CRIMINAL JUSTICE IN THE AGE OF AI:

ADDRESSING BIAS IN PREDICTIVE ALGORITHMS USED BY COURTS

Rahulrajan Karthikeyan, Chieh Yi, and Moses Boudourides

Rahulrajan Karthikeyan, Arizona State University, rkarthi5@asu.edu

Chieh Yi, Arizona State University, chiehyi@asu.edu

Moses Boudourides, Northwestern University and Arizona State University, Moses.Boudourides@northwestern.edu

Author/s Biographies: Moses Boudourides served as Professor of Practice at the Arizona State University School of Public Affairs and is currently a member of the faculty of the Northwestern University School of Professional Studies Data Science Online Graduate Program. Previously, he held the position of Professor of Computational Mathematics at the University of Patras in Greece and the Department of Electrical and Computer Engineering at the Democritus University of Thrace, also in Greece. His expertise lies in Applied and Computational Mathematics, Network Science, and Computational Social Science. Rahulrajan Karthikeyan and Chieh Yi are graduate students enrolled in the Arizona State University School of Computing and Augmented Intelligence Computer Systems Engineering (CSE) program. During the spring semester of 2023, Karthikeyan and Yi attended the CSE 575 course on Statistical Machine Learning, which was taught by Boudourides. Subsequently, they have continued to collaborate with him on various machine learning projects.

ABSTRACT:

As AI and machine learning become increasingly integrated into daily life, both individuals and institutions are growing dependent on these technologies. However, it's crucial to acknowledge that such advancements can introduce potential flaws or vulnerabilities. A case in point is the investigation conducted by the non-profit organization ProPublica into the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) risk assessment tool—a tool widely used by US courts to assess the likelihood of a defendant reoffending. To address the issue of underlying biases, including racial biases, which can lead to inaccurate predictions and significant social harm, we are delving into the current literature on algorithmic bias in decision systems. We are also exploring the evolving considerations of fairness and accountability in machine learning. Specifically, within the realm of predictive policing algorithms employed in the criminal justice system, our focus is on recent studies aimed at mitigating biases in algorithmic decision-making. This involves reassessing recidivism rates and implementing adversarial debiasing in conjunction with fairness metrics.

KEYWORDS:

1. Bias
2. Fairness
3. Algorithmic decision
4. Criminal justice
5. Recidivism
6. Machine learning

Main Body:

1. Introduction

The concept of justice has long been regarded as a fundamental tenet of any society, wherein impartiality and objectivity are expected to guide judicial decision-making. However, the inherent biases of human judgment have historically undermined the efficacy of the criminal justice system, leading to disparities and inequalities in the administration of justice. That was why, in recent years, there was an increasing interest and research on a category of explorations of Artificial Intelligence (AI) and Machine Learning (ML) that are now referred to as "algorithmic fairness" or "fair ML" (Corbett-Davies and Goel, 2018; Cowgill & Tucker, 2019; Kleinberg et al., 2018; Barocas & Selbst, 2016). In particular, our aim here is to examine the use of AI and ML in addressing biases in the criminal justice system, highlighting their future.

The COMPAS algorithm serves as a notable example of potential biases inherent in AI-driven risk assessment tools used in the criminal justice system. As a matter of fact, COMPAS is a software tool, which uses various algorithms and data analysis techniques to assess a defendant's likelihood of reoffending (risk of recidivism) and it helps guide decisions related to criminal sentencing/supervision/parole, etc. Typically, COMPAS is processing datasets of information about defendants' backgrounds including several different attributes/features of individuals such as age, gender, criminal history, etc. Nevertheless, COMPAS has been found to exhibit significant racial and gender biases, with black and female defendants disproportionately affected. In 2016, an investigation by ProPublica revealed that the dataset was based on prior judgments made by biased judges, undermining the tool's objectivity and fairness. These findings underscore the need for greater scrutiny of AI and machine learning applications in the criminal justice system to ensure that they do not perpetuate or amplify existing biases.

This contribution aims to discuss existing fair and unbiased models of analysis in the case study of the COMPAS tool, given that the fairness and accuracy of algorithms applied to justice have frequently raised doubts among the public (Angwin et al., 2016). Before arguing about COMPAS, we are examining some common domain areas in which decision-making algorithms are habitually adopted. Additionally, we are conducting a literature review on existing/previous risk assessment tools.

Within the computational framework, while being engaged in replicating the existing analysis of the COMPAS recidivism algorithm (Larson et al., 2016), we have been employing a special toolkit AI Fairness 360 (AIF360) and the technique called adversarial debiasing. In principle, such an ML analysis involves two trained models: the first one is a Random Forest classifier and the other is the AIF360 debiasing model. Furthermore, a method called in-processing adversarial debiasing is often employed to make the model fair and unbiased even when some biases exist in the data. As it is done in ML processing, one needs to divide the data into two parts: a training set and a test set. This division helps the analyst to see how accurate and fair each model is and to be able to employ fairness metrics for the understanding of models that treat different groups of people fairly. Moreover, these metrics facilitate measuring the balance between fairness and accuracy. In this way, one can assess whether the model is making fair decisions for everyone while still being accurate in its predictions. By using these tools, one can build a better and more trustworthy model.

Finally, to make the results even more reliable and understandable a technique called LIME is often employed.

Studies, such as ProPublica's scrutiny of the COMPAS tool, have important implications for the use of artificial intelligence and machine learning in the criminal justice system. The approaches that we are discussing here show that one may use several techniques to overcome racial bias in the data so that one may create fair models that are still very accurate.

2. Domains of Algorithmic Decision Systems

In the rapidly advancing digital age, the integration of algorithmic decision systems (ADS) into various facets of our lives powered by machine learning and artificial intelligence holds the promise of enhancing efficiency, objectivity and being cost-effective in decision-making processes across multiple domains. However, as the adoption of algorithmized processes continues to expand, an increasingly urgent question emerges: Can we genuinely trust algorithmic decision systems to make critical choices on our behalf? (Emspak, 2016) This is a concern, which is not merely theoretical. As has been already argued (Vartan, 2019), algorithms employed in ADS have the potential to perpetuate and even amplify society's deeply ingrained biases, contributing to widespread inequities. However, challenges involved in the practice of algorithmic decision-making systems extend far beyond this well-documented problem.

Incidentally, a recent commentary (Ananya, 2023) has focused on a previously underexplored aspect of their functioning in terms of how sensitive ADS are in minuscule variations of the way that humans annotate data used to train them. The latter issue delves into the complex landscape of ADS and its implications for various sectors from employment and healthcare to criminal justice and entertainment. It is an issue related to how these systems are already shaping who gets hired, who receives priority in medical care, how bail is determined, and even what content is consumed. While algorithmic systems promise to expedite decision-making, alleviate backlogs, and provide objective evaluations, there are plenty of news reports and many research findings uncovering the existence of certain alarming shortcomings in ADS, which can possibly have profound and long-lasting consequences in the lives of individuals affected by algorithmic judgments.

A recent study (Balagopalan et al., 2023) provides a particularly insightful perspective, which reveals how the nuances of human annotation during training can significantly influence algorithmic decisions. By examining how training in machine learning models is implemented to assess rule violations (such as dress code adherence or meal standards etc.), Balagopalan et al. have uncovered some cases in research exhibiting a remarkable disparity in the outcomes of the employed models. Even when both versions of an algorithm were trained on the same rules, their judgments diverged, based on whether humans provided descriptive labels or directly evaluated the rule violations. This discrepancy, it turns out, reflects the fact that, when confronted with potential consequences of their decisions, humans may label data in multiple ways. Furthermore, the above discrepancy highlights the broader implications of relying on historical data for training algorithms. As ADS tend to infer the future based on past data, they may inadvertently perpetuate past biases and fail to adapt to evolving societal values. Thus, the following fundamental question emerges: Can algorithmic decision-making systems truly imagine a different future when their foundation is rooted in the past?

