use of high-quality, unbiased data.[170] This initiative ensures that the datasets fed into AI systems are as accurate and representative as possible, thereby reducing the risk of biases being encoded into AI algorithms from the outset.[171] Through the institution of comprehensive risk management systems, the AIA enforces the early identification and mitigation of potential biases within AI systems. This preemptive approach ensures that biases are addressed at their root, which significantly decreases the possibility of biased outcomes.

---

166. Raposo, *Ex Machina, supra* note 153; Raposo, *Houston*, supra note 153, at 68-69.

167. *Artificial Intelligence Act Article 3*, ACCESIBLE L., https://artificialintelligenceact.com/title-i/article-3/ (last visited Mar. 27, 2024).

168. EUR. PARL. DOC. P9_TA(2024)0138, *supra* note 160, ch. III, art. 43.

169. *See* Hadrien Pouget & Ranj Zuhdi, *AI and Product Safety Standards Under the EU AI Act*, CARNEGIE ENDOWMENT FOR INT'L PEACE (Mar. 5, 2024), https://carnegieendowment.org/2024/03/05/ai-and-product-safety-standards-under-eu-ai-act-pub-91870; *see also* Raposo, *Look at the Camera*, *supra* note 156.

170. *See* EUR. PARL. DOC. P9_TA(2024)0138, *supra* note 160, ch. III, art. 10.

171. IÑIGO DE MIGUEL BERIAIN ET AL., EUR. PARL. RSCH. SERV., AUDITING THE QUALITY OF DATASETS USED IN ALGORITHMIC DECISION-MAKING SYSTEMS 37–39 (2022), https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729541/EPRS_STU(2022)729541_EN.pdf.

Moreover, the AIA demands detailed technical documentation[172] and rigorous record-keeping.[173] This requirement enhances the transparency and accountability of AI systems by providing insight into their decision-making processes, which in turn facilitates the identification and correction of biases.[174]

Additionally, human oversight is a critical component of the AIA, ensuring that AI systems are continuously monitored and evaluated by individuals capable of contextual understanding and ethical judgment processes where AI systems may falter.[175] This human layer of analysis and intervention, in conjunction with other mitigation techniques, is key to identifying and rectifying biased decisions that AI systems make. While AI systems can process information and make decisions at scales and speeds unattainable by humans, they cannot fully understand ethical nuances and societal contexts. Human intervention cannot completely bridge this gap but it can spot some biased outcomes and correct them.[176] By prioritizing the accuracy, robustness, and cybersecurity of AI systems, the AIA indirectly combats bias by ensuring systems perform reliably and are safeguarded against external manipulations that could introduce or amplify biases.[177]

The AIA also significantly focuses on making AI system operations transparent, traceable, and explainable.[178] The emphasis on transparency, traceability, and explainability ensures that stakeholders can understand the decision-making process, thereby enabling the identification and rectification of biases.[179] Traceability and transparency

172. EUR. PARL. DOC. P9_TA(2024)0138, *supra* note 160, ch. III, art. 11.

173. *Id.* ch. III, art 12.

174. Ferrara, *supra* note 140, at 3.

175. EUR. PARL. DOC. P9_TA(2024)0138, *supra* note 160, ch. III, art. 14.

176. For example, in the specific case of predictive AI for recidivism, recurrent audits led by human experts are important to analyse the outcomes and detect concerning trends, such as the recurrent provision of high prediction rates to individuals pertaining to a marginalized group. On the importance of audits see Corinna Herweck et al., *FairnessLab: A Consequence-Sensitive Bias Audit and Mitigation Toolkit*, *in* EWAD'23: EUROPEAN WORKSHOP ON ALGORITHMIC FAIRNESS (2023). *See also* Ferrara, *supra* note 140, at 10.

177. Gabriele Carovano & Alexander Meinke, *Improving Fairness and Cybersecurity in the Artificial Intelligence Act*, *in* EWAD'23: EUROPEAN WORKSHOP ON ALGORITHMIC FAIRNESS (2023), https://ceur-ws.org/Vol-3442/paper-43.pdf.

178. EUR. PARL. DOC. P9_TA(2024)0138, *supra* note 160, ch. III, art. 13.

179. *See generally* Ali et al., *supra* note 122.

further support how the AIA requires replicability and true understanding by ensuring the entire decision-making process, from data selection to algorithm application, is open to scrutiny.[180] This allows others to attempt to reproduce the process through simulations and evaluate the system's veracity.

CONCLUSION

Biases in AI predictive modeling—specifically in AI used in risk assessment and recidivism prediction—are impossible to extinguish, especially when applied in a criminal setting. The problem does not lie in the prediction itself but in the underlying processes used to make AI predictions that suffer from the same types of biases as humans do because they view the past as indicative of the future. Biased predictions are inevitable because human biases taint the historical data used to make those predictions.[181] Therefore, one can logically conclude that AI predictions are no different considering AI decision-making relies tremendously on human decision-making.

The worry lies with the impact of AI on people. AI's ability to affect—both positively and negatively—many more individuals than one human's decision-making, and at a much faster rate, is not to beunderestimated. Instead of having an individual impacted by biased human decision-making, thousands risk being affected by AI that feels no remorse or repercussions. Still, the root of the problem is not AI itself, but its prediction methods relying on biased data.

> The deep problem is the nature of prediction itself. All prediction looks to the past to make guesses about future events. In a racially stratified world, any method of prediction will project the inequalities of the past into the future. This is as true of the subjective prediction that has long pervaded criminal justice as it is of the algorithmic tools now replacing it.[182]

---

180. *Id.*

181. Naroa Martínez et al., *Human Cognitive Biases Present in Artificial Intelligence*, 67 REVISTA INTERNACIONAL DE LOS ESTUDIOS VASCOS [REV. INT. ESTUD. VASCOS.] 1, 6–10 (2022).

182. Mayson, *supra* note 80, at 2218.