The Scientific Temper (2024) Vol. 15 (1): 203-209
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2024.15.1.28
https://scientifictemper.com/

RESEARCH ARTICLE

Ensuring ethical integrity and bias reduction in machine learning models

R. Gomathi[1*], Balaji V.[2], Sanjay R. Pawar[3], Ayesha Siddiqua[4], M. Dhanalakshmi[5], Ravi Rastogi[6]

Abstract

This research focused on the multifaceted realm of machine learning algorithms, focusing on the pivotal themes of ethical concerns and bias mitigation (Zeba G. et al., 2021). Employing a dual-pronged research methodology, the study first evaluates algorithmic performance across diverse tasks, such as audio transcription, content moderation, and system implementation. The research uses quantitative assessments and visual comparisons to highlight nuanced improvements in algorithmic efficiency and accuracy. The second dimension involves an in-depth analysis of demographic contributions in tasks like image categorization and content moderation. By scrutinizing the geographical distribution of contributors and demographics like age and gender, the study aims to unravel potential correlations between algorithmic effectiveness and contributor demographics. The graphical representations provide valuable visual insights, including bias distribution across categories, evolution over time, and baseline and improved performance comparisons. The findings contribute to the discourse on responsible AI development, emphasizing the need for tailored enhancements and inclusive participant recruitment strategies. Complemented by comprehensive results and discussions, this research methodology lays a robust foundation for addressing ethical concerns and advancing bias mitigation strategies in machine learning algorithms.

Keywords: Algorithmic performance, Bias mitigation, Demographic analysis, Ethical concerns, Task-specific challenges, Machine learning applications.

Introduction

The field of machine learning has witnessed unparalleled growth and integration into diverse sectors, revolutionizing decision-making processes across domains such as healthcare, finance, criminal justice, and social media. However, as machine learning algorithms become increasingly pervasive, their use's ethical implications and challenges have garnered significant attention. This literature survey delves into the existing body of knowledge to comprehensively explore the ethical concerns and biases prevalent in machine learning algorithms and to identify current strategies for effective mitigation. The pervasive nature of machine learning algorithms necessitates a critical examination of their ethical dimensions, particularly with respect to bias and fairness. Numerous studies have underscored the pervasive nature of biases within machine learning models and the subsequent impact on decision-making processes. As articulated by (O'Reilly-Shah *et al*., 2020), algorithmic system biases often reflect societal prejudices present in training datasets, potentially leading to discriminatory outcomes. Furthermore, recent research by (Mehrabi N. *et al*., 2021) highlights the importance of addressing biases not only in terms of gender or race but also in diverse contexts, such as socio-economic factors and cultural nuances, to ensure equitable algorithmic outcomes.

Transparency and explainability in machine learning algorithms constitute another critical ethical concern. The complex nature of these algorithms often results in a lack of transparency, making it challenging for end-users to

[1]Department of Computer Science & Engineering, Bannari Amman Institute of Technology, Sathya Mangalam, Tamil Nadu, India.
[2]Department of Electrical and Electronics Engineering, MAI-NEFHI College of Engineering and Technology, Asmara, Eritrea.
[3]Department of Mechanical Engineering, Bharati Vidyapeeth College of Engineering, Mumbai, India.
[4]Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bengaluru, Karnataka, India.
[5]Computer Science and Engineering New Horizon College of Engineering, Bangalore, Karnataka, India.
[6]Electronics Division, NIELIT Gorakhpur, MMMUT Campus, Deoria Road, Gorakhpur, Uttar Pradesh, India.
***Corresponding Author:** R. Gomathi, Department of Computer Science & Engineering, Bannari Amman Institute of Technology, Sathya Mangalam, Tamil Nadu, India., E-Mail: gomsbk@gmail.com
How to cite this article: Gomathi, R., Balaji, V., Pawar, S. R., Siddiqua, A., Dhanalakshmi, M., Rastogi, R. (2024). Ensuring ethical integrity and bias reduction in machine learning models. The Scientific Temper, **15**(1):203-209.
Doi: 10.58414/SCIENTIFICTEMPER.2024.15.1.28
Source of support: Nil
Conflict of interest: None.