When decisions are solely based on patterns and principles derived from past data and predefined rules, ADS, unlike humans, are doomed to be inherently devoid of imaginative capacity and creativity. Although algorithmic decisions process historical data to make informed choices in the present, their ability to foresee a different alternative future is limited. Aspiring to make ADS consider a different future inadvertently necessitates adjustments to the employed underlying algorithms and data inputs, which are typically made available through human intervention. In other words, the stake of algorithmic fairness lies in a feasible rectification of machine learning algorithms that typically learn from an already biased past. Because ADS trained on questionable past data are prone to make unfair decisions possibly propagating discrimination in future cases (Pfeiffer et al., 2023).

Moreover, from an epistemic point of view, Holm (2023) is examining a related but neglected problem of algorithmic decision-making about how appropriate it is to allocate resources based on purely statistical evidence provided by algorithmic predictions. Stated formally, let us give the example when a malevolent event Y occurs, in such a way that all but one among N individuals might have played a role in Y. Thus, anyone among the N individuals has a probability of guilt equal to 0.99. Now, suppose the individual X is picked at random. Concluding that X was guilty, based only on the available purely statistical evidence, would be relying on a so-called "actuarial inference." However, is an actuarial inference an adequate basis for punishing X? Most people would answer no because purely statistical evidence is epistemically deficient and inadequate to justify a verdict. This is not a criticism against the use of statistical generalizations and probabilistic inferences as such, but rather it is a criticism of the epistemic value of statistical evidence as justification for believing propositions of the form "X is guilty of Y" (Holm, 2023, p.28).

After having reviewed some of the general opportunities and risks related to the use of ADS and discussed the need to take into consideration other legal, ethical and social dimensions too, in what follows in this section, we are going to delve deeper into the findings of four different sectors of applications of ADS: (i) education, (ii) healthcare, (iii) employment, and (iv) bank lending, leaving aside the case of criminal justice, which is going to be examined separately on its own in the next section. From shaping the allocation of resources and opportunities to influencing the delivery of services, algorithmic decision systems reach far and wide. Moreover, they are posing intricate challenges and ethical considerations. Surveying through the literature in these four sectors, we are aiming to gain a comprehensive understanding of many existing or potential benefits and, at the same time, to highlight the need for critical assessment in the deployment of ADS across diverse sectors of human lives.

(i) Education is a sector in which ADS has gained prominence, particularly in the admissions process intending to enhance efficiency and data-driven decision-making. However, algorithmic decision-making in education has also introduced the risk of bias, which can result in unfair outcomes, particularly for underrepresented groups. In what follows below, the significance of recognizing and addressing biases in educational institution admissions is underscored. Kizilcec and Lee (2022) offer the following valuable insights into bias and fairness issues within education decision algorithms.

As algorithmic bias in education might arise when machine learning models are trained on non-representative student data, it can result in models prone to making inaccurate predictions for specific student groups, such as the ones with minority backgrounds or students with disabilities. ADS also impacts the admissions process in educational institutions. For instance, low-income applicants who often have lower SAT scores and college GPAs may face higher rejection rates, and a model trained to predict student college success may unjustly predict failure for low-income students, solely because the employed training data might indicate lower graduation rates for this group (Kizilcec and Lee, 2022, pp.174-102).

To mitigate algorithmic bias in education, several actions can be taken. First, it is crucial to ensure that the employed training data for machine learning models is representative of the entire student population. Achieving this involves data collection from diverse sources and oversampling data from underrepresented groups. This also entails assessing model performance across various student groups and identifying bias patterns. Once identified, steps can be taken to mitigate bias such as adjusting model parameters or employing post-processing techniques on model output (Kizilcec and Lee, 2022, pp.174-102).

(ii) Biomedical research and healthcare are some of the most important areas of applications of AI. In an important recent paper, Baumgartner et al. (2023) shed light on the challenges and opportunities that AI presents in these fields, emphasizing the importance of social science perspectives to ensure fairness, equity, and ethical use of AI.

In this perspective, one of the central concerns raised is the perpetuation of inequalities due to biased data and models. Biomedical AI systems heavily rely on data for their functioning and if the data used is biased or unrepresentative these systems can lead to inaccurate predictions and decisions. Demographic factors, such as race, gender, and socioeconomic status, play a significant role in shaping health outcomes and when not properly accounted for they can lead to health disparities. For instance, as Baumgartner et al. (2023, p.3) argue, AI algorithms that are not trained on diverse datasets may struggle to accurately assess the needs of minority and marginalized groups.

Discrimination, exclusion, prejudice, and stereotyping are deeply rooted in social inequalities, which can also affect the algorithms used in healthcare AI (Baumgartner et al., 2023, p.4). Consequently, there is always the risk that processing, labelling, and classification of data may inadvertently target certain populations while neglecting others. Thus, Baumgartner et al. (2023, p.4) advocate for a justice-oriented design of AI algorithms, emphasizing the need for independent oversight and auditing by experts trained in science and technology studies (STS) to ensure that AI systems are fair, accountable, and transparent.

Transparency in AI systems is another critical concern, in cases when many AI algorithms used in healthcare are complex and difficult to audit or explain (Baumgartner et al., 2023, p.4). The authors of this paper conclude that the lack of transparency may leave both healthcare providers and patients in the dark about how decisions are made and this is something that may erode trust in AI systems. Therefore, the authors stress the importance of making AI systems more explainable and understandable for users.

As Beaulieu and Leonelli argue in their book (2021), the issue of distorted data and human errors in interpreting results is also a growing concern. It highlights the limitations of binary classifications particularly in gender and sex classifications and the need to recognize that neither sex nor gender is binary (Baumgartner et al., 2023, p.4). Hence, overcoming these limitations requires a more nuanced approach to data collection and analysis.

Furthermore, within this realm, Chen et al. (2021, p.2) insist on acknowledging the persistent challenge of discrimination based on race and gender during the evaluation of such critical factors as image acquisition, genetic variation, and intra-observer labelling variability in current clinical workflows. These discriminatory practices may inadvertently introduce bias into the decision-making processes. As one harnesses the transformative potential of machine learning and artificial intelligence within the medical field, it is incumbent to shift the focus toward identifying and addressing design flaws inherent in AI algorithms. These flaws require meticulous handling and appropriate mitigation strategies.

Therefore, in a nutshell, many scholars emphasize the urgency of addressing fairness in AI for biomedical research and healthcare calling for an interdisciplinary collaboration (including perspectives from social sciences, race and gender studies, STS, and medical ethics) to develop and apply AI in a way that is equitable and beneficial for all (Baumgartner et al., 2023, p.7).

(iii) In the literature on algorithmic fairness in information systems, when it comes to issues of employment and processes of job hiring (Rieskamp et al., 2023), gender biases remain a persistently recurrent issue. For instance, Lavanchy (2018) has revealed a tendency among hiring managers to prefer male applicants over their female counterparts. Such a deeply ingrained bias raises serious concerns regarding fairness and equality in the recruitment process.

Nonetheless, the adoption of AI-based algorithms in Human Resource (HR) departments of organizations represents a significant advancement in streamlining and expediting the candidate evaluation process. In computerized HR management systems, resumes and applications are subject to a certain initial machine assessment before undergoing comprehensive manual review by human recruiters and reviewers. Notwithstanding, there are serious concerns among computer scientists and advocates of fair hiring practices on whether algorithmic hiring may make problematic decisions (Langenkamp et al., 2019).

That gender bias influences personnel selection processes has been already documented in numerous field experiments (for the Spanish case study, Martinez, 2021), which examine the causes resulting in gender inequalities and discrimination in job access, hiring decisions, selection of leaders, salaries, etc. (World Bank, 2020).

