Discrimination, Bias, Fairness, and Trustworthy AI

Subjects: Computer Science, Artificial Intelligence | Computer Science, Software Engineering

Contributor: Daniel Varona Cordero , Juan Luis Suárez

It has been identified that there exists a set of specialized variables, such as security, privacy, responsibility, etc., that are used to operationalize the principles in the Principled AI International Framework. Bias, discrimination, and fairness are mainly approached with an operational interest by the Principled AI International Framework.

discrimination bias fairness trustworthy ADMS principled AI

1. Analysis of the Variable Discrimination

Automated learning aims to mimic some of the natural learning processes existing in nature, the difference being that in automated learning, the learning is mainly based on a set of examples rather than following defined indications and rules that describe a given context. Similar to what happens with humans, ML often produces predictions and recommends decisions that end up being discriminatory to individuals or groups.

Among the available definitions of “Discrimination” in the context of ML and AI systems [1] is Verma and Rubin’s approach describing discrimination as the direct or indirect relation between a protected attribute and the resulting prediction/classification/suggested decision. This is seconded by Mehrabi [2], where direct discrimination is distinguished by the direct relation between protected attributes and the produced prediction/classification/decision with a negative consequence for the object being targeted by the decision. It expands by declaring that indirect discrimination not only relates to an indirect relation between the mentioned taxonomy but is also manifested when the implicit effects of protected attributes are considered. For instance, the use of an individual postal code in loan and insurance premium calculations are two examples showing how apparently less sensitive individual features may lead to a discriminatory decision.

It can be said that discrimination, in the context of ML and AI systems, has a statistical root when the information learned, by means of pattern discoveries, frequency measure, correlations among attributes, etc., about a group is used to judge an individual with similar characteristics. Hence, the importance of data and data collection procedures is carried out according to the scope of the intended decision or prediction.

The continued use of statistical methods in decision-making and/or the arrival of predictions leads to systematization of discrimination. Therefore, it can be understood that ML has scaled the impact of discrimination and “unintentionally institutionalized” these discriminatory methods through AI, and it has created a perpetual cycle where the object of discrimination itself becomes part of the knowledge base used in subsequent estimates, that,

hence, become equally discriminatory. That is, a recommending software used within an enterprise with a given gender distribution will tend to reproduce the same unbalanced current gender distribution in their selection process while hiring new candidates. The referred distribution might not only be fit in correspondence to the enterprise's training base but also in correspondence with available knowledge about the top performers' distribution in the guild; the particular enterprise is part of what will result in perpetuating the gender distribution in the workforce and conditioning future hiring if the same method is used over time. This is the reason why discriminatory decisions are nowadays generally attributed to prediction, selection/estimation algorithms, etc. [3][4], and not to other equally important aspects such as data gathering, data cleaning, and data processing.

As can be appreciated, discriminating upon the characteristics of an object is not intrinsic to humans. Technology reproduces and amplifies such behavior. The specialized literature exhibits a tendency to hold machine learning algorithms accountable for the problem created by their inability to adequately deal with bias, as analyzed in [5]; however, the data used in training and the data collection methods are equally responsible for discriminatory predictions and recommendations.

Lastly, it can be highlighted that discrimination has both an origin and cause of bias once the outcomes of today's discriminatory decisions based on yesterday's biases populate tomorrow's datasets. A visual aid can be found in **Figure 1** below. In the field of the software industry, both variables, discrimination and bias, are closely related because of the speed at which the whole cycle occurs and because of the cycle's many iterations. The following section presents bias as a variable of analysis.

