STM JOURNALS

International Journal of
Information Security Engineering

Volume 1, Issue 2, 2023
July—December
DOI (Journal): 10.37591/IJISE

https://journals.stmjournals.com/ijise

Review IJISE

Unveiling Fairness: A Quest for Ethical Artificial Intelligence and Bias Mitigation

Ushaa Eswaran[1,*], Vivek Eswaran[2], Keerthna Murali[3], Vishal Eswaran[4]

Abstract

Artificial intelligence (AI) systems have become ubiquitous across areas like finance, healthcare, employment, and criminal justice. However, they suffer from issues of unfair bias, lack of transparency, and broad ethical implications impacting vulnerable societal groups disproportionately. This paper reviews key challenges around AI ethics and bias while proposing data-driven guidelines mitigating such algorithmic harms through rigorous statistical testing, predictive modeling ensembles adjusting distortion vectors and AI audits by domain experts analyzing source codes, training data curation and model card documentations ensuring responsible development. A tiered regulatory framework is envisioned spanning self-regulation, external audits, professional codes of ethics, and government oversight balancing innovation impacts with public safeguards.

Keywords: Algorithmic bias, artificial intelligence (AI) ethics, mitigation techniques, model transparency, regulation

INTRODUCTION

Rise of Ubiquitous Artificial Intelligence Systems

The advent of modern artificial intelligence (AI) techniques like machine learning and neural networks has elevated automated decision-making efficiencies across pivotal societal functions like loan approvals, recruitment screening, judicial rulings, and clinical diagnostics assisting human experts with immense data processing abilities [1]. Global investments into AI solutions stand poised to reach $500 billion by 2024 signaling lucrative opportunities that continue attracting private enterprises and public agencies alike keen on augmenting organizational intelligence and competencies at scale using such smart algorithms more pervasively.

***Author for Correspondence**
Ushaa Eswaran
E-mail: drushaaeswaran@gmail.com

[1]Principal and Professor, Department of Electrical Communication Engineering, Indira Institute of Technology and Sciences, Markapur, Andhra Pradesh, India
[2]Senior Software Engineer, Tech Lead, Medallia, Austin, Texas, United States
[3]Site Reliability Engineer II (SRE), Cybersecurity, Dell EMC, CKAD, AWS CSAA, United States
[4]Senior Data Engineer, CVS Health Centre, Dallas, Texas, United States

Received Date: November 23, 2023
Accepted Date: November 29, 2023
Published Date: December 06, 2023

Citation: Ushaa Eswaran, Vivek Eswaran, Keerthna Murali, Vishal Eswaran. Unveiling Fairness: A Quest for Ethical Artificial Intelligence and Bias Mitigation. International Journal of Information Security Engineering. 2023; 1(2): 28–31p.

Emerging Issues with Artificial Intelligence Systems

However, accompanying its spread, disturbing incidents of algorithmic failures reflecting coded societal biases, lack of transparency around inner workings and broad ethical implications have surfaced frequently violating public trust [2]. Studies reveal resume screening tools discounting female applicants disproportionately, facial analysis technologies misclassifying non-Caucasian subjects persistently while sentencing recommendation systems suggesting heightened incarcerations for specific races relatively [3]. Such revelations indicate how historically accumulated human prejudices percolate into AI models through their development environments demanding urgent mitigations protecting fair public participation.

Importance of Artificial Intelligence Ethics and Algorithmic Audits

Hitherto predominantly unregulated AI advancements necessitate principled assessments through an interdisciplinary ethics lens encompassing technological techniques, social theories and policy paradigms collectively ushering an accountable innovation ecosystem guiding sustainable progress [4]. Specifically bias detection using statistical tests and predictive modeling ensembles on fleet wide datasets coupled with explain ability methods and ongoing audits by multidisciplinary teams would be crucial intervention avenues promising transparent and equitable future AI diffusion benefitting communities inclusively.

LITERATURE REVIEW

Theory and Practice of Artificial Intelligence Bias

Substantial research attributes unfair AI algorithm outputs to three primary forms of bias including inherited, technical, and emergent types manifesting from human prejudices percolating into engineering design choices for data and model architectures that get further compounded by feedback loops entrenching historical distortions during continuous redeployments [5]. Studies caution how considerations for benchmark test suits overwhelmingly center around efficiency metrics optimized for homogenous majority demographics overlooking tail minorities through preferential curations skewing encoded worldviews eventually [6].

Mitigating and Governing Artificial Intelligence Harms

To address such challenges, scholars advocate statistical tests assessing model performance variances across relevance subgroups calculating bias indices like demographic parity, equality of odds, and calibration scores that indicate distortions suitable for corrective interventions [7]. Complementing quantitative estimates, qualitative techniques incorporating diverse expert audits analyzing code structures, reviewing input data compositions, studying falsified counterfactuals, and interrogating model card documentations ensure accountable improvisations minimizing algorithmic harms continually [8]. Such ongoing governance further necessitates multi-layered policy formulations addressing self-regulations, professional codes of conduct, external impact audits and government oversight balancing innovation trajectories ethically [9].

Limitations of Current Literature

Despite extensive AI ethics scholarship debating principles, technical toolkits for auditing complex neural networks lag considerably constrained still by computational demands, lack of fleet wide standardized datasets, inadequate multidisciplinary collaborations across social and technical fields and insufficient policy implementations guiding responsible scaling sustainably [10]. Significant research gaps persist around evaluation metrics encompassing often ignored minorities, participative assessment protocols accommodating pluralistic worldviews and regulation regimes incentivizing voluntary adoptions at scale.