Thus, recent research and systematic reviews in the field of HR recruitment and development have shed light on the challenges, causes, and consequences of algorithmic decision-making. Algorithms have been shown to produce discriminatory or biased outcomes when trained on inaccurate or invasive exploitation of sensitive personal data (Kim, 2017), on societally prejudiced cases (Barocas and Selbst, 2016), or input data instrumental in representational harm (Suresh and Guttag, 2021). As a matter of fact, all these issues underscore the vulnerability of algorithms to replicate biased decisions if their training data is inherently flawed or biased.

Moreover, the employment of Natural Language Processing (NLP) and Machine Learning tools in the evaluation of such data as resumes or interview transcripts has demonstrated biases against women and individuals with disabilities (Engler, 2021). Speech recognition models have exhibited clear biases against African Americans and have struggled with dialectical and regional variations in speech (Koenecke et al., 2020). The use of commercial AI facial analysis has revealed disparities across skin colour and raises significant concerns, particularly for individuals with disabilities (Engler, 2019). Algorithms employed in the advertisement of job postings have been found to unintentionally lead to biased outcomes, including bias against young women seeking STEM jobs, as well as ageism against older candidates (Lambrecht and Tucker, 2019). The cumulative impact of all these algorithmic biases on hiring decisions is particularly troubling since the accumulation of small biases within algorithms can create larger structural effects.

Ensuring the fairness of AI systems in job hiring is paramount, especially considering that job applicants can be directly rejected by certain inadequate algorithmic systems. Applicants' data attributes like sex, ethnicity, or age may be responsible for biases harming and potentially disfavouring certain groups of individuals (Chakraborty et al., 2021). Biases in AI-driven hiring processes can stem from the so-called cognitive biases, such as the microeconomic home bias, similarity bias in recruitment, or some other stereotypes that can lead recruiters to favour specific applicants over others (Rieskamp et al., 2023). Measurement bias relates to general errors in data collection, while representation bias occurs when data fails to accurately depict the relevant population resulting in an underrepresentation of certain groups (Suresh and Guttag, 2021, pp.4-5).

Algorithms can also exhibit bias due to specific optimization functions or the use of biased estimators (Baeza-Yates, 2018). Historical biases inherited through the transition to AI-based algorithms can further exacerbate these issues (Australian Human Rights Commission, 2020).

Efforts to mitigate gender and ethnicity bias in AI-based hiring processes can be implemented throughout various stages of data processing. According to Rieskamp et al. (2023, p.223), the following checklist of computational actions might ensure algorithmic fairness:

- Pre-processing involves costly efforts to enhance accuracy and denoise protected attributes.
- In-process tweaks introduce adversarial networks to mitigate gender and ethnicity bias with evidence suggesting the existence of increased fairness, albeit at the cost of reducing model accuracy.
- Post-processing approaches aim to rectify bias by reranking processed data to include candidates from protected groups. For instance, in the case of gender bias, high-ranked candidates can be combined with a gender distribution over qualified candidates to achieve a new and fair ranking.
- Feature selection methods have also been employed to remove features containing information about candidates' gender, resulting in a substantial reduction in bias.

In a nutshell, the previous scrutiny of the existing literature suggests that the prevalence of gender bias in hiring processes necessitates a concerted effort to leverage AI technologies to eliminate biases and promote fairness and equity in recruitment. Organizations must adopt strategies and techniques at various stages of the hiring process to ensure that AI-based algorithms treat all candidates equally regardless of their demographic (or other) personal attributes. By addressing these biases, one can move closer to achieving fair and equitable hiring practices that are solely based on candidates’ qualifications and abilities.

(iv) In the field of bank lending and mortgage credit, racial bias casts a long shadow over bank lending, resulting in unequal treatment among different ethnicities (Counts, 2018). This bias becomes evident when examining the disparities in mortgage credit approval rates for various racial groups. For instance, individuals of European descent often find themselves with a higher likelihood of mortgage approval compared to their counterparts of Asian descent (Martinez & Kirchner, 2021).

Bhutta et al. (2022) have unveiled troubling disparities in mortgage approval recommendations and processes. They discovered that Black and Hispanic applicants tend to face significantly lower mortgage credit scores compared to their white counterparts. Additionally, their research revealed that Black and Hispanic applicants are less likely to secure loans from government-backed sources. This disheartening pattern is further underscored by a persistent Black-white denial gap of two percentage points alongside residual gaps for Hispanic and Asian applicants of approximately one percentage point. These residual disparities are aptly referred to as excess denials which further highlights the existence of racial bias within mortgage lending practices. While algorithms have been increasingly employed in the lending process, it is crucial to recognize that they too must adhere to fair lending regulations. This means that these algorithms cannot factor in race or ethnicity or proxies such as neighbourhood location or ZIP code. While the integration of algorithms into lending processes has the potential to mitigate discrimination compared to face-to-face lenders it falls short of completely eliminating discrimination in loan pricing (Bhutta et al., 2022, p.7).

Bartlett et al. (2019) have shed additional light on this issue. They have estimated that lenders exhibit discriminatory practices against Latinx and African-American applicants, resulting in higher rejection rates and the imposition of higher percentage interest rates for purchase mortgages. This alarming revelation underscores the systemic nature of racial bias in the way that ADS operates within the lending industry. Furthermore, the authors have argued that, even with the advent of algorithmic lending, discrimination remains a persistent issue, especially concerning loan pricing (Bartlett et al., 2019, pp.2-5). Lenders must adhere to a stringent set of criteria to ensure that their lending practices are fair and unbiased. This includes demonstrating that variables such as high school attendance were correlated with historical data related to fundamental lifecycle variables such as income growth and they did not predict protected characteristics after orthogonalizing them to the lifecycle variables (Bartlett et al., 2019, p.22).

After having described how algorithmically-driven decision-making systems are employed in four representative sectors and having discussed certain problems caused by the inherent biasing that algorithms used in these systems may display, we are moving to our main focus, which is ADS in criminal justice. For this purpose, we are starting, in the next section, with a discussion of the questionable role that algorithms of recidivism prediction play in certain frequently utilized assessment tools in the criminal justice system.

3. Recidivism risk assessment instruments

Recidivism risk assessment instruments serve as pivotal tools in the criminal justice system aiding judges/jury in predicting the likelihood of defendants reoffending after their initial encounter with the legal system. In this respect, Fazel et al. (2022) have identified the following eleven widely used risk assessment instruments in criminal sentencing.

(i) One of the first instruments of this sort was the COMPAS tool, which was developed by the private company Northpointe Inc. (now Equivant), where the acronym COMPAS stands for “Correctional Offender Management Profiling for Alternative Sanctions.” This is a standard instrument used in the US criminal justice system to forecast a defendant's likelihood of reoffending (recidivism) and it has relied upon the sentencing decisions made by certain judges. Moreover, COMPAS has attracted the ground-breaking investigative journalism work of the non-profit organization ProPublica (Barenstein, 2019). Since the COMPAS tool is going to be discussed separately in the next section, we are now examining the remaining ten tools in more detail.

(ii) Douglas, Hart, Webster, and Belfrage (2013) have introduced the assessment tool of Historical Clinical Risk Management-20 (HCR-20), which is a comprehensive instrument that amalgamates historical clinical and risk management factors to gauge the potential for violent or aggressive behaviour. This is a multidimensional approach acknowledging the significance of a holistic comprehension of an offender's background and circumstances in predicting the risk of recidivism. HCR-20 evaluates variables such as past criminal activities, mental health, and contextual triggers. Thus, it offers a nuanced perspective on a complex interplay among factors contributing to tendencies of reoffending.