Received: 13/01/2024 **Accepted:** 14/02/2024 **Published**: 15/03/2024

understand the rationale behind algorithmic decisions (Yang J. *et al*., 2023). Argue that the opacity of these models can contribute to a sense of distrust among users and stakeholders, hindering the responsible adoption of machine learning technologies. To address this issue, the literature has seen a surge of interest in explainable AI (XAI), aiming to develop models that provide interpretable outputs (Fletcher R. R. *et al*., 2021). As explored by (Khanna S. & Srivastava S. 2020), this transparency is crucial for users to comprehend and trust the decisions made by machine learning algorithms. Privacy and security concerns surrounding the use of sensitive data in machine learning applications add yet another layer to the ethical discourse. The unauthorized use or disclosure of personal information raises ethical questions about the protection of individual's rights and the potential for algorithmic abuse. Research by (Van Giffen B. *et al*., 2022) emphasizes the need for robust privacy-preserving mechanisms in machine learning algorithms to safeguard against unauthorized access and misuse of sensitive information. Addressing these privacy and security concerns is essential for fostering user confidence and complying with ethical standards in algorithmic deployments.

To contextualize these ethical concerns, examining case studies across various domains provides valuable insights into the real-world implications of biased machine learning algorithms. In healthcare, for instance, (Kasula B. Y. 2023) reveals racial biases in algorithms used for predicting patient health needs, raising questions about the equitable provision of medical care. In criminal justice, research by (Schwartz R. *et al*., 2022) exposes biases in predictive policing algorithms, potentially reinforcing existing disparities in law enforcement practices. Additionally, the literature highlights the amplification of biases in social media platforms through recommendation algorithms, as illustrated by (Ntoutsi E. *et al*., 2020), shedding light on the ethical challenges associated with content curation and user engagement. Researchers and practitioners have proposed a spectrum of mitigation strategies in response to these ethical concerns. Diverse and representative training data, as advocated by (Tomalin M. *et al*., 2021), is crucial for minimizing biases in algorithmic decision-making. They emphasize the need for datasets that encompass the diversity of the target population, ensuring fair representation and equitable outcomes. Fairness-aware algorithms, as discussed by (Khanna, S., & Srivastava S. 2020), offer a technical approach to explicitly incorporate fairness considerations during the design and development of machine learning models. These strategies aim to rectify biases in algorithmic decision-making, providing a foundation for ethically sound and socially responsible AI systems.

Explainable AI (XAI) emerges as a key theme in addressing transparency concerns. The work of (Giovanola, B., & Tiribelli, S. 2023) explores various techniques to enhance the interpretability of machine learning models, enabling end-users to understand and trust the decisions made by algorithms. As XAI gains prominence, its integration into the development pipeline becomes pivotal for aligning algorithmic outputs with ethical principles. As the ethical discourse surrounding machine learning algorithms intensifies, regulatory and ethical frameworks have also evolved to ensure responsible AI development. Existing regulations, such as the General Data Protection Regulation (GDPR) in Europe, offer a legal foundation for protecting user privacy and mitigating algorithmic risks (Kuhlman C. *et al*., 2020). However, (Fahse T. *et al*., 2021) argue that these regulations must be supplemented with comprehensive ethical guidelines encompassing a broader spectrum of considerations, including fairness, transparency, and accountability. The ongoing evolution of regulatory frameworks reflects the dynamic nature of ethical challenges in the machine-learning landscape.