RESEARCH QUESTIONS

Considering limitations of current literature, this analysis focuses on the following key questions enriching AI ethics and bias mitigation research:

- *Research Question 1 (RQ1):* What taxonomic frameworks characterize algorithmic harms multidimensionally across purposes, privileges and people for assessment scope consistency?
- *Research Question 2 (RQ2):* How can ensemble quantification combine computational evaluations and participative audits balancing efficiency and generalizability?
- *Research Question 3 (RQ3):* Which policy interventions effectively incentivize self-initiated mitigation practices matching governance capabilities contextually?

METHODOLOGY

Mitigation Taxonomy

To address RQ1 systematically, an exploratory study develops a multilayered mitigation taxonomy categorizing algorithmic harms across functional aspects like data collection purposes, processing

privileges, and impacted demographic groups facilitating targeted redressals using document analysis and expert interviews ($n = 35$) spanning technologists, ethicists, sociologists, lawyers, and policymakers helping characterize key issues multi-dimensionally aiding audit scopes [11].

Bias Quantification

For RQ2, computational techniques and participative audits get combined into evaluation ensembles assessing model disparities effectively using staff-level banking dataset ($n = 20{,}000$) evaluating loan defaults quantitatively using metrics like statistical parity difference, calibration score variance and disaggregated performance metrics between ethnicities followed by code reviews, documentation analysis and counterfactual simulations ($n = 1000$) gauging model behaviors qualitatively minimizing blind spots through expert collaborations.

Policy Formulation

Lastly, for RQ3, large-scale surveys across private enterprises ($n = 500$) and public agencies ($n = 800$) assess motivations and barriers influencing voluntary adoptions of ethical AI practices using regression techniques determining effective incentives like financial schemes, technical tooling and reputational outcomes guiding formal policy formulations federally, professionally, and organizationally.

POTENTIAL CONTRIBUTIONS

Mitigation Taxonomy

The exploratory mitigation taxonomy enriches assessment frameworks characterizing algorithmic issues multi-dimensionally aiding targeted and modular redressals across identified privileges, purposes and people matching context granularities. The dimensions further boost audit participation accommodating diverse pluralistic perspectives systematically.

Ensemble Quantification

Novel ensemble methodologies combining computational efficiency and qualitative sensitivities minimize evaluation blind spots through expert collaboration balancing accuracy metrics and generalizability concerns cohesively minimizing bias. Effective ensembles pave template pathways fostering deeper interdisciplinary cooperation.

Incentivizing Adoptions

Large-scale perception analysis determines motivating factors and prevailing barriers behind voluntary mitigation adoptions tailored for public and private sector contexts distinctively guiding policy formulations incentivizing ethical AI practices effectively through financial schemes, tooling grants and reputation systems lowering compliance overheads.

TIMELINE

The timeline of key research stages is outlined in Table 1.

Table 1. Timeline outlining key research stages.

Months	Tasks
Months 1–2	Develop mitigation taxonomy through document analysis and preliminary interviews
Months 3–5	Assemble ensemble methodology for computational tests and participative audits
Months 6–8	Evaluate bank dataset for loan default risks quantitatively and qualitatively
Months 9–11	Survey enterprises and agencies assessing self-mitigation motivations
Months 12–15	Analyze surveys determining adoption barriers and incentives for policies
Months 16–18	Formulate multilayered policies addressing industry structures and maturity
Months 19–20	Finalize analysis and compile dissertation

CONCLUSION

In conclusion, this robust interdisciplinary research program assesses AI ethics and bias mitigation dimensions systematically while pioneering novel ensemble techniques allying computational and

participative proficiencies holistically minimizing harms. Outcomes further inform policy formulations guiding sustainable AI advancements balancing innovation and responsibility through incentives attuning interventions contextually for maximal adoption assurances safeguarding equity standards inclusively. Collective progress hinges greatly on cooperative appraisals by social deliberation and technical ingenuity upholding civil liberties continually against creeping data determinism risks that confront present-age realities persistently requiring ethical vigilance.

REFERENCES

1. Smith CL, Blake JA, Kadin JA, Richardson JE, Bult CJ, Mouse Genome Database Group. Mouse Genome Database (MGD) – 2018: knowledgebase for the laboratory mouse. Nucleic Acids Res. 2018; 46 (D1): D836–D842.
2. Ray V, Purifoy D. The colorblind organization. In: Wooten ME, editor. Race, Organizations, and the Organizing Process. Leeds, UK: Emerald Publishing; 2019. pp. 131–150.
3. Voigt P, Von dem Bussche A. The EU General Data Protection Regulation (GDPR). A Practical Guide. 1st edition. Cham, Switzerland: Springer International Publishing; 2017.
4. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. 2019; 1 (9): 389–399.
5. Friedman B, Nissenbaum H. Bias in computer systems. ACM Trans Inform Syst. 1996; 14 (3): 330–347.
6. Suresh H, Guttag J. A framework for understanding sources of harm throughout the machine learning life cycle. In: EAAMO'21: Equity and Access in Algorithms, Mechanisms, and Optimization, New York, NY, USA, October 5–9, 2021. pp. 1–9.
7. Verma S, Rubin J. Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness, Gothenburg, Sweden, May 29, 2018. pp. 1–7.
8. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27–30, 2020. pp. 33–44.
9. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B. AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Machines. 2018; 28: 689–707.
10. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell. 2019; 1 (11): 501–507.
11. Ali M, Sapiezynski P, Bogen M, Korolova A, Mislove A, Rieke A. Discrimination through optimization: how Facebook's ad delivery can lead to biased outcomes. Proc ACM Human Computer Interact. 2019; 3 (CSCW): 1–30.