(iii) Latessa, Lovins, and Makarios (2013) delve into the Indiana Risk Assessment System (IRAS), which is made up of five instruments designed to assess recidivism risk and guide intervention strategies. This tool takes a comprehensive approach by considering various elements including criminal history, behaviour, and the social environment. Notably, the IRAS incorporates dynamic factors such as an individual’s participation in treatment programs, thus, recognizing the offender's potential for positive change in evolving circumstances.

(iv) Andrews, Bonta, and Wormith (2004) have introduced an essential recidivism risk assessment instrument, called Level of Service/Case Management Inventory (LS/CMI). This instrument takes into account a range of dynamic and static factors, thus, providing a holistic perspective on an offender's risk profile. Its comprehensive nature allows it to evaluate aspects such as criminal history, family and marital relationships, employment, substance abuse, and offender’s attitude and orientation. LS/CMI is particularly notable for its emphasis on case management and intervention planning. Indeed, it goes beyond risk assessment by guiding professionals in developing tailored case management strategies based on the identified risk factors. In this instrument, Andrews, Bonta, and Wormith (2004) developed the Risk-Need-Responsivity (RNR) model emphasizing the importance of addressing an individual's criminogenic needs to reduce the likelihood of reoffending effectively. In this way, LS/CMI becomes a valuable asset in promoting various interventions (mostly, cognitive-behavioural ones), which may differ in the corresponding degree of effectiveness to reduce recidivism. Furthermore, the model of LS/CMI may encompass various attributes of offenders, such as gender, cognitive capacities, motivations, etc.

(v) Andrews and Bonta (1995) have introduced the actuarial assessment tool, Level of Service Inventory-Revised (LSI-R). Like LS/CMI, LSI-R takes a comprehensive approach, taking into account a range of dynamic and static factors providing again a holistic perspective on an offender's risk profile. It acknowledges that an individual's circumstances and behaviours can change over time. Therefore, LSI-R incorporates ongoing assessment and reassessment ensuring that interventions remain relevant and effective in reducing the risk of recidivism. Like LS/CMI, its approach aligns with the Risk-Need-Responsivity (RNR) model and it also may develop tailored case management strategies on the identified risk factors.

(vi) Garrett, Jakubow, and Monahan (2019) established the Nonviolent Risk Assessment (NVRA), which is a valuable addition to the field of risk assessment tools within the criminal justice system. This tool is specifically designed to provide a targeted approach in the assessment of risk factors associated with nonviolent or low-level offending conduct (nonjail alternatives). In this context, nonviolent behaviour is understood to include factors, such as mental health, substance abuse, family, social considerations, and individual behaviour patterns. While NVRA excels in assessing nonviolent risks, it may require complementary tools or interventions for case management and rehabilitation.

(vii) Howard (2006) has developed the Offender Assessment System (OASys) used by the prison and probation services in England and Wales. This is an assessment tool that evaluates both the risk of harm and it identifies–classifies offending-related needs, including basic personality characteristics and cognitive behavioural problems needs of offenders. OASys focuses on analysing dynamic factors such as attitudes, employment status, and substance abuse, which can significantly influence an individual’s likelihood of reoffending. It underscores the importance of tailored interventions that address specific needs, which ultimately aim to reduce the risk of reoffending by addressing underlying factors contributing to criminal behaviour.

(viii) Latessa, Lemke, Makarios, and Smith (2010) have discussed the Ohio Risk Assessment System (ORAS), which is a state-wide comprehensive instrument tailored to assess recidivism risk of adult probationers. This system offers a structured framework to guide supervision strategies, thus, ensuring that resources are allocated effectively to mitigate the risk of reoffending. ORAS elaborates on the notion that interventions should be targeted based on an individual's risk profile aligning with the broader principle of risk-needs-responsivity.

(ix) The Canadian forensic psychologist Robert D. Hare (1993) has elaborated the Psychopathy Checklist-Revised (PCL-R), an essential assessment tool in the field of forensic psychology and criminology. It is a tool specifically designed to assess individuals diagnosed with psychopathy vs. ones diagnosed with antisocial personality disorder. PCL-R is based on a detailed checklist of clinical cases of psychopathy (Huchzermeier et al., 2007). It is a multidimensional approach to assess various factors including interpersonal, affective, and lifestyle dimensions to determine the presence of severity of psychopathy in individuals. What sets PCL-R apart is its utility in assessing individuals who may pose a higher risk of engaging in serious criminal behaviour due to psychopathy. The PCL-R yields a total score that ranges from 0 to 40 with higher scores signifying a heightened level of psychopathy (Hare, 1993). In professional assessment, a total score within the range of 10 to 19 on the PCL-R denotes a mild degree of psychopathy, while a score falling between 20 and 29 indicates a moderate level and a score between 30 and 40 signifies severe psychopathic. This finely-tuned scoring system aids the precise categorization of individuals based on the severity of their psychopathy and provides valuable insights for risk assessment and intervention strategies (Scarlet, 2011).

(x) Lowenkamp, Johnson, Holsinger, van Benschoten, and Robinson (2011) have developed the Post Conviction Risk Assessment (PCRA) tool, which is a valuable instrument in the hands of the federal supervision system in the USA. This instrument is designed to assess the risk of recidivism in individuals, who have already been convicted and sentenced and it plays a role in informing decisions related to parole, re-entry programs, and supervision. It takes into account a wide range of factors including an individual's criminal history, institutional behaviour, and other dynamic variables that may influence the likelihood of reoffending. This approach seamlessly aligns with the Risk-Need-Responsivity (RNR) model that emphasizes the importance of addressing an individual's criminogenic needs for successful reintegration into society.

(xi) Another actuarial tool for assessing recidivism risk was introduced by Helmus, Thornton, Hanson, and Babchishin (2011): the Static-99 Revised (Static-99(R)), an updated iteration of the original Static-99, which retains the core principles of its predecessor. Static-99(R) addresses a critical need in risk assessment by evaluating unalterable factors contributing to an individual's risk of recidivism, specifically in cases of individuals with prior sexual offense convictions, and it incorporates refinements and adjustments informed through empirical research and clinical insights (Melton et al., 2018). As it is described in the manual of the third published version of the Static-99 (Phenix et al., 2016), what sets the Static-99(R) apart is its focus on evaluating static risk factors, such as an individual's criminal history, age at release and victim gender. By emphasizing these unchangeable variables, Static-99(R) provides a valuable perspective on an individual's long-term risk profile. This approach complements other tools that assess dynamic risk factors thus enabling a comprehensive evaluation of an individual's overall risk. Overall, the significance of the Static-99(R) lies in its potential to inform decisions related to sentencing, parole, and post-conviction supervision, thereby contributing to the safety and fairness of the criminal justice system (Phenix et al., 2016, p.30 and p.35). In terms of offenders' relative risk for sexual recidivism, the Static-99(R) utilizes readily available demographic and criminal history information known to correlate with sexual recidivism among adult male sex offenders. In this way, such a categorization facilitates law enforcement in developing tailored supervision strategies for each individual (Phenix et al., 2016, p.10, p.13, and p.23). The main strengths of the Static-99(R) include setting up several explicit rules and following strictly the relative objectivity of a scoring system. The total risk score is determined through specific risk factors outlined in a scoring table (given as the tally sheet on page 99 of Phenix et al., 2016) that provides a ranked assessment of each sexual offender. Hence, by adhering to the coding manual and the established rules, evaluators can achieve a more objective and less biased decision-making process and, eventually, contribute to enhancing the tool's reliability.

In a nutshell, concluding the discussion of the above eleven assessment instruments, one may remark that they all provide valuable insights into the realm of recidivism risk assessment. Overall, these tools offer multifaceted perspectives on risk, encompassing historical, clinical, dynamic, and comprehensive factors. While each instrument approaches risk assessment both from a distinct angle and collectively, these eleven tools underscore the importance of holistic evaluations and tailored interventions in promoting a fairer and more effective criminal justice system.