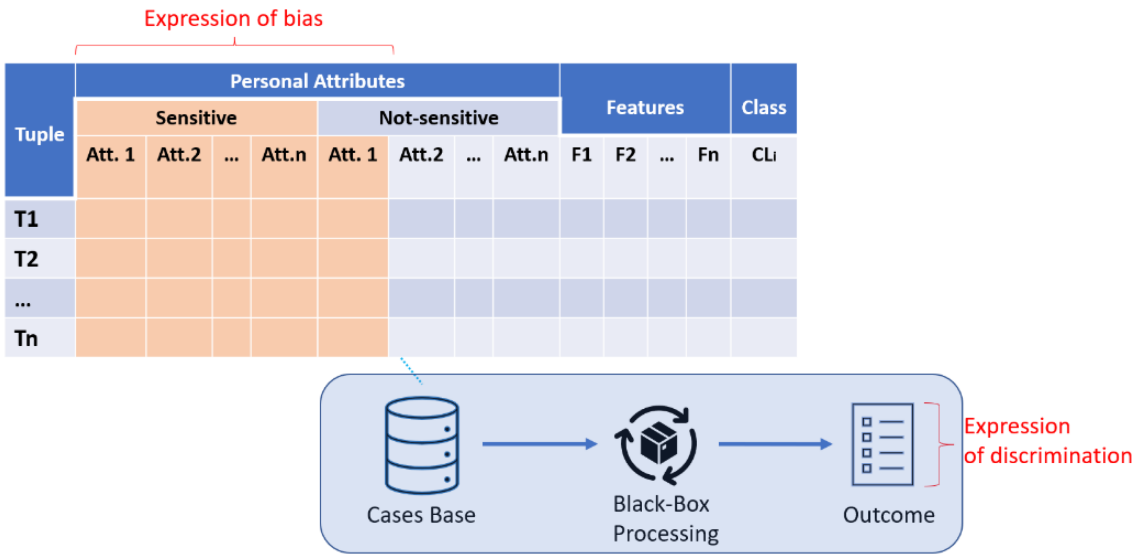

Figure 1. Simplified representation of an automated decision-making process.

2. Analysis of the Variable Bias

Similar to what occurs with human prejudice, the bias in ML leads to discriminatory predictions and recommendations. Consequently, many researchers are pursuing optimization of the methods in which ML

identifies and eliminates bias. There are two marked methodological trends in that regard. The first trend pertains to algorithm calibration [6][7][8][9][10][11][12][13], while the most recent trends [14][15][16][17][18][19] aim to tackle the problem from the early stages of AI algorithms/model design.

Among the documents forming the Principled AI International Framework [20], the UNI Global Union 2017 report [21] describes bias as the action of using features such as gender, race, sexual orientation, and others as discriminatory elements in a decision with a negative impact somehow harmful to the human being. Then, the difference between bias concerning “Discrimination” is that “bias” represents the action while discrimination manifests itself in the result of using certain attributes in the decision-making process, as exhibited in **Figure 1**.

Figure 1 show how bias can be expressed in the inclusion of a subset of attributes oriented to the subject identification from the set of attributes describing a particular individual. Those attributes marked as an expression of bias in **Figure 1** can be both sensitive attributes, also referred to as protected (by researchers promoting the exclusion of such attributes from the decision), and insensitive attributes. The consideration of those attributes in the decision can result in a discriminatory outcome, as previously stated and also represented in the figure.

The dependence among these two variables could be located in this relation. It is also important to note that such a definition emphasizes the negative impact of the decision so that it seems not to consider “bias” when such an effect might be positive.

In that respect, the obligation of fairness defined by Access Now Organization [22] and The Public Voice Coalition [23] first suggests the existence of two benchmarks for the definition of bias in AI. The statistical reference is expressed as the deviation of the prediction in contrast with the event’s actual occurrence, and the social reference is from the evidence of statistical bias within the data representing a social bias. Second, it recognizes that decisions/predictions reflecting bias and discrimination should not be normatively unfair. This means that decisions which are unfair and reflect biases must not only be assessed quantitatively but also evaluated with regard to their context with a case-by-case approach. This is to understand how to avoid them and create a norm/standard rather than being the exception to the rule. Additionally, third, it clarifies that the single evaluation of the outcomes (previously mentioned algorithm calibration) is not enough to determine the fairness of the algorithm or model. This idea was first explored in [5]. Consequently, Access Now Organization [22] and The Public Voice Coalition [23] propose the evaluation of pre-existing conditions in the data that can be further amplified by the AI system before its design is even considered. This report shows an inclination towards the emerging trend of recognizing in the data an origin for discriminatory and biased decisions, in contrast with the rooted trend of solely holding the algorithms accountable for the negative outcomes produced by AIS.