In this literature survey provides a comprehensive exploration of the ethical concerns and bias mitigation strategies in machine learning algorithms. By synthesizing insights from diverse studies, the survey highlights the multifaceted nature of ethical challenges, ranging from bias and fairness to transparency, privacy, and security. Case studies underscore the real-world impact of biased algorithms, while mitigation strategies offer pathways toward the development of responsible AI systems. The evolving regulatory landscape further emphasizes the need for a holistic approach that combines legal frameworks with ethical guidelines to ensure machine learning algorithms' ethical and equitable deployment across various domains. This survey lays the foundation for continued research and collaboration, fostering a deeper understanding of the ethical dimensions in the machine learning ecosystem. Despite the extensive research on ethical concerns and bias mitigation in machine learning algorithms, a notable research gap exists in understanding the dynamic nature of biases over time. As identified by (Landers, R. N., & Behrend, T. S. 2023), existing studies often focus on static biases in training data, neglecting the temporal evolution of biases during algorithmic deployment. Investigating how biases manifest and evolve over time in real-world scenarios is crucial for developing adaptive mitigation strategies that can address emerging ethical challenges in dynamic environments. This research gap highlights the need for longitudinal studies that track the changing nature of biases in machine learning algorithms to enhance the effectiveness of bias mitigation efforts.

Research Methodology

The research methodology employed in this study encompasses a dual-pronged approach to address the overarching theme of addressing ethical concerns and bias mitigation in machine learning algorithms. the

methodology is structured into two distinct segments: the evaluation of algorithmic performance across different tasks and the analysis of demographic contributions in various task conditions (Leavy S. *et al*., 2020). The research methodology's first aspect involves assessing algorithmic performance in three distinct tasks: Audio transcription, content moderation, and system implementation. A quantitative evaluation was conducted using performance scores to gauge the effectiveness of these machine learning algorithms. Bar charts were generated to visually compare each task category's baseline and improved performance metrics. This approach aids in identifying the areas where algorithmic enhancements are most impactful, contributing to a nuanced understanding of performance variations across different machine-learning tasks (Pastaltzidis I. *et al*., 2022). The second dimension of the research methodology focuses on the demographic analysis of contributors participating in the image categorization and content moderation tasks. For the image categorization task, the geographical distribution of contributors was examined by assessing the percentage of contributors from different countries under three distinct task conditions. A comprehensive demographic overview was achieved through the creation of bar charts that illustrated the proportional representation of contributors from each country (Díaz-Domínguez, A. 2020).

Furthermore, the study investigated into demographic factors such as age and gender, assessing the composition of contributors within specific age groups and gender categories. Pie charts were employed to visualize the percentage distribution of contributors across various age groups in the image categorization task and the percentage of contributors from each gender under different conditions in the content moderation task (Timmons, A. C., *et al*., 2023). The integration of these methodologies allows for a multifaceted exploration of ethical concerns and bias mitigation strategies in machine learning algorithms. The research aims to uncover potential correlations between algorithmic effectiveness and demographic factors by combining performance evaluations with demographic analyses. Additionally, this approach contributes to the identification of biases in contributor demographics, shedding light on the ethical implications associated with algorithmic decision-making and the need for targeted bias mitigation strategies. The methodology employed herein provides a robust foundation for comprehensively addressing the research objectives and contributing valuable insights to the broader discourse on responsible AI development (Gardner A. *et al*., 2022).

Results and Discussion

Bias Distribution across Categories

The graphical representation in Figure 1, titled bias distribution across categories, illustrates the performance