4. ProPublica's study COMPAS

Going back to the first of the eleven instruments, COMPAS, let us mention that in 2016, ProPublica, a non-profit news organization specializing in investigative journalism, investigated the use, fairness, and accuracy of the COMPAS algorithm (Larson et al., 2016). ProPublica's investigation revealed that the tool was racially biased against African-American defendants, resulting in harsher and more frequent sentences. As a matter of fact, in their investigation of the COMPAS algorithm's predictions for over 10,000 defendants in Broward County, Florida, ProPublica discovered troubling evidence of racial bias (Angwin et al., 2016, p.2). ProPublica's analysis revealed that the algorithm disproportionately labelled black defendants as high-risk for reoffending, even if they did not actually re-offend, while white defendants were more likely to be labelled as low-risk, even when they did go on to commit further crimes (Angwin et al., 2016, p.2).

Moreover, the examination of the COMPAS instrument by ProPublica uncovered that the accuracy of the algorithm was significantly lower for female defendants, particularly for black women who were disproportionately misclassified as high-risk (Larson et al., 2016, p.5). These findings reveal the gender and racial biases inherent in the algorithm and underscore the urgent need to identify and address such biases in machine learning tools employed in the criminal justice system. In particular, ProPublica has shown that 45% of black defendants were misclassified as high-risk compared to only 23% of white defendants. In addition, white defendants who reoffended were misclassified as safe 48% of the time compared to only 28% for black defendants (Angwin et al., 2016, p.7). These differences are even more significant when controlling for relevant criminal history, age, and gender, with black defendants being 45% more likely to be given a higher risk score than their white counterparts (Angwin et al., 2016, p.7).

ProPublica's investigation prompted significant attention and generated discussions about the implementation of algorithms in the criminal justice system (Washington, 2019, pp.11-12). Moreover, it illuminated the importance of transparency and accountability when it comes to decision-making based on algorithms, particularly concerning sensitive issues such as criminal justice. The results of the investigation raised concerns regarding the fairness of algorithms and the potential for unintended bias and discrimination, emphasizing the need for further research and development of more equitable machine learning practices.

Nonetheless, as has been already noted (Barenstein, 2019), there exists a significant flaw in the COMPAS system. Rather than mitigating the systemic racism inherent in the American criminal justice system, it served to perpetuate it, since the system was trained on data from biased judicial decisions, it effectively reinforced existing biases (Park, 2019). Judges who sought to remain impartial relied on COMPAS's predictions, but in doing so unwittingly replicated the flawed decision-making of their biased peers, something which amplified the issue, leading to even greater disparities in the system (Larson et al., 2016).

Thus, the biased data used to train the COMPAS model resulted in flawed decision-making in the criminal justice system. For instance, black defendants charged with drug possession were more likely to be labelled as high-risk compared to white defendants with a more serious criminal history (Angwin et al., 2016, p.4). Additionally, a black female with only juvenile misdemeanours and no new crimes after two years was rated as a higher risk than a seasoned criminal with prior convictions and a lower risk of reoffending (Angwin et al., 2016, pp.1-2). The use of COMPAS in the criminal justice system highlights the grave danger of turning a blind eye toward the biased data that might have been used to train the model, regardless of the good intentions behind its development.

ProPublica's rigorous investigation of COMPAS revealed the grave dangers of biased data and underscored the pressing need for ethical considerations in the development and deployment of machine learning models in critical decision-making processes. As a result of ProPublica's reporting, COMPAS has been widely contested (Goel et al., 2021; Bao et al., 2021, p.11; Taylor A., 2020), signalling a significant victory in the fight against systemic racism in the criminal justice system. This serves as a cautionary tale that, to foster fair and equitable outcomes, it is imperative to scrutinize and rectify the biases that may be embedded in the data used to train machine learning models.

5. Algorithmic biases and debiasing

The COMPAS algorithm's issues of bias and fairness in algorithmic decision-making have been highlighted by numerous studies. In particular, Dressel and Farid (2018) revealed that African American defendants were twice as likely to be labelled as high-risk compared to white defendants with similar backgrounds. Meanwhile, Chouldechova's (2017) study showed that the use of COMPAS led to higher rates of false positives for African-American defendants, resulting in longer sentences and increased incarceration rates. Additionally, research by Angwin et al. (2016) and Larson et al. (2016) demonstrated that the use of risk assessment tools like COMPAS can lead to racial disparities in sentencing, with African-American defendants receiving harsher sentences than their white counterparts for similar offenses. ProPublica's ground-breaking investigation into the COMPAS algorithm adds to this body of work, underscoring the need for ethical considerations in the development and deployment of machine learning models in critical decision-making processes.

In summary, these studies highlight the pressing need for algorithmic decision-making tools to be developed and assessed with fairness and accountability in mind. ProPublica's investigation into the COMPAS algorithm has prompted a broader dialogue regarding the use of algorithmic decision-making in the criminal justice system, emphasizing the crucial importance of transparency and accountability in this field.

Adversarial debiasing refers to a set of techniques used to mitigate the effects of biases in a dataset during the construction of a fair machine learning model. In-processing adversarial debiasing, a prominent method within this family of techniques, involves training the primary model to predict the target variable, while simultaneously training an adversary to predict the biased attribute of the input data. The goal of this method is to avoid a correlation between the protected attribute and the primary model's decision-making, which can lead to biased outcomes (Yang et al., 2023).

In the process of adversarial debiasing, the primary model is trained to maximize accuracy while confusing the adversary. This approach enables the model to learn how to focus on relevant input features while disregarding biased attributes. In addition, the model is designed not to base decisions on features that overly correlate with the protected attribute, thereby minimizing the potential for perpetuating biased outcomes in the decision-making process. For example, ZIP codes have historically been used as a surrogate for race due to the history of segregation in the United States, and hence, adversarial debiasing applied to a model for removing racial bias can also help mitigate ZIP code bias.

Adversarial debiasing is a flexible and effective tool for addressing bias in datasets, particularly in cases where correlations between the protected attribute and other features in the data could potentially perpetuate biased outcomes. By removing this correlation, adversarial debiasing can produce a model that is free from bias and can be used in critical decision-making processes without fear of perpetuating systemic bias.

Correa et al. (2023) have in particular addressed biases in the context of medical image analysis. They proposed a two-step framework that incorporated adversarial debiasing and partial learning. In the first step, they used a Generative Adversarial Network (GAN) to achieve statistical learning of a latent representation of medical images that is less influenced by biased factors. In the second step, they trained separate models on a debiased subset and a biased subset to identify and correct biases in the latter.

In another influential paper, Zhang et al. (2018) explored the use of adversarial learning to address bias in machine learning models. They proposed an adversarial learning framework that consisted of a classifier and an adversary. The classifier aimed to predict the target variable accurately while the adversary strived to identify the sensitive attributes used by the classifier to make predictions. By simultaneously training the classifier and the adversary, they achieved a model that is both accurate and unbiased. Their work has highlighted the effectiveness of adversarial learning in debiasing machine learning models.

6. The problems with the COMPAS algorithm

The COMPAS algorithm originally processed a dataset containing information on 11,757 criminal defendants, who were assessed in Broward County, Florida, between 2013 and 2014 (Blomberg et al., 2010). The dataset included demographic information, such as age, gender, and race, as well as data on criminal history, socioeconomic status, and neighbourhood characteristics. Additionally, the dataset incorporated the predicted scores for recidivism for each defendant, which was a key variable used in the analysis of the dataset.