Additionally, the House of Lords Select Committee on Artificial Intelligence [24] and Martinho-Truswell et al. [25] criticize the methods of learning developed in machine learning, specifically how data is used during training. Per the House of Lords Select Committee on Artificial Intelligence [24], while learning, systems are designed to spot patterns, and if the training data is unrepresentative, then the resulting identified patterns will reflect those same patterns of prejudice and, consequently, they will produce unrepresentative or discriminatory decisions/predictions

as well. Martinho-Truswell et al. [25] highlight that good-quality data is essential for the widespread implementation of AI technologies; however, the study argues that if the data is nonrepresentative, poorly structured, or incomplete, then there exists the potential for the AI to make the wrong decisions. Both reports define bias over the basis of misleading decisions produced from such compromised datasets.

Acknowledging the role of data in the introduction of bias is a relatively new approach (This is different from the Garbage In Garbage Out (GIGO) approach to explain the relation of trashy data input with faulty outputs. The GIGO approach links specific data issues such as duplicity of information, absence of information, and noise in information, just to provide a few examples, and bad programming with faulty output from systems. The relatively new approach of pointing out the datasets as an origin for discriminatory decisions refers to those datasets that, even when not being trashy, are biased and triggers discriminatory patterns in ADM systems. It is a new approach as the origin of discriminatory ADM systems' outcomes were mainly linked to biased algorithms, ignoring that datasets and the development team had a role in introducing bias into the system.). Mehrabi's [2] comprehensive survey provides several definitions of types of biases originating in the data. The author enriches upon the already mentioned historical and representation biases by providing further classifications.

IBM [26] adds a human edge to the binomial data-algorithmic bias origin while presenting a set of unconscious bias definitions expressed in terms of their manifestation among the general population that engineers need to be consciously aware of when designing and developing for AI.

3. Analysis of Variable Fairness

By definition, heavy methodologies for software projects help developers and stakeholders to understand that efforts are needed along the software project lifecycle for verification and validation tasks. The automation of bias, the incapacity of AI systems to bring neutrality to the decisions they produce, the perpetuation of bias, and the amplification of the historical discrimination are leading to concerns about how to ensure fairness. On one side, software practitioners strive to prevent intentional discrimination or failure, avoid unintended consequences, and generate the evidence needed to give stakeholders justified confidence that unintended failures are unlikely. On the other side, policymakers work to regulate the design and consumption of such systems so they are not harmful to human beings and that the necessary amendments are made in case they are required.

From a technical point of view, ref. [27] fairness is defined as the actions performed to optimize search engines or ranking services without altering or manipulating them for purposes unrelated to the users' interest. Expanding on that idea, in [21], it is acknowledged that fairness tasks should be planned during the design and maintenance phase of software development and that those tasks should seek to control negative or harmful human bias so that they are not propagated by the system.

Some studies [28][29] relate fairness to inclusion. For instance, ref. [28] stresses that fairness is expressed by means of inclusion and diversity by ensuring equal access through inclusive design and equal treatment. In [29], it is stated that AI systems should make the same recommendations for everyone with similar characteristics or qualifications.

In consequence, software developers and software operators should be required to test the deployed solutions in the workplace on a regular basis to ensure that the system is built for its purpose and it is not harmfully influenced by bias of any kind—gender, race, sexual orientation, age, religion, income, family status, and so on—exposing the variable character of fairness over time. The report also states that AI solutions should adopt inclusive design efforts to anticipate any potential deployment issues that could unintentionally exclude people. Both studies believe necessary the involvement of all affected stakeholders along the project lifecycle. This is a work philosophy that is shared by companies such as Telefónica [30], based in Spain, and one of the main telecommunication operators in Europe. Several of the techniques and metrics available describing how ML pursues fairness are mathematically formalized in the literature [1][2]. A critical analysis of metrics and techniques such as those formalized in both studies were criticized in [5].