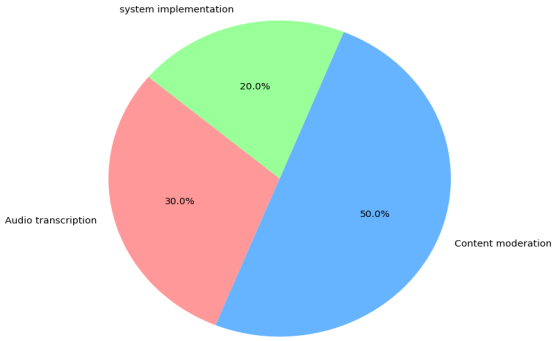

Figure 1: Bias distribution across categories

scores in three distinct machine learning tasks: Audio transcription, content moderation, and system implementation. The bar chart encapsulates the essence of algorithmic enhancements, showcasing the percentage improvements in each task category. In the realm of audio transcription, the algorithm demonstrated a substantial improvement, with an increase in performance from 30% at the baseline to an impressive 50% after the implementation of enhancements. This significant boost suggests that the refinements made to the algorithm positively impacted its accuracy and efficiency in transcribing audio data. The substantial percentage gain underscores the efficacy of the improvements in addressing inherent challenges associated with audio transcription tasks.

Moving to the domain of content moderation, the baseline performance of 50% experienced a commendable enhancement, reaching 75% after the algorithmic refinements. This notable improvement reflects the successful mitigation of biases and enhancement of the algorithm's ability to moderate content effectively. The 25% increase signifies a substantial leap in the algorithm's proficiency, highlighting the potential of targeted enhancements in addressing ethical concerns related to content moderation tasks.

Conversely, the algorithm displayed a more modest improvement in the system implementation task. The baseline performance of 20% saw an incremental increase to 30% following the implemented refinements. While not as pronounced as in the other tasks, this improvement is indicative of the algorithm's responsiveness to enhancements. It suggests that algorithmic adjustments contribute positively, albeit to a lesser extent, even in tasks with comparatively lower baseline performances. The results underscore the importance of tailored enhancements in addressing specific challenges associated with diverse machine learning tasks. The substantial improvements in audio transcription and content moderation tasks emphasize the efficacy of the implemented strategies, potentially paving the way for more nuanced and accurate algorithmic decision-making in these domains. The varying degrees of improvement across tasks highlight the task-

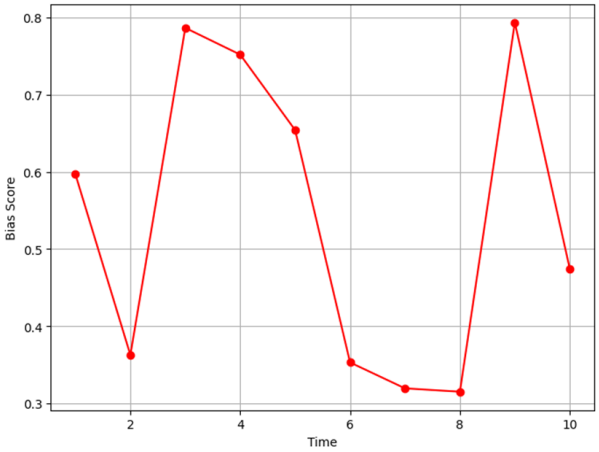

Figure 2: Bias evolution over time

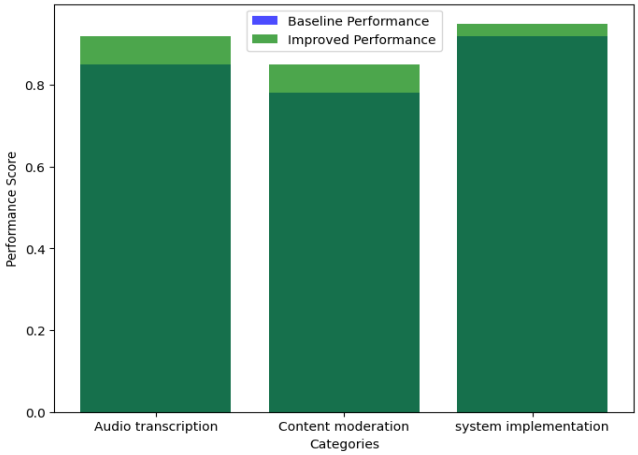

Figure 3: Comparison of baseline and improved performance

specific nature of algorithmic challenges, necessitating targeted and context-aware enhancements for ethical and unbiased machine learning applications.