Nonetheless, the use of the COMPAS system has been widely criticized for its bias in predictions, particularly in its tendency to misclassify black defendants as higher risk and white defendants as lower risk. This bias is believed to be partially attributed to the use of historical data on criminal offenders that reflect societal biases and discrimination, as well as the inclusion of variables that are correlated with race (Israni, 2017). In general, including demographic information such as race in the dataset is not inherently problematic, but rather the issue arises when such information is used to make biased predictions as in the case of the COMPAS algorithm.

Although it is important to note that the COMPAS algorithm’s specifics are not available to the public, we have attempted to reprocess the COMPAS analysis in a course project (Karthikeyan et al., 2023). After applying a range of data clean-up techniques to ensure the data was consistent and reliable, we were left with approximately 7,000 criminal defendant records. Thus, the dataset included demographic details, such as age, gender, and race, alongside more relevant information such as prior criminal history and the degree of the criminal charge that a defendant was facing. Subsequently, we ran three primary experiments, two with the baseline Random Forest model (Hastie et al., 2009, pp.587-604) and one with an adversarial debiasing model. In particular, we have been also using two machine learning tools:

- LIME (being the acronym for “Local Interpretable Model-agnostic Explanations”), a technique that can trustworthily explain the predictions of any classifier or regressor by approximating it locally with an interpretable model (Ribeiro et al., 2016).
- AIF360 (AI Fairness 360) is a toolkit comprising a set of open-source tools developed by IBM (including fairness metrics for datasets and models together with the corresponding explanations and algorithms). According to Bellamy et al. (2019), “the main objectives of this toolkit are to help facilitate the transition of fairness research algorithms for use in an industrial setting and to provide a common framework for fairness researchers to share and evaluate algorithms.”

In the first experiment (Random Forest model with LIME), the objective was to find the baseline model accuracy with the biased dataset and identify the primary relevant data attributes. For this purpose, a minimalistic 100-estimator Random Forest model was trained to predict real two-year recidivism. Eighty percent of the pruned dataset was used for training, while the remaining twenty percent was used for validation. The result found the model's accuracy at 76.99%, where the five primary features used for prediction were: (i) the date defendants left custody (if they were recently in custody), (ii) the date that they were screened, (iii) how long since they were most recently put into custody, (iv) the date of their most recent arrest and (v) how many days between their screening date and their arrest date.

In the second experiment (adversarial debiasing with LIME), the same selection and division of training and test data were utilized and a 200-hidden-unit network was constructed using the AI Fairness 360 library (AIF360) to predict two-year recidivism while treating both race and sex as protected attributes that should not be recognizable given the other features and the final result. The found model's accuracy was 76.09%, a very slight reduction from the Random Forest implementation. The most relevant primary features for prediction were now: cases (i), (iii), (iv) of the first experiment, (vi) the date defendants most recently left jail, and (vii) the defendants' juvenile felony count.

In the third experiment (Random Forest with fairness metric), the objective was to gain information about the fairness metrics specifically regarding race. Following a similar methodology, the observed demographic parity was found to be approximately 0.149 while equalized odds were approximately 0.106.

Based on the above experiments, we found that both the baseline Random Forest model and the adversarial debiasing model performed similarly in predicting two-year recidivism. However, there was an incredibly minimal accuracy drop and a much higher performance increase than what was expected from the debiased model.

In terms of feature importance, the Random Forest model relied on a peculiar collection of seemingly unrelated dates, while the adversarial debiasing model found that defendants' recent arrest record and their criminal history were crucial for predictions. Interestingly, we found that neither made a significant priority of race. Because the Random Forest regression model was trained to predict real-world recidivism rates and not to recreate the original COMPAS scores, it did not fully replicate the mistakes of its predecessor due to the increased knowledge of the dataset. Despite this, the nonsensical priorities of the Random Forest model suggest that there might remain an issue with the data. The fact that these trends disappear when using adversarial debiasing indicates that they were trends having significant correlations with race, even if that relationship was unclear from a pragmatic view.

Overall, our findings suggest that the adversarial debiasing model can achieve similar accuracy to the baseline Random Forest model while mitigating bias. The use of fairness metrics can also be useful in understanding the impact of sensitive attributes on model predictions. These results have important implications for developing fairer and more accurate machine-learning models for predicting two-year recidivism.

7. Conclusions and further studies

As Artificial Intelligence and Machine Learning are entering human lives, many applications have leveraged the benefits of the technology. However, as mentioned previously, along with the influence of AI and ML becoming stronger, scientists are questioning their reliability and fairness. Especially, an algorithmic decision system like COMPAS that can impact not only the criminals but also other potential victims has come to public attention. In our empirical work (Karthikeyan et al., 2023), we aimed to analyse and suggest improvements to the COMPAS algorithm from different aspects through three experiments.

In the study of our course project (Karthikeyan et al., 2023), we compared the performance of a baseline Random Forest model with an adversarial debiasing model, which was predicting recidivism. The results showed that both models performed similarly in accuracy, with the Random Forest model having a slightly higher accuracy than the adversarial debiasing model. However, the adversarial debiasing model identified more relevant features for prediction, including recent arrest records and criminal history, while the Random Forest model prioritized a bizarre collection of seemingly unrelated dates that bore unexpected correlations with the defendant's race. In addition, the fairness metrics that we have employed showed relatively good values for demographic parity and equalized odds in the Random Forest model, although there is still room for improvement to achieve a fairer system using techniques such as the often-mentioned adversarial debiasing method. These findings have important implications for the development of fairer and more accurate machine-learning models that will predict recidivism. The adversarial debiasing model demonstrated that it is possible to achieve similar accuracy while mitigating bias, and the use of fairness metrics can be useful in understanding the impact of sensitive attributes on model predictions.

Future computational studies that we intend to undertake are going to continue to explore ways to improve the fairness and accuracy of machine learning models in the criminal justice system, paying attention to the human bias that often makes its way into seemingly impartial statistics. One direction that we intend to follow is that of the reduction of bias in the data using methods like Gradient Penalty, Disparate Impact Remover, and SMOTE.

- Gradient Penalty has been used in the context of improving the performance of the Wasserstein Generative Adversarial Networks (WGANs) (Arjovsky et al., 2017), which generate only poor samples or fail to converge. Gulrajani et al. (2017) found that using a certain weight clipping in WGANs to enforce a Lipschitz constraint on the discriminator was the source of many problems leading to undesired behaviour. Thus, by penalizing the norm of the gradient of the discriminator, Gulrajani et al. proposed the Gradient Penalty (WGAN-GP), which does not suffer from the same problems.
- Disparate Impact Remover (Feldman et al., 2015) is a pre-processing technique that by editing feature values increases group fairness while preserving rank-ordering within groups. The motivation behind the work of Feldman et al. was that, in the legal system of the U.S.A., unintentional bias is often encoded via the so-called "disparate impact," which occurs when the outcomes of a selection process for different groups (e.g., over ethnicity, gender, etc.) turn out to be disparate. In their 2015 publication, Feldman et al. have managed to link disparate impact to an underrated measure of classification accuracy and proposed a test for disparate impact based on how well a class can be predicted from existing empirical attributes in data. In this way, they developed a methodology on how data might be made unbiased.
- The Synthetic Minority Over-sampling Technique (SMOTE) is one of the very often used methods to mitigate bias in AI models, which happen to result in unfair decisions. Chawla et al. (2002), who established the SMOTE method, were concerned with the construction of classifiers from imbalanced datasets (a dataset is imbalanced if the classification categories are not approximately equally represented). They found that a combination of their method of over-sampling the minority class and under-sampling the majority class can achieve better classifier performance than only under-sampling the majority class. Recently, the

mathematical justification of SMOTE was presented by Zhou et al. (2023), who showed that synthetic data generated by oversampling underrepresented groups can mitigate algorithmic bias while keeping the predictive errors bounded.

