The Use of Fairness and Machine Learning in Judicial Decision-Making Processes: a Systematic Literature Review

Gabriel Santana Furtado Soares
CESAR School
Recife, Brazil
email:gsfs@cesar.school

Rafael Ferreira Mello
CESAR School
Recife, Brazil
email:rflm@cesar.school

***Abstract*—The use of fairness has been pointed out as an alternative to avoid or mitigate biases in decision-making processes that use machine learning. This work aims to identify, evaluate and interpret studies that present details about the method, tools and use of fairness tests. This is a systematic review study by selecting articles on fairness and machine learning in judicial decision-making processes. As a result, it was possible to understand the state of the art of the use of fairness in the context of the justice system, analysing the quality, challenges and difficulties found in the literature.**

***Index Terms*—Fairness, Machine Learning, Court Lawsuits, Bias.**

I. Introduction

In recent years, research and the practical adoption of machine learning techniques, including deep learning, have advanced exponentially in the legal domain [1]. For instance, between 2018 and 2019, the number of works published in the area of Artificial Intelligence (AI) applied to Natural Language Processing (NLP) was almost three times greater than the works published in the same area between 2010 and 2017.

The increase in the hypotheses of using AI applications, adding to the access to massive data and information, has dramatically impacted the decision-making process in the most diverse areas. In the context of the North American criminal justice system, there are examples of use in virtually all procedural steps, including the formulation of sentences and calculation of the risk of recidivism [2].

The application of these techniques in decision-making systems, a fact that has become common in the United States, with models such as PREDPOL and Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), has found acceptance by the justice system due to the increase in efficiency in the production of decisions or at least, as a way of providing assistance to them [2], [3].

Still, in the US context, Kehl and Kesser [3] discuss the substantial increase in the use of predictive tools using AI techniques, especially machine learning, in recent years. In turn, Zavrsnik cites the use of predictive algorithms in about 1.5 million criminal cases in 21 different jurisdictions [4]. Given this context, recent literature has shown the high impact of the use of machine learning and artificial intelligence techniques in the Justice System. However, recent papers also point out critical issues, such as the bias of those decisions [2], [4], [5]. The use of decision-making software can also produce, even if unintentionally, disparities and discrimination [2]. Thus, it becomes essential to debate the consequences of what can happen when cultural codes are embedded in technical software codes.

In this context, this article presents a systematic review of the literature focusing on applying fairness techniques to avoid or mitigate decision bias. In this sense, this review aims to identify, evaluate and interpret studies that present details on the method, tools and use of fairness tests and answer the following research questions:

- **RQ1**: What are the goals of using fairness in decision-making algorithms of the Justice System?
- **RQ2**: Which machine learning techniques have been used to support fairness in decision-making algorithms of the Justice System?
- **RQ3**: What evidence, if any, shows that using fairness improves the quality of decisions made with machine learning techniques?
- **RQ4**: What are the challenges and difficulties in using fairness in algorithmic decision-making applications in the Justice System?

II. Theoretical Background

The concept of fairness is difficult to characterise. Srivastava et al. state that, despite the interest in the topic, within the context of machine learning, there is no consensus on its definition [6]. According to the authors, there is an impossibility of finding a definitive concept. Therefore, they propose that the notion of fairness must be established within the context in which the algorithmic model is used to reflect the impacted population's sense of justice.

Following the same idea, Russel [7] points out several difficulties in obtaining a precise definition of fairness in the most diverse contexts of application of machine learning techniques, whether in the field of the justice system or as a job search system, loans, among others. In addition, Chouldechova and Roth [8] stated that most works on fairness are based on

 ISBN: 978-1-61208-997-3

classifications with little data (one-shot classification), while algorithmic systems work with several layers and components combined, dramatically increasing their complexity. It is also possible to find other distortions in using systems based on machine learning techniques, which will not necessarily be the object of fairness tests, such as the existence of biases embedded in the data collected [8], [9].

The PREDPOL predictive system against crimes involving illicit drugs, in the U.S. Context, particularly in the city of Oakland, noticed that black people had an approach rate almost twice as high as white people, demonstrating the persistence and influence of stereotypes and cultural prejudices in the construction of machine learning models, generating negative feedbacks, even if there is no deliberate will to discriminate [10]. An example of this type of bias can be seen in a paper that analysed the COMPAS software in the USA, which concluded that black defendants had a 77% chance of being classified as at high risk for committing future crimes compared to white defendants [5].

