THE PHILOSOPHICAL FOUNDATION OF FAIRNESS IN MACHINE LEARNING: AN AFRICAN PERSPECTIVE

Cyriacus O. Emedolu (Rev. Fr) | Madonna University, Nigeria Okija Campus | frcyriacus@madonnauniversity.edu.ng

Remi Chukwudi Okeke, Ph.D | Department of Public Administration Madonna University, Nigeria Okija Campus | okekerc@madonnauniversity.edu.ng

CONCEPTUALIZING FAIRNESS IN MACHINE LEARNING

Fairness in machine learning (ML) refers to the principle that an algorithm should make decisions in a way that is just and unbiased, avoiding discrimination against individuals or groups based on characteristics such as race, gender, age, or other sensitive attributes (Barocas et al., 2023; Pessach & Shmueli, 2022; Mehrabi et al., 2021). Fairness is therefore crucial in ensuring ethical and equitable use of ML systems, particularly in applications like hiring, lending, healthcare, and criminal justice. Key Concepts in Fairness in Machine Learning include the following:

- Bias: Bias in ML can occur at different stages, from data collection to model deployment. On one hand, data bias is about when training data reflects historical inequalities or societal prejudices. On the other hand, algorithmic bias refers to when the design of the algorithm inadvertently favors one group over another.
- Protected Attributes: Characteristics such as race, gender, age, or disability that should not influence the outcomes of an ML model.
- Discrimination: Unfair treatment of individuals or groups in the decisions made by ML systems. This can be direct (explicit bias) or indirect (disparate impact).
- Fairness Metrics: Several metrics are used to measure fairness in ML models, including:
- Demographic Parity: Ensures equal outcomes across groups, regardless of sensitive attributes.
- Equalized Odds: Ensures equal error rates (false positive and false negative) across groups.
- Fairness through Awareness: Considers contextual knowledge about fairness in the specific application.

- Calibration: Ensures the likelihood of an outcome matches the predicted probability across groups.
- Trade-offs: Achieving fairness often involves trade-offs with other objectives, such as accuracy or utility. For example, optimizing for equalized odds might slightly reduce overall accuracy.
- Mitigation Techniques: To address fairness concerns, several techniques can be used:
- Pre-processing: Adjusting the data before training to remove biases.
- In-processing: Modifying the learning algorithm to enforce fairness constraints.
- Post-processing: Adjusting the model's predictions to ensure fairness after training.
- Fairness-Aware Design: Incorporating fairness considerations throughout the ML lifecycle, from data collection and preprocessing to model evaluation and deployment (Barocas et al., 2023; Chouldechova & Roth, 2020; Pessach & Shmueli, 2022; Makhlouf et al., 2021; Mehrabi et al., 2021; Mhasawade et al., 2021).

EXAMPLES OF UNFAIRNESS IN MACHINE LEARNING

Examples of unfairness in machine learning arise in various domains where biased data, algorithmic design, or unintended consequences lead to discriminatory outcomes. For example, an ML-based recruitment tool was found to favor male candidates over female candidates for technical roles. This was because the training data reflected historical hiring practices, where men were more likely to be hired for technical jobs, leading the model to associate higher suitability with male candidates. An ML tool used in the U.S. for risk assessment in criminal justice was found to predict higher recidivism rates for Black defendants than for white defendants. This was because of bias in training data, which may have reflected systemic racial disparities in policing and sentencing (Chen, 2023; Kawamleh, 2024).

Furthermore, facial recognition systems have been shown to perform poorly on individuals with darker skin tones compared to lighter skin tones. This was because; **t**raining datasets were predominantly composed of lighter-skinned individuals, leading to biased performance. In lending and credit scoring, an illustration is where ML algorithm used by financial institutions to assess creditworthiness denied loans to certain minority groups at higher rates. This was essentially because, historical lending data encoded discriminatory practices, leading the model to perpetuate those biases (Leal, 2022; Yucer et al, 2024)

It therefore stands to reason that many instances of unfairness in ML stem from biased or incomplete training data that reflects existing societal inequities. Furthermore, unfairness can also arise from the unintended consequences of optimization objectives,

proxies, or assumptions in model design. Addressing unfairness accordingly requires deliberate interventions, such as diverse datasets, fairness-aware algorithms and continuous monitoring of outcomes.