A cultural attachment is also presented in [2] while defining the fairness variable when the authors state that different preferences and outlooks within different cultures condition the current situation of having multiple concepts for the term. The situation is aggravated by the fact that available definitions of fairness in philosophy, psychology, and computer science supporting algorithmic constraints are mostly based on Western culture. This led the authors to define fairness as the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making.

An even broader definition is being proposed by the Vatican [31] while using impartiality to explain fairness. The Vatican's working concept gathers the development and consumption of AI systems when it says, "*do not create or act according to bias*", and it connects the outcome of working to ensure fairness with its human focus when it says, "*safeguarding fairness and human dignity*".

The analysis evidences a steering of the majority of the elements describing machine learning's traditional approach [6][9][12] to cope with bias and discrimination, moving away from its reactive character towards a more proactive style. Hence, it is appropriate to state that, in order to produce less discriminatory outcomes, in the context of AIS, the engineering focus needs to commute from fairness (as a nonfunctional requirement) to trustworthy AI as a business model.

4. Analysis of the Variable Trustworthiness

Several studies [32][33][34][35][36] agree that it requires human agency, oversight, and the use of a set of overlapping properties to define trustworthiness in the context of AI systems development and consumption. Among the most frequent highlighted properties across the studied bibliography, the following can be found:

- Reliability is when the system does the right thing it was designed to and is available when it needs to be accessed.
- Reproducibility is when the systems produce the same results in similar contexts.

- Safety is when the system induces no harm to people as a result of their outcomes.
- Security is when the systems are invulnerable or resilient to attacks.
- Privacy is when the system protects a person's identity and the integrity of data, indicates access permission and methods, data retention periods, and how data will be destroyed at the end of such period, which ensures a person's right to be forgotten.
- Accuracy is when the system performs as expected despite new unseen data compared to data on which it was trained and tested.
- Robustness is when the system is sensitive to the outcome and to a change in the input.
- Fairness is when the system's outcomes are unbiased.
- Accountability is when there are well-defined responsibilities for the system's outcome such as the methods for auditing such outcomes.
- Transparency is when it is clear to an external observer how the system's outcome was produced, and the decisions/predictions/classifications are traceable to the properties involved.
- Explainability is when the decisions/predictions/classifications produced by the system can be justified with an explanation that is easy to be understood by humans while being also meaningful to the end-user.
- Other variables such as data governance, diversity, societal and environmental well-being/friendliness, sustainability, social impact, and democracy.

Altogether, as supported by Brundage et al. [37], it can help build a trustworthy methodology to ensure users are able to verify the claims made about the level of privacy protection guaranteed by AI systems, regulators are able to trace the steps leading to a decision/prediction/classification and evaluate them against the context described by the modeled business, academics are able to research the impacts associated with large-scale AI systems, and developers are able to verify best practices are set for each of the AI development stage within the project lifecycle.

In order to achieve Trustworthy AI, the Independent High-Level Expert Group on AI [28] recommends enabling inclusion and diversity throughout the entire AI system's development project's life cycle involving all affected stakeholders throughout the process. Along with Abolfazlian [35], both studies describe three components trustworthy AI should comply with throughout the system's entire life cycle: it should be lawful, complying with all applicable laws and regulations; it should be ethical, ensuring adherence to ethical principles and values; and it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm. Similarly, Gagnon [32] proposes three other main components trustworthy AI systems should consist of the following:

- Ethics of algorithms (respect for human autonomy, prevention of harm, fairness, explicability);
- Ethics of data (human-centered, individual data control, transparency, accountability, equality), and;
- Ethics of Practice (responsibility, liability, codes, regulations).