Furthermore, in the current context of increased use of predictive models, it is necessary to apply fairness tools during their design and construction since such systems can be used in sensitive areas, making impactful decisions in people's lives, to ensure that decisions do not reflect discriminatory behaviour towards certain groups or populations [11]. Therefore, Propositional works, such as the Python-based toolkit and AI Fairness 360 (AIF360), reinforce the need for more research on the topic [12].

III. METHOD

We started the research with the methodology proposed by Kitchenham et al. for systematic literature reviews [13]. We define our objectives in the survey and, consequently, define them as prime questions. Afterwards, the search string and exclusion and inclusion criteria were defined.

We defined the search strategy based on the research questions and objectives presented in the introduction. We defined the following string – "fairness" and "machine learning" and "decision", to be searched in the abstract of each article and then applied the string to four databases: IEEE, ACM, Springerlink, and Scopus, from 2018 to 2022.

To define the string, we use keywords from the research questions. In turn, the databases were chosen for the relevance of publications and because they are multidisciplinary. In the end, we selected the following keywords: (Fairness or ethics) + (Machine Learning or Natural Language Processing or deep learning or clustering).

After running the string, a set of 518 articles was found, at which point, using the Mendeley tool [14], we removed duplicate articles and applied the following inclusion/exclusion criteria:

- Full and short papers;
- Legal decision-making context;
- Published in English;
- The grey literature was excluded.

Finally, we applied the quality assessment, with 6 questions defined based on the objectives and research questions: (i) Are the results well presented and meaningful? (ii) Are the work objectives clearly defined? (iii) Was the research design adequate to achieve the objectives? (iv) Is there an analysis of data obtained from real problems? (v) Were the results of the study adequately validated? (vi) Does the work address a real problem?

After reading the entire paper, the authors assigned a grade of 0 or 1 to each item of the quality assessment. In the end, the articles that obtained 50% or more of the possible points were kept. Table I presents the quality criteria for each paper. After removing duplicates and applying the inclusion and exclusion criteria, 11 articles remained, as shown in column round 1 of table II.

TABLE I
QUALITY ASSESSMENT

Paper	Q1	Q2	Q3	Q4	Q5	Q6	total
[15]	1	1	1	0	0	1	4
[16]	1	1	1	0	0	0	3
[17]	1	1	1	1	0	1	5
[18]	1	1	0	0	0	1	3
[19]	0	1	0	0	0	0	1
[20]	1	1	0	0	1	1	4
[21]	0	1	0	1	0	1	3
[22]	1	1	1	0	0	0	3
[23]	1	1	1	1	1	1	6
[24]	1	1	1	1	1	1	6
[25]	1	1	1	1	0	1	5

Finally, Table II shows the final selection of the papers after applying the methodology proposed by Kitchham et al. [13]. In the end, the final number of papers analysed in this literature review was 10.

TABLE II
NUMBER OF PAPERS SELECTED

Database	Direct search	Round 1 (title and abstract)	Round 2 (quality assessment)
ACM	150	3	2
Scopus	264	6	6
IEEE	60	0	0
Springerlink	44	2	2
Total	518	11	10

IV. RESULTS

A. RQ1: What are the goals of using fairness in decision-making algorithms of the Justice System?

The articles highlighted in this review raise several concerns about the use of data-oriented technologies in the justice system, even questioning the current state-of-the-art since the lack of ethical and legal values, such as accountability and justice, adequately embedded in the design of tools, can prevent their use [23]. Moreover, Goel et al. defines the objectives of using fairness in machine learning applications to develop a system that is both accurate and nondiscriminatory [23].

 ISBN: 978-1-61208-997-3

Rodolfa et al., on the other hand, refer us to the system's concern about the ethical implications of using machine learning applications in the justice system, pondering the results with COMPAS racial submission [24].

B. RQ2: Which machine learning techniques have been used to support fairness in decision-making algorithms of the Justice System?

The search for better decisions in the area of Justice, using machine learning techniques that respect values such as fairness, has been a growing area of studies and applications, whether in an attempt to create adequate frameworks, models to mitigate bias, performance testing and post-processing steps to mitigate disparities [15]. The search for fairness in decision-making algorithms of the Justice System should not necessarily be a reactive task. Instead, it can be ensured by a process in which values such as fairness must be built into the application design, even being a solution quality criterion [16].