SITUATING FAIRNESS IN MACHINE LEARNING WITHIN A PHILOSOPHICAL FOUNDATION

Fairness is a fundamental concept in human societies, rooted in principles of justice, equity, and moral reasoning. As machine learning (ML) increasingly shapes decisions in critical domains such as hiring, healthcare, criminal justice and education, the philosophical foundation of fairness becomes a pivotal lens through which to evaluate its impact. Then while philosophy offers diverse frameworks to interpret fairness, in the context of ML, three key perspectives provide a foundation:

- Aristotelian Justice: Aristotle distinguished between distributive justice (allocating resources based on merit or need) and corrective justice (addressing harm caused by wrongdoing) (Maurya, 2021). In ML, distributive justice underpins algorithms that ensure equitable access to opportunities, while corrective justice informs mechanisms to rectify biases in decision-making.
- Rawlsian Theory of Justice: Philosopher John Rawls proposed the "veil of ignorance," advocating for fairness in social systems by designing rules as if unaware of one's position in society (Kalaitzidis, 2024). This perspective inspires ML fairness metrics that prioritize disadvantaged groups and strive for equity in outcomes, irrespective of historical or systemic biases.
- Utilitarianism vs. Deontology: Utilitarianism emphasizes maximizing overall welfare, which may justify trade-offs in fairness to benefit the majority. Deontological ethics, on the other hand, focuses on principles and rights, asserting that fairness should not be compromised, even if it reduces overall efficiency (Nesseler & Adelstein, 2024). These philosophical foundations illuminate the complexity of fairness, as it involves balancing competing values like equity, equality, and utility.

THE ETHICAL DIMENSIONS

Fairness in ML transcends technical considerations, encompassing profound ethical implications. Central in many ethical theories is the principle of non-maleficence (avoiding harm). In ML, fairness mechanisms aim to prevent disparate impacts on marginalized groups. Respecting autonomy is also important. Respecting individual autonomy involves ensuring that ML decisions do not unjustly restrict personal freedom or agency. Fairness in ML therefore requires transparency and accountability, empowering users to understand and challenge decisions. Furthermore, fairness in Ml is deeply tied

to addressing historical inequities and addressing systemic injustices embedded in historical data. Philosophically, this aligns with reparative justice, emphasizing the need to correct the effects of past discrimination. This invariably relates with the need for ensuring inclusivity. Ethical theories highlight the importance of inclusivity, requiring ML systems to reflect diverse perspectives and avoid perpetuating narrow, exclusionary worldviews.

CHALLENGES IN ENSURING FAIRNESS

While fairness is a noble goal, it is fraught with challenges rooted in philosophical and practical dilemmas. Defining fairness is actually a dicey matter. Different philosophical traditions suggest varied definitions of fairness. Should ML systems aim for equal outcomes, equal opportunities, or proportional representation? These interpretations often conflict, necessitating context-specific solutions. Fairness is not static; societal norms and values evolve. ML systems must then adapt to changing expectations while maintaining their ethical grounding. Different stakeholders may therefore have varying definitions of what fairness means in a given context. There is also the issue of conflicting metrics: Achieving one type of fairness may violate another. Training data often reflects societal biases, challenging the ideal of impartiality. Philosophically, this raises questions about whether ML can ever be truly objective. Dynamic environments are accordingly involved. Fairness needs may evolve as societal norms and data distributions change. Limited data is also a challenge. Insufficient or unbalanced data for certain groups can lead to biased models for measuring fairness in Ml (Ferrara et al., 2024; Jui & Rivas, 2024, Zhang, 2024).

Fairness in ML may then require trade-offs with other objectives, such as accuracy, efficiency, or privacy. For instance, prioritizing fairness for one group may inadvertently disadvantage another. In the final analysis, fairness in ML is about creating systems that are ethical, equitable, and inclusive. It requires a combination of technical interventions, thoughtful design, and continuous monitoring, to address biases and ensure just outcomes for all stakeholders.

THE AFRICAN PERSPECTIVE: UNDERSTANDING FAIRNESS IN THE AFRICAN CONTEXT

As machine learning (ML) systems become deeply embedded in decision-making processes across the globe, the need to ensure fairness has become a critical concern. Fairness in ML, a concept grounded in ethical, philosophical, and cultural values, then requires careful contextualization, particularly in African societies where histories of colonialism, systemic inequities, and diverse cultural norms uniquely shape the interpretation and application of fairness. This section of the paper explores the philosophical

foundation of fairness in ML through an African lens, emphasizing indigenous ethical systems, the challenges of bias, and the need for inclusive approaches to technology.

Fairness, as a philosophical and ethical concept, has universal appeal but is also profoundly shaped by cultural and societal values. Hence, in the African context, fairness often aligns with communalism, restorative justice, and the emphasis on the collective well-being. For instance, Ubuntu, a foundational African philosophy, emphasizes interconnectedness, compassion, and the principle that "a person is a person through other persons" (umuntu ngumuntu ngabantu). In the context of ML fairness, Ubuntu suggests that algorithms should prioritize collective welfare, inclusivity and the mitigation of harm to vulnerable communities.