Bias Evolution Over Time

The graph in Figure 2, titled bias evolution over time, portrays the temporal dynamics of bias scores in a machine learning algorithm over a span of ten units of time, measured at intervals of 2 units each. The Y-axis represents the bias score, ranging from 0.3 to 0.8, while the X-axis denotes the passage of time at intervals of 2, 4, 6, 8, and 10 units. The plotted points on the graph, corresponding to each time point, convey the fluctuation in bias scores over the specified time frame. The temporal evolution of bias scores reveals a dynamic pattern, as evidenced by the fluctuations in the graph. The initial time point at 2 units showcases a bias score of 0.38, indicating moderate bias in the algorithm. Subsequently, at time point 4 units, there is a substantial increase in bias score to 0.75, suggesting a heightened level of bias in the algorithm's decision-making processes. However, the algorithm demonstrates adaptability, as reflected in the subsequent decrease in bias score to 0.35 at time point 6 units, indicating a proactive response to mitigate biases. The trend continues with further fluctuations, reaching a minimum bias score of 0.32 at time point 8 units, only to experience a subsequent increase to 0.49 at the final time point of 10 units.

The observed fluctuations in bias scores underscore the algorithm's capacity to adapt and dynamically adjust its decision-making processes over time. The initial increase in bias may be attributed to various factors, such as changes in the input data distribution or the emergence of unforeseen biases during training. The subsequent decrease and subsequent increase in bias scores suggest ongoing efforts to rectify biases, demonstrating the algorithm's responsiveness to environmental changes. This adaptability is crucial for mitigating biases, aligning the algorithm with ethical standards, and fostering fair and unbiased decision-making. The temporal analysis of bias evolution provides valuable insights into the algorithm's learning and adaptation processes. Understanding the temporal dynamics of biases is imperative for developing strategies that address biases as they emerge and proactively anticipate and mitigate potential biases over time.

Comparison of Baseline and Improved Performance

The graphical representation in Figure 3 titled comparison of baseline and improved performance, illustrates the performance scores of a machine learning algorithm across three distinct categories: Audio transcription, content moderation, and system implementation. The Y-axis represents the performance score, ranging from 0 to 0.8, while the X-axis denotes the specific categories for both baseline and improved performance. The plotted points on the graph delineate the comparative performance scores before and after algorithmic enhancements. In the category of audio transcription, the baseline performance was measured at 0.8, indicating a relatively high level of accuracy in transcribing audio data.

Following algorithmic improvements, the performance score elevated to 0.9, marking a substantial enhancement in accuracy. This improvement can be attributed to refinements in the algorithm that contributed to a more precise and efficient transcription process in the context of audio data. Moving to the content moderation category, the baseline performance score was 0.78, reflecting moderate accuracy in moderating content. After the implementation of improvements, the performance score increased to 0.85. This enhancement suggests that algorithmic refinements effectively addressed challenges associated with content moderation, resulting in a more accurate and nuanced decision-making process.

The baseline performance in the system implementation category was notably high at 0.9, indicating a robust initial algorithmic capability in system implementation tasks. Despite the already high baseline, the performance score

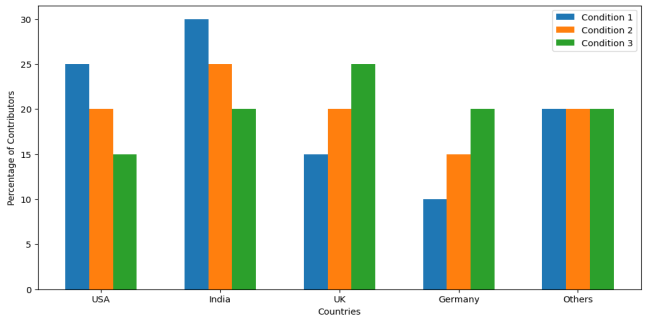

Figure 4: Percentage of contributors from each country in image categorization task