This actually represents an attempt to harness unintended discrimination produced by AIS, from the perspective of the policymaking and legal norms, specifically with a basis on the International Law of Human Rights. Given that engineering methods alone could not be sufficient enough to protect, according to Fjeld et al. [20], the fundamental rights from unintended harms of AI systems. As seen above, the Principled AI International Framework presented by Fjeld et al. [20] gathers a global effort to establish a set of policies and guidelines informed by principles as a methodological reference when designing AI. Despite the progress that this mechanism might represent from the legal point of view, it is yet insufficient as a methodological mechanism manageable by AI designers given their background and the language [17][18] discrepancies among legal jargon and the software profession, better detailed in [38][39].

References

1. Verma, S.; Rubin, J. Fairness Definitions Explained. In Proceedings of the International Workshop on Software Fairness (FairWare 2018), Gothenburg, Sweden, 29 May 2018.
2. Mehrabi, N.; Morstatter, F.; Saxena, N.A.; Lerman, K.; Galstyan, A.G. A survey on bias and fairness in machine learning. Mach. Learn. 2019, 54, 1–35.
3. Jago, A.S.; Laurin, K. Assumptions about algorithms' capacity for discrimination. Personal. Soc. Psychol. Bull. 2021, 1–14.
4. Loi, M.; Christen, M. Choosing how to discriminate: Navigating ethical trade-offs in fair algorithmic design for the insurance sector. Philos. Technol. 2021, 34, 967–992.
5. Varona, D.; Lizama-Mue, Y.; Suarez, J.L. Machine learning's limitations in avoiding automation of bias. AI Soc. 2020, 36, 197–203.
6. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data 2017, 5, 153–163.
7. Feldman, M.; Friedler, S.A.; Moeller, J.; Scheidegger, C.E.; Venkatasubramanian, S. Certifying and Removing Disparate Impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sidney, Australia, 10–13 August 2015.
8. Fish, B.; Kun, J.; Lelkes, Á.D. A Confidence-Based Approach for Balancing Fairness and Accuracy. In Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, FL, USA, 5–7 May 2016.

9. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. Adv. Neural Inf. Process. Syst. 2016, 29, 3315–3323.
10. Reich, C.L.; Vijaykumar, S. A Possibility in Algorithmic Fairness: Calibrated Scores for Fair Classifications. In Proceedings of the 2nd Symposium on Foundations of Responsible Computing (FORC 2021), Virtual Event, 9–11 June 2021. arXivLabs:2002.07676.
11. Pedreschi, D.; Ruggieri, S.; Franco, T. Discrimination-Aware Data Mining Technical Report: TR-07-19; Dipartimento di Informatica, Universitμa di Pisa: Pisa, Italy, 2007.
12. Solon, B.; Selbst, A.D. Big data's disparate impact. Calif. L. Rev. 2016, 104, 671–732.
13. Zafar, M.B.; Valera, I.; Rodriguez, M.G.; Gummadi, K.P. Fairness Constraints: Mechanisms for Fair Classification. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), San Diego, CA, USA, 9–12 May 2015. arXiv:1507.05259.
14. Holstein, K.; Vaughan, J.W.; Daumé, H.; Dudík, M.; Wallach, H.M. Improving Fairness in Machine Learning Systems: What do Industry Practitioners Need? In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019.
15. Varona, D. La responsabilidad ética del diseñador de sistemas en inteligencia artificial. Rev. Occidente 2018, 446–447, 104–114.
16. Varona, D. AI systems are not racists just because. In Proceedings of the T-13 hours: Building Community Online in CSDH/SCHN2020, Virtual Event, 1–5 June 2020.
17. Varona, D. (Western University, Canada). Artificial Intelligence Design Guiding Principles: Review of "European Ethical Charter on the Use of AI in Judicial Systems and Their Environment". 2020. Available online: https://www.danielvarona.ca/2020/06/17/artificial-intelligence-design-guiding-principles-review-of-european-ethical-charter-on-the-use-of-ai-in-judicial-systems-and-their-environment/ (accessed on 1 July 2021).
18. Varona, D. (Western University, Canada). Artificial Intelligence Design Guiding Principles: Review of "Recommendation of the Council on Artificial Intelligence". 2020. Available online: https://www.danielvarona.ca/2020/06/28/artificial-intelligence-design-guiding-principles-review-of-recommendation-of-the-council-on-artificial-intelligence/ (accessed on 1 July 2021).
19. Veale, M.; Van Kleek, M.; Binns, R. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–27 April 2018.
20. Fjeld, J.; Achten, N.; Hilligoss, H.; Nagy, A.; Srikumar, M. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI; Berkman Klein Center for Internet & Society: Cambridge, MA, USA, 2020.