Another option to explore new and more fair ways of algorithmic decision-making would be to use a method called SHAP instead of LIME. The SHapley Additive exPlanations (SHAP) tool is a black-box interpretation approach employed to elucidate machine learning predictions (Lundberg & Lee, 2017). It uses the concept of the "Shapley value" from cooperative game theory to compute explanations of model predictions in three cases: regression values, sampling values, and Quantitative Input Influence (QII) measures (Lundberg & Lee, 2017, p.3).

To conclude, the aim of our contribution was twofold: (i) to provide a survey of algorithmic-decision systems (used mainly in criminal justice, but in other sectors too) and (ii) to discuss the importance and the crucial issues around certain machine learning tools that may contribute to a better understanding of algorithmic bias and effectively address issues of AI fairness promoting equity and diversity in society.

References:

Ananya. (2023). Algorithms are making important decisions. What could possibly go Wrong? Retrieved from https://www.scientificamerican.com/article/algorithms-are-making-important-decisions-what-could-possibly-go-wrong/

Andrews, D. A., & Bonta, J. (1995). The Level of Service Inventory - Revised. Toronto: Multi-Health Systems.

Andrews, D., Bonta, J., & Wormith, J. S. (2004). Level of service/case management inventory. [Data set]. PsycTESTS Dataset. doi:10.1037/t05029-000

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning, PMLR, 70 (pp. 214-223). Retrieved from https://proceedings.mlr.press/v70/arjovsky17a.html

Australian Human Rights Commission. (2020). Using artificial intelligence to make decisions: Addressing the problem of algorithmic bias. Retrieved from https://humanrights.gov.au/sites/default/files/document/publication/ahrc_technical_paper_algorithmic_bias_2020.pdf

Baeza-Yates, R. (2018). Bias on the web. Communications of the ACM, 61(6), 54–6. doi:10.1145/3209581

Balagopalan, A., Madras, D., Yang, D., Hadfield-Menell, D., Hadfield, G. K., & Ghassemi, M. (2023). Judging facts, judging norms: Training machine learning models to judge humans requires a modified approach to labeling data. Science Advances, 9(19). doi:10.1126/sciadv.abq0701

Bao, M., Zhou, A., Zottola, S. A., Brubach, B., Desmarais, S. L., Horowitz, A., Lum, K., & Suresh Venkatasubramanian. (2021). It's COMPASlicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. arXiv:2106.05498. https://doi.org/10.48550/arXiv.2106.05498

Barenstein, M. (2019). ProPublica's COMPAS data revisited. arXiv:1906.04711. https://doi.org/10.48550/arXiv.1906.04711

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. California Law Review. 104(3), 671-732. doi:10.15779/Z38BG31

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2019). Consumer-lending discrimination in the FinTech era. National Bureau of Economic Research. doi:10.3386/w25943

Baumgartner, R., Arora, P., Bath, C., Burljaev, D., Ciereszko, K., Custers, B., Ding, J., Ernst, W., Fosch-Villaronga, E., Galanos, V., Gremsl, T., Hendl, T., Kropp, C., Lenk, C., Martin, P., Mbelu, S., dos Santos Bruss, S. M., Napiwodzka, K., Nowak, E., Roxanne, T., Samerski, S., Schneeberger, D., Tame-Mai, K., Vlantoni, K., Wiggert, K., & Williams, R. (2023). Fair and equitable AI in biomedical research and healthcare: Social science perspectives. Artificial Intelligence in Medicine, 144, 102658. doi:10.1016/j.artmed.2023.102658

Beaulieu, A., & Leonelli, S. (2021). Data and society: A critical introduction. Los Angeles: SAGE Publications.

Bellamy, R. K. E., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., Zhang, Y., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., & Mehta, S. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), (pp. 4:1–4:15). doi:10.1147/jrd.2019.2942287

Bhutta, N., Hizmo, A., & Ringo, D. (2022). How much does racial bias affect mortgage lending? Evidence from human and algorithmic credit decisions. Board of Governors of the Federal Reserve System. Finance and Economics Discussion Series, (2022-067), 1-44. doi:10.17016/FEDS.2022.067

Blomberg, T., Bales, W., Mann, K., Meldrum, R., & Nedelec, J. (2010). Validation of the COMPAS risk assessment classification instrument. Retrieved from https://criminology.fsu.edu/sites/g/files/upcbnu3076/files/2021-03/Validation-of-the-COMPAS-Risk-Assessment-Classification-Instrument.pdf

Chakraborty, J., Majumder, S., & Menzies, T. (2021). Bias in machine learning software: Why? How? What to do? In D. Spinellis (Ed.), Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (pp. 429-440). New York: Association for Computing Machinery. doi:10.1145/3468264.3468537

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. The Journal of Artificial Intelligence Research, 16, 321–357. doi:10.1613/jair.953

Chen, R. J., Wang, J. J., Drew, Chen, T. Y., Lipkova, J., Lu, M. Y., Sahai, S., & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. Nature Biomedical Engineering, 7, 719–742. doi:10.1038/s41551-023-01056-8

Chouldechova A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153-163. doi:10.1089/big.2016.0047

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In KDD ‘17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 797-806). New York: Association for Computing Machinery. doi:10.1145/3097983.3098095

Correa, R., Jeong, J. J., Patel, B., Trivedi, H., Gichoya, J. W., & Banerjee, I. (2023). A robust two-step adversarial debiasing with partial learning - medical image case-studies. In B. J. Park & H. Yoshida (Eds.), Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications, 1246908. doi:10.1117/12.2647285

Counts, L. (2018). Minority homebuyers face widespread statistical lending discrimination. Retrieved from https://newsroom.haas.berkeley.edu/minority-homebuyers-face-widespread-statistical-lending-discrimination-study-finds/

Cowgill, B., & Tucker, C. E. (2019). Economics fairness and algorithmic bias. SSRN Electronic Journal. doi:10.2139/ssrn.3361280

Douglas, K. S., Hart, S. D., Webster, C. D., Belfrage, H., Guy, L. S., & Wilson, C. M. (2014). Historical-Clinical-Risk Management-20, Version 3 (HCR-20V3): Development and overview. The

International Journal of Forensic Mental Health, 13(2), 93–108. doi:10.1080/14999013.2014.906519

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580. doi:10.1126/sciadv.aao5580

Emspak, J. (2016). How a machine learns prejudice. Retrieved from https://www.scientificamerican.com/article/how-a-machine-learns-prejudice/

Engler, A. (2019). For some employment algorithms, disability discrimination by default. Retrieved from https://www.brookings.edu/articles/for-some-employment-algorithms-disability-discrimination-by-default/

Engler, A. (2021). Auditing employment algorithms for discrimination. Retrieved from https://www.brookings.edu/articles/auditing-employment-algorithms-for-discrimination/

Fazel, S., Burghart, M., Fanshawe, T., Gil, S. D., Monahan, J., & Yu, R. (2022). The predictive performance of criminal risk assessment tools used at sentencing: Systematic review of validation studies. Journal of Criminal Justice, 81(101902), 101902. doi:10.1016/j.jcrimjus.2022.101902

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In KDD '15: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 259–268). New York: Association for Computing Machinery. doi:10.1145/2783258.2783311

Garrett, B. L., Jakubow, A., & Monahan, J. (2019). Judicial Reliance on Risk Assessment in Sentencing Drug and Property Offenders: A Test of the Treatment Resource Hypothesis. Criminal Justice and Behavior, 46(6), 799-810. doi:10.1177/0093854819842589

Goel, S., Shroff R., Skeem J., & Slobogin, C. (2021) The accuracy, equity, and jurisprudence of criminal risk assessment. In Roland Vogl (Ed.), Research Handbook on Big Data Law (pp. 9-28). Cheltenham, United Kingdom: Edward Elgar Publishing. doi:10.4337/9781788972826