Furthermore, African societies have been shaped by histories of colonial exploitation and systemic injustices. Fairness in ML, therefore, must address these historical inequities by ensuring that algorithms do not perpetuate the marginalization of African voices or exclude the continent from global technological advancements. Then Africa's diversity, spanning languages, traditions, and social structures, requires fairness frameworks that are adaptable and sensitive to local contexts. A one-size-fits-all approach to fairness risks erasing cultural nuances and reinforcing hegemonic biases.

The African philosophical tradition indeed provides rich insights into fairness, which can inform the development and evaluation of ML systems. These underpinnings are anchored on restorative justice, holistic approaches and notions of economic and social equity. The target is the removal of all forms of underrepresentation in data, occasioning economic disparities, language marginalization and colonial legacies. Then in the final analysis, achieving fairness in ML within African contexts requires a blend of philosophical reflection, technical innovation and policy intervention, buildable on the following actionable strategies:

- Indigenous Data Sovereignty: African nations must assert control over their data, ensuring that datasets used to train ML systems are representative of their populations and adhere to ethical standards rooted in local values.
- Culturally-Informed Algorithms: Building of ML models that incorporate cultural knowledge and linguistic diversity; which leverages African epistemologies, to design systems that are contextually relevant and fair.
- Capacity Building: Investing in education and research is essential to empower African scholars, engineers, and policymakers to lead the development of fair ML systems.
- Regulatory Frameworks: African governments and regional bodies, such as the African Union, should establish robust policies to govern the ethical use of AI and ML, incorporating fairness principles that reflect the continent's unique values and challenges.

- Community Engagement: Fairness must be co-created with the communities affected by ML systems. Engaging diverse stakeholders ensures that technology aligns with local needs and aspirations.

CONCLUSION

An African perspective on fairness enriches the global discourse on ML ethics by emphasizing in the first place, the primacy of humanity. The emphasis of Ubuntu on human dignity and collective welfare challenges profit-driven or efficiency-focused ML paradigms, advocating for systems that center on human values. Furthermore, African contributions to ML fairness highlight the importance of addressing global power imbalances in technology development and deployment. Additionally, Africa's cultural and linguistic diversity offers valuable lessons on the importance of inclusivity, adaptability, and pluralism in ML design. The philosophical foundation of fairness in machine learning therefore, viewed through an African perspective, underscores the need for systems that prioritize collective welfare, address historical inequities, and reflect the continent's rich cultural heritage. By drawing on principles such as Ubuntu, restorative justice, and holistic thinking, Africa can contribute uniquely to the global effort to create ethical and equitable ML systems. Fairness in ML, therefore, is not just a technical challenge but a moral imperative that demands context-sensitive solutions and the inclusion of diverse worldviews. As Africa's role in the digital age grows, its philosophical insights will be indispensable in shaping a fairer, more inclusive technological future.

REFERENCES

Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT press.

Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. Humanities and Social Sciences Communications, 10(1), 1-12.

Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. Communications of the ACM, 63(5), 82-89.

Ferrara, C., Sellitto, G., Ferrucci, F., Palomba, F., & De Lucia, A. (2024). Fairness-aware machine learning engineering: how far are we?. Empirical software engineering, 29(9), 1-46.

Jui, T. D., & Rivas, P. (2024). Fairness issues, current approaches, and challenges in machine learning models. International Journal of Machine Learning and Cybernetics, 15(8), 3095-3125.

Kalaitzidis, G. (2024). Review of Amartya Sen's Book "The Idea of Justice". Open Journal of Philosophy, 14(4), 731-735.

Kawamleh, S. (2024). Algorithmic evidence in US criminal sentencing. AI and Ethics, 1-14. https://doi.org/10.1007/s43681-024-00473-y

Leal, A. A. (2022, April). Algorithms, creditworthiness, and lending decisions. In International Conference on Autonomous Systems and the Law (pp. 321-353). Cham: Springer

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2021). On the applicability of machine learning fairness notions. ACM SIGKDD Explorations Newsletter, 23(1), 14-23.

Maurya, S. K. (2021). The concept of justice in reference with philosophies of Plato and Aristotle: A critical study. Journal of Liberty and International Affairs, 7(3), 250-266.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.

Mhasawade, V., Zhao, Y., & Chunara, R. (2021). Machine learning and algorithmic fairness in public and population health. Nature Machine Intelligence, 3(8), 659-666.

Nesseler, A., & Adelstein, M. (2024). Utilitarianism, Deontology, and Ethical Veganism. Journal of Animal Ethics, 14(1), 1-8.

Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. ACM Computing Surveys (CSUR), 55(3), 1-44.

Yucer, S., Tektas, F., Al Moubayed, N., & Breckon, T. (2024). Racial bias within face recognition: A survey. ACM Computing Surveys, 57(4), 1-39.

Zhang, W. (2024). AI fairness in practice: Paradigm, challenges, and prospects. Ai Magazine, 45(3), 386-395.