21. UNI Global Union. The Future World of Work. Top 10 Principles for Ethical Artificial Intelligence; UNI Global Union: Nyon, Switzerland, 2017.
22. Access Now Organization. Human Rights in the Age of AI. Technical Report. AccessNowOrg. November 2018. Available online: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf (accessed on 19 May 2021).
23. The Public Voice Coalition. Universal Guidelines for AI; The Public Voice Coalition: Geneva, Switzerland, 2018.
24. House of Lords Select Committee on Artificial Intelligence. AI in the UK: Ready, Willing and Able? Technical Report. Authority of the House of Lords. 2018. Available online: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf (accessed on 19 May 2021).
25. Martinho-Truswell, E.; Miller, H.; Nti Asare, I.; Petheram, A.; Stirling, R.; Gomez Mont, C.; Martinez, C. Towards an AI strategy in Mexico: Harnessing the AI revolution. White Pap. 2018, 23.
26. IBM. Everyday Ethics for AI; IBM: Armonk, NY, USA, 2019.
27. Demiaux, V.; Si, A.Y. How Can Humans Keep the Upper Hand? The Ethical Matters Raised by Algorithms and Artificial Intelligence; The French Data Protection Authority (CNIL): Paris, France, 2017.
28. Independent High Level Expert Group on AI. AI Ethics Guidelines for Trustworthy AI; European Commission: Brussels, Belgium, 2019.
29. Task Force 7. The Future of Work and Education for the Digital Age. Technical Report. T20. 2020. Available online: https://t20japan.org/task-forces/the-future-of-work-and-education-for-the-digital-age/ (accessed on 19 May 2021).
30. Telefónica. AI Principles of Telefónica; Telefónica: Madrid, Spain, 2018.
31. Servizio Internet Vaticano. Rome Call for AI Ethics; Servizio Internet Vaticano: Vatican City, Vatican, 2020.
32. Gagnon, G.P.; Henri, V.; Fasken; Gupta, A. Trust me!: How to use trust-by-design to build resilient tech in times of crisis. WJCOMPI 2020, 38, 1–6.
33. Wickramasinghe, C.S.; Marino, D.L.; Grandio, J.; Manic, M. Trustworthy AI Development Guidelines for Human System Interaction. In Proceedings of the 13th International Conference on Human System Interaction, Tokyo, Japan, 6–8 June 2020.
34. Smith, C.J. Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development. In Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.

35. Abolfazlian, K. Trustworthy AI needs unbiased dictators! Artif. Intell. Appl. Innov. 2020, 584, 15–23.
36. Wing, J.M. Trustworthy AI. arXiv 2020, arXiv:2002.06276.
37. Brundage, M.; Avin, S.; Wang, J.; Belfield, H.; Krueger, G.; Hadfield, G.K.; Khlaaf, H.; Yang, J.; Toner, H.; Fong, R.; et al. Toward trustworthy AI development: Mechanisms for supporting verifiable claims. arXiv 2020, arXiv:2004.07213.
38. Varona, D.; Suarez, J.L. Analysis of the principled-AI framework's constraints in becoming a methodological reference for trustworthy-AI design. In Handbook of Computational Social Science; Engel, U., Quan-Haase, A., Xun Liu, S., Lyberg, L.E., Eds.; Routledge Taylor and Francis Group: Oxfordshire, UK, 2022; Volume 1, ISBN 9780367456528.
39. Varona, D.; Suarez, J.L. Principled AI Engineering Challenges Towards Trust-worthy AI. Ethics Inf. Technol. 2022, submitted.