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). Improved training of Wasserstein GANs. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, & R. Fergus (Eds.), NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 5769–5779). Red Hook, NY: Curran Associates Inc. doi:10.5555/3295222.3295327

Hare, R. D. (1993). Without Conscience: The Disturbing World of Psychopaths Among Us. New York: Pocket Books.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. New York: Springer. doi:10.1007/978-0-387-84858-7

Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2011). Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders: Revised age weights. Sexual Abuse: A Journal of Research and Treatment, 24(1), 64–101. doi:10.1177/1079063211409951

Holm, S. (2023). Statistical evidence and algorithmic decision-making. Synthese, 202(1). doi:10.1007/s11229-023-04246-8

Howard, P. D., (2006). The Offender Assessment System: An evaluation of the second pilot. Retrieved from https://www.ojp.gov/ncjrs/virtual-library/abstracts/offender-assessment-system-evaluation-second-pilot

Huchzermeier, C., Geiger, F., Bruss, E., Godt, N., Köhler, D., Hinrichs, G., & Aldenhoff, J. B. (2007). The relationship between DSM-IV cluster B personality disorders and psychopathy according to Hare's criteria: Clarification and resolution of previous contradictions". Behavioral Sciences & the Law, 25(6), 901–11. doi:10.1002/bsl.722.

Israni, E. T. (2017). When an Algorithm Helps Send You to Prison. Retrieved from https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html

Karthikeyan, R., Moore, G., Yi, C., Naveen, Y., Madhugondu, D., & Chang Y.-Y. (2023). Mitigating bias in judicial ML models to promote fairness in criminal justice system. Unpublished manuscript of CSE 575 course project, Arizona State University.

Kim, P. (2017). Data-driven discrimination at work. William & Mary Law Review, 58(3), 857, https://scholarship.law.wm.edu/wmlr/vol58/iss3/4/

Kizilcec, R. F. & Lee, H. (2022). Algorithmic fairness in education. In W. Holmes & K. Porayska-Pomsta (Eds.), Ethics in Artificial Intelligence in Education (pp. 174-202), New York: Routledge.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. AEA Papers and Proceedings. American Economic Association, 108, 22-27. doi:10.1257/pandp.20181018

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences of the United of America, 117(14), 7684-7689. doi:10.1073/pnas.1915768117

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Management Science, 65(7), 2966–2981. doi:10.1287/mnsc.2018.3093

Langenkamp, M., Costa, A., & Cheung, C. (2019). Hiring fairly in the age of algorithms. SSRN Electronic Journal. doi:10.2139/ssrn.3723046

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Latessa, E. J., Lemke, R., Makarios, M., Smith, P., & Lowenkamp, C. T. (2010). The creation and validation of the Ohio Risk Assessment System (ORAS). Federal Probation, 74(1), 16–22.

Lavanchy, M. (2018). Amazon's sexist hiring algorithm could still be better than a human. Retrieved from https://theconversation.com/amazons-sexist-hiring-algorithm-could-still-be-better-than-a-human-105270

Lowenkamp, C. T., Johnson, J. L., Holsinger, A. M., VanBenschoten, S. W., & Robinson, C. R. (2013). The federal Post Conviction Risk Assessment (PCRA): A construction and validation study. Psychological Services, 10(1), 87–96. https://doi.org/10.1037/a0030343

Lundberg, S., & Lee S. I. (2017). A unified approach to interpreting model predictions. arXiv:1705.07874. doi:10.48550/arXiv.1705.07874

Martinez, E., & Kirchner, L. (2021). The secret bias hidden in mortgage-approval algorithms. Retrieved from https://publicintegrity.org/inequality-poverty-opportunity/bias-mortgage-approval-algorithms/

Martínez, N., Vinas, A., & Matute, H. (2021). Examining potential gender bias in automated-job alerts in the Spanish market. PloS One, 16(12), e0260409. doi:10.1371/journal.pone.0260409

Melton, G. B., Petrila, J., Poythress, N. G., Slobogin, C., Otto, R. K., Mossman, D., & Condie, L. O. (2018). Psychological Evaluations for the Courts: A Handbook for Mental Health Professionals and Lawyers (4th ed.). New York: Guilford Press.

Park, A. (2019). Injustice Ex Machina: Predictive algorithms in criminal sentencing. Retrieved from https://www.uclalawreview.org/injustice-ex-machina-predictive-algorithms-in-criminal-sentencing

Pfeiffer, J., Gutschow, J., Haas, C., Möslein, F., Maspfuhl, O., Borgers, F., & Alpsancar, S. (2023). Algorithmic fairness in AI. Business & Information Systems Engineering. 65(2), 209-222. doi:10.1007/s12599-023-00787-x.

Phenix, A., Fernandez, Y., Harris, A. J. R., Helmus, M., Hanson, R. K., & Thornton, D. (2016). Static-99R Coding Rules Revised – 2016. Retrieved from https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/sttc-2016/sttc-2016-en.pdf

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp.1135-1144). New York: Association for Computing Machinery. doi:10.1145/2939672.2939778

Rieskamp, J., Hofeditz, L., Mirbabaie, M. & Stieglitz, S. (2023). Approaches to improve fairness when deploying AI-based algorithms in hiring - Using a systematic literature review to guide future research. In X. B. Tung (Ed.), Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS). Retrieved from https://scholarspace.manoa.hawaii.edu/items/606e3cb6-9ab8-44a5-9bfc-4430410bc29d

Scarlet, J. (2011). An introduction to the Psychopathy Checklist-Revised (PCL-R). Retrieved from https://psychopathyinfo.wordpress.com/2011/12/31/an-introduction-to-the-psychopathy-checklist-revised-pcl-r/

Latessa, E., Lovins, B. & Makarios, M. (2013). Validation of the Indiana Risk Assessment System Final Report. Retrieved from https://www.in.gov/courts/iocs/files/prob-risk-iras-final.pdf

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the Machine Learning life cycle. Equity and Access in Algorithms, Mechanisms, and Optimization. Presented at the EAAMO '21: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (pp. 1–9). New York: Association for Computing Machinery. doi:10.1145/3465416.3483305

Taylor, A. M. (2020). AI Prediction Tools Claim to Alleviate an Overcrowded American Justice System... But Should they be Used? Retrieved from https://stanfordpolitics.org/2020/09/13/ai-prediction-tools-claim-to-alleviate-an-overcrowded-american-justice-system-but-should-they-be-used/

Vartan, S. (2019). Racial bias found in a major health care risk algorithm. Retrieved from https://www.scientificamerican.com/article/racial-bias-found-in-a-major-health-care-risk-algorithm/

Washington, A. L. (2019). How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. Colorado Technology Law Journal, 17(1), 131. http://ctlj.colorado.edu/wp-content/uploads/2021/02/17.1_4-Washington_3.18.19.pdf

World Bank Group Study. Women, Business and the Law 2020. Washington, DC: World Bank Group. 2020.

Yang, J., Soltan, A. A. S., Eyre, D. W., Yang, Y., & Clifton, D. A. (2023). An adversarial training framework for mitigating algorithmic biases in clinical machine learning. Npj Digital Medicine, 6(1), 55. doi:10.1038/s41746-023-00805-y

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In AIES ‘18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, (pp. 335-340). New York: Association for Computing Machinery. doi:10.1145/3278721.3278779

Zhou, Y., Kantarcioglu, M., & Clifton, C. (2023). On improving fairness of AI models with synthetic minority oversampling techniques. In Proceedings of the 2023 SIAM International Conference on Data Mining (SDM) (pp. 874–882). doi:10.1137/1.9781611977653.ch98