Nevertheless, the same author points out that the most common technique is the Demographic Party Technique, followed by the Supervised Learning Technique [16]. He also cites Calibration checks, Predictive parity, Error rate balance and Statistical Parity as less commonly used, and Unsupervised learning and Deep learning techniques are rarely used. Regarding the techniques above, for conceptualization purposes, demographic parity is defined as a technique for obtaining fairness that works with the category of protected attributes, which requires an equal proportion of positive predictions in each group, with a prevalence of at least 80% hit [26]. In turn, in the Supervised learning technique, a labelled training dataset is applied to the decision model to ensure nondiscriminatory (regular or inverse) decisions [15].

Conceptualisation of the calibration checks algorithms is defined as the pursuit of "not being discriminatory", while in the error rate balance technique, both protected and unprotected groups have equal false positive rates [27]. In the predictive value parity technique, it is sought that both positive predictive value-parity (ppv) and negative predictive value - parity (NPV) are satisfied [27]. And finally, statistical parity a classifier satisfies this definition if subjects in both protected and unprotected groups have an equal probability of being assigned to the positive predicted class [27].

However, the decision on which technique to use will depend on the context being modelled, and even the use of the protected attributes technique, if poorly applied, can lead to inverse discrimination [16]. In the same direction, Green and Chen [20] suggest that the technique or mechanism to be used with greater ethical efficiency will depend on the situation under analysis, with the need for further studies to improve the decision of the technique to be used.

There are reports about using the linear regression technique, especially in the COMPAS system and in risk analysis tools [17], [22]. Logic regression is also cited in one case [25]. In turn, Goel et al. propose using a weighted sum of logs to reduce the possibility of discriminatory decisions [23].

Finally, Chiao reports difficulty in adopting algorithmic decision-making in high-stakes areas like criminal justice due to the need for better knowledge of the area and the consequences of these decisions [22].

C. RQ3: What evidence, if any, shows that using fairness improves the quality of decisions made with machine learning techniques?

Two articles pointed to direct evidence on the quality of decisions and results obtained by applications. Goel et al. use the weighted sum of logs technique in the COMPAS system to achieve a nondiscriminatory result better than the one obtained by the original algorithm without significant loss of accuracy [23].

Rodolfa et al., although still in an initial form, manage to develop a predictive model with greater accuracy for social and nonincarcerating interventions in potentially relapsing individuals [24].

Although the review did not find many references about practical tests that would imply the improvement of decisions taken by algorithmic systems, the authors [24] express their concern with the quality issue, placing it as a key factor for using these systems. One author suggests that the greater the impact of a decision, the less acceptable the use of machine learning or other equivalent techniques to carry out the decision-making [21].

D. RQ4: What are the challenges and difficulties in using fairness in algorithmic decision-making applications in the Justice System?

To answer this research question, it is first important to consider that studies in AI ethics are still in development and conceptualisation, with many challenges to overcome, especially regarding accountability, transparency and fairness [23].

As stated in the answer to questions 1 and 2, the solution to this debate does not necessarily lie in a single technique or development phase. It is necessary to analyse the context of the problem to be solved, as well as the possibility of thinking about values such as fairness both at the time of application design and in its production, quality control and performance analysis [15], [23].

In turn, Chiao points out two major challenges to overcome in using technology for the justice system, especially the criminal justice system. The first refers to the belief that algorithms are purely objective when they contain, encoded, a series of preferences and discrimination by those who designed them. A second challenge for the author also resides in the impossibility of algorithms to simulate the full range of factors humans consider when making a decision [22].

In the context of justice systems, Rodolfa et al. show particular concern since the data samples can be historically biased [24]. Furthermore, Chiao criticises the situation of the criminal justice system as a whole, questioning the feasibility of automated decision-making processes when the context to be modelled is still precarious [22].

 ISBN: 978-1-61208-997-3

V. Conclusion

There is a consensus in the literature addressed by this systematic review that fairness is crucial for developing systems that involve decision-making in the justice system. Other values mentioned, such as accuracy, transparency, and accountability, that were not the object of this research due to the reduced scope, will be the subject of future studies. The theme of fairness must be addressed throughout the application development process, from conception and design to the testing and validation stages [16].