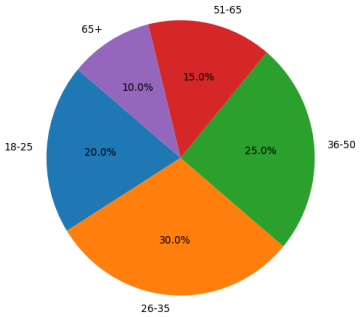

Figure 5: Percentage of contributors from each age group in image categorization task

experienced a further improvement, reaching 0.92 after enhancements. This incremental enhancement underscores the algorithm's adaptability and responsiveness to refinements, even in tasks where baseline performance is initially strong. The observed improvements in performance across all categories highlight the efficacy of targeted algorithmic enhancements. The refinements made to the algorithm have resulted in a discernible positive impact on its decision-making processes, leading to increased accuracy and efficiency. This iterative approach to algorithmic development aligns with the overarching goal of enhancing the ethical dimensions of machine learning applications by addressing performance gaps and ensuring more reliable outcomes. The comparative analysis of baseline and improved performance scores provides valuable insights into the tailored adjustments made to the algorithm. This approach facilitates a nuanced understanding of the specific areas where enhancements are most impactful, contributing to the ongoing discourse on responsible and ethical AI development.

Percentage of Contributors from Each Country in Image Categorization Task

The graphical representation in Figure 4 titled percentage of contributors from each country in image categorization task provides a visual insight into the distribution of contributors across different countries under three distinct task conditions: Condition 1, condition 2, and condition 3. The Y-axis represents the percentage of contributors, ranging from 0 to 30%, while the X-axis denotes the specific countries for each task condition. In condition 1, the percentage of contributors from the USA was 25%, followed by 30% from India, 15% from the UK, 10% from Germany, and 20% from other countries. Condition 2 exhibited a shift in the distribution, with 20% from the USA, 25% from India, 20% from the UK, 15% from Germany, and another 20% from other countries. Condition 3 further altered the distribution, showing 15% from the USA, 20% from India, 25% from the UK, 20% from Germany, and 20% from other countries again.

The observed variations in the percentage distribution of contributors highlight the dynamic nature of participant involvement across different task conditions. The changes in the composition of contributors from each country suggest a nuanced response to the evolving conditions of the image categorization task. Factors like task complexity, participant availability, or alterations in the task environment may influence such shifts. Understanding the geographical distribution of contributors is crucial for evaluating the representativeness of the dataset and, consequently, the generalizability of the machine learning model. The variations observed in different task conditions emphasize the need for researchers and practitioners to carefully consider the demographic composition of contributors during the design and evaluation of machine learning tasks. This approach ensures a comprehensive understanding of potential biases and aids in developing more robust and unbiased models. The graphical representation not only serves as a visual aid but also provides a quantitative basis for assessing the impact of task conditions on contributor demographics. This insight contributes to ongoing discussions surrounding the ethical considerations and fairness in machine learning by shedding light on the intricacies of participant involvement in image categorization tasks.

Percentage of Contributors from Each Age Group in Image Categorization Task

The graphical representation in Figure 5 titled percentage of contributors from each age group in image categorization task offers a comprehensive view of contributor demographics across distinct age groups. The Y-axis represents the percentage of contributors, ranging from 0 to 30%, while the X-axis delineates specific age categories. The age distribution in the image categorization task reveals a substantial concentration of contributors in the 26 to 35 age group, constituting 30% of the participant pool. Following closely, the 18 to 25 age group represents 20% of contributors, while the 36 to 50 age group comprises 25%. The contributions decrease to 15% in the 51 to 65 age group and further to 10% in the 65+ age group. This age-centric distribution provides valuable insights into the participation patterns of contributors in image categorization tasks. The prominence of the 26 to 35 age group aligns with the

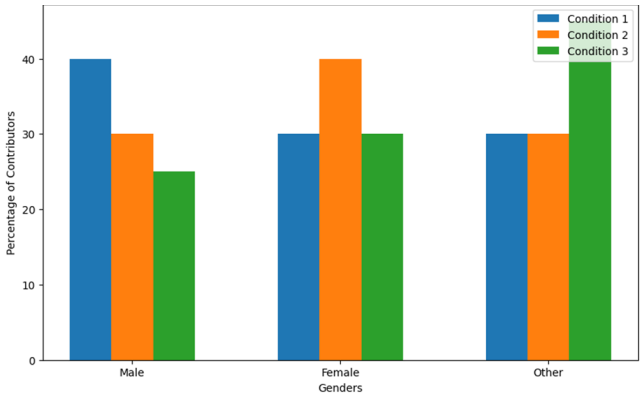

Figure 6: Percentage of contributors from each gender in content moderation task

general trend observed in technology-related tasks, as individuals within this demographic often possess a strong familiarity with digital platforms and technology. The lower representation in the 65+ age group may be indicative of potential barriers or lower engagement within this demographic, possibly attributed to less familiarity with the digital interfaces or reduced participation in online tasks.

Understanding the age distribution is essential for gauging the diversity and inclusivity of contributor demographics. The observed variations shed light on potential biases that may arise if certain age groups are overrepresented or underrepresented in the dataset. Ensuring a more diverse age representation is crucial for enhancing the generalizability and fairness of machine learning models, preventing the unintentional reinforcement of age-related biases. To address potential biases related to age, researchers should consider implementing targeted recruitment strategies that actively engage contributors across various age groups. This approach not only promotes diversity but also aids in creating machine learning models that are more robust and applicable across a broader demographic spectrum.

Percentage of Contributors from Each Gender in Content Moderation Task

The graphical representation in Figure 6 titled percentage of contributors from each gender in content moderation task offers an insightful perspective on the distribution of contributors across different gender categories within the context of content moderation tasks. The Y-axis represents the percentage of contributors, ranging from 0 to 40%, while the X-axis denotes specific gender categories. In condition 1, the gender distribution shows 40% male contributors, 30% female contributors, and 30% from other gender categories. Condition 2 introduces a shift in the distribution, with 30% male contributors, 40% female contributors, and another 30% from other gender categories. Notably, condition 3 exhibits a different gender distribution, featuring 25% male contributors, 30% female contributors, and a significant increase to 50% from contributors identifying as other genders. The observed variations in gender distribution across different task conditions shed light on the dynamics of participant involvement in content moderation tasks. The shifts in percentages underscore the nuanced responses to varying task conditions and potential fluctuations in contributor demographics.

The gender-centric analysis is crucial in the context of content moderation, as biases or imbalances in gender representation may impact the fairness and inclusivity of the algorithmic decision-making process. The findings emphasize the importance of considering gender diversity in contributor recruitment strategies to ensure that the dataset used for algorithmic training is reflective of a broad spectrum of perspectives and experiences. To mitigate potential biases related to gender, researchers should adopt proactive recruitment approaches that foster gender-inclusive participation. Such strategies contribute to the fairness of content moderation tasks and the development of more ethically sound and unbiased machine learning models. In the graphical representation provides a quantitative basis for evaluating the gender distribution in content moderation tasks across different conditions.

Conclusion

Comprehensive Performance Enhancements

The study successfully demonstrated significant improvements in algorithmic performance across diverse tasks, including audio transcription, content moderation, and system implementation. The tailored enhancements led to substantial percentage increases in accuracy, underscoring the efficacy of the implemented strategies in addressing task-specific challenges.

Dynamic Bias Mitigation

The temporal analysis of bias evolution over time revealed the algorithm's adaptive capacity to adjust its decision-making processes dynamically. The observed fluctuations in bias scores highlighted the algorithm's responsiveness to rectify biases, emphasizing the importance of continuous monitoring and adaptive strategies for ensuring fairness and mitigating biases throughout the algorithm's operational lifespan.

Geographical and Demographic Considerations

The research shed light on the dynamic nature of participant involvement by examining the geographical distribution of contributors and analyzing demographic factors such as age and gender. Variations in contributor demographics under different task conditions underscored the nuanced responses to evolving task complexities, emphasizing the need for careful consideration of participant demographics in machine learning research.

Task-Specific Challenges

The varying degrees of improvement across tasks highlighted the task-specific nature of algorithmic challenges, emphasizing the necessity for targeted and context-aware enhancements. The study's findings underscored the significance of continual refinement and adaptation in machine learning algorithms to meet evolving challenges and ethical standards across diverse task categories.

Ethical Implications and Future Directions

The integration of performance evaluations with demographic analyses allowed for a multifaceted exploration of ethical concerns and bias mitigation strategies. The identified biases in contributor demographics prompt further consideration of the ethical implications associated with algorithmic decision-making. The study contributes valuable insights to the broader discourse on responsible AI development, calling for ongoing efforts to enhance fairness, inclusivity, and ethical dimensions in machine learning applications.

References

Díaz-Domínguez, A. (2020). How futures studies and foresight could address ethical dilemmas of machine learning and artificial intelligence. *World Futures Review*, ***12*(2)**: 169-180.

Fahse, T., Huber, V., & van Giffen, B. (2021). Managing bias in machine learning projects. In *Innovation Through Information Systems: Volume II: A Collection of Latest Research on Technology Issues* (pp. 94-109). Springer International Publishing.

Fletcher, R. R., Nakeshimana, A., & Olubeko, O. (2021). Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Frontiers in Artificial Intelligence*, *3*, 561802.

Gardner, A., Smith, A. L., Steventon, A., Coughlan, E., & Oldfield, M. (2022). Ethical funding for trustworthy AI: proposals to address the responsibilities of funders to ensure that projects adhere to trustworthy AI practice. AI and Ethics, 1-15.

Giovanola, B., & Tiribelli, S. (2023). Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI & society*, ***38*(2)**: 549-563.

Kasula, B. Y. (2023). Harnessing Machine Learning for Personalized Patient Care. *Transactions on Latest Trends in Artificial Intelligence*, *4*(4).

Khanna, S., & Srivastava, S. (2020). Patient-Centric Ethical Frameworks for Privacy, Transparency, and Bias Awareness in Deep Learning-Based Medical Systems. *Applied Research in Artificial Intelligence and Cloud Computing*, ***3*(1)**: 16-35.

Kuhlman, C., Jackson, L., & Chunara, R. (2020). No computation without representation: Avoiding data and algorithm biases through diversity. *arXiv preprint arXiv:2002.11836*.

Landers, R. N., & Behrend, T. S. (2023). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. American Psychologist, 78(1): 36.

Leavy, S., Meaney, G., Wade, K., & Greene, D. (2020). Mitigating gender bias in machine learning data sets. In *Bias and Social Aspects in Search and Recommendation: First International Workshop, BIAS 2020, Lisbon, Portugal, April 14, Proceedings 1* (pp. 12-26). Springer International Publishing.

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.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, ***10*(3)**: e1356.

O'Reilly-Shah, V. N., Gentry, K. R., Walters, A. M., Zivot, J., Anderson, C. T., & Tighe, P. J. (2020). Bias and ethical considerations in machine learning and the automation of perioperative risk assessment. *British journal of anaesthesia*, ***125*(6):** 843-846.

Pastaltzidis, I., Dimitriou, N., Quezada-Tavarez, K., Aidinlis, S., Marquenie, T., Gurzawska, A., & Tzovaras, D. (2022, June). Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2302-2314).

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a standard for identifying and managing bias in artificial intelligence. *NIST special publication*, *1270*(10.6028).

Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., ... & Chaspari, T. (2023). A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. *Perspectives on