The Alan Turing Institute[1] also recommends this method through its guide for the responsible design and implementation of AI systems in the public sector, in which it is recommended to observe fairness in all stages of the development of applications that use artificial intelligence.

The use of tools or the definition of specific algorithms for legal decisions will be the object of further study. Another point to be the subject of future research is the normative regulation on the subject under analysis. Regulatory bodies, limits of use, and minimum requirements to avoid bias, among others, can significantly help design and develop systems and applications that respect human rights.

References

[1] N. Bansal, A. Sharma, and R. Singh, "A review on the application of deep learning in legal domain," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 2019, pp. 374–381.

[2] S. Galhotra, Y. Brun, and A. Meliou, "Fairness testing: testing software for discrimination," in *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*, 2017, pp. 498–510.

[3] D. L. Kehl and S. A. Kessler, "Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing," 2017.

[4] A. Završnik, "Algorithmic justice: Algorithms and big data in criminal justice settings," *European Journal of criminology*, vol. 18, no. 5, pp. 623–642, 2021.

[5] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[6] M. Srivastava, H. Heidari, and A. Krause, "Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2459–2468.

[7] J. Russell, "Machine learning equity and accuracy in an applied justice setting," in *2021 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 2021, pp. 215–221.

[8] A. Chouldechova and A. Roth, "A snapshot of the frontiers of fairness in machine learning," *Communications of the ACM*, vol. 63, no. 5, pp. 82–89, 2020.

[9] S. Barocas, M. Hardt, and A. Narayanan, "Fairness and machine learning: Limitations and opportunities," *Fairmlbook. org*, 2018.

[10] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," in *Ethics of Data and Analytics*. Auerbach Publications, 2016, pp. 254–264.

[11] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović *et al.*, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.

[12] Z. Stapic, E. G. López, A. G. Cabot, L. de Marcos Ortega, and V. Strahonja, "Performing systematic literature review in software engineering," in *Central European Conference on Information and Intelligent Systems*. Faculty of Organization and Informatics Varazdin, 2012, p. 441.

[13] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering–a systematic literature review," *Information and software technology*, vol. 51, no. 1, pp. 7–15, 2009.

[14] J. Reiswig, "Mendeley," *Journal of the Medical Library Association: JMLA*, vol. 98, no. 2, p. 193, 2010.

[15] D. Varona, Y. Lizama-Mue, and J. L. Suárez, "Machine learning's limitations in avoiding automation of bias," *AI & SOCIETY*, vol. 36, no. 1, pp. 197–203, 2021.

[16] F. M. Zennaro, "A left realist critique of the political value of adopting machine learning systems in criminal justice," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2020, pp. 92–107.

[17] G. Mohler and M. D. Porter, "A note on the multiplicative fairness score in the nij recidivism forecasting challenge," *Crime Science*, vol. 10, no. 1, pp. 1–5, 2021.

[18] K. Alikhademi, E. Drobina, D. Prioleau, B. Richardson, D. Purves, and J. E. Gilbert, "A review of predictive policing from the perspective of fairness," *Artificial Intelligence and Law*, pp. 1–17, 2021.

[19] G. Harrison, J. Hanson, C. Jacinto, J. Ramirez, and B. Ur, "An empirical study on the perceived fairness of realistic, imperfect machine learning models," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 392–402.

[20] B. Green and Y. Chen, "Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 90–99.

[21] S. Lo Piano, "Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward," *Humanities and Social Sciences Communications*, vol. 7, no. 1, pp. 1–7, 2020.

[22] V. Chiao, "Fairness, accountability and transparency: notes on algorithmic decision-making in criminal justice," *International Journal of Law in Context*, vol. 15, no. 2, pp. 126–139, 2019.

[23] N. Goel, M. Yaghini, and B. Faltings, "Non-discriminatory machine learning through convex fairness criteria," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[24] K. T. Rodolfa, E. Salomon, L. Haynes, I. H. Mendieta, J. Larson, and R. Ghani, "Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 142–153.

[25] N. Grgić-Hlača, C. Engel, and K. P. Gummadi, "Human decision making with machine assistance: An experiment on bailing and jailing," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–25, 2019.

[26] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *International Conference on Machine Learning*. PMLR, 2018, pp. 60–69.

[27] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the international workshop on software fairness*, 2018, pp. 1–7.

[1] https://www.turing.ac.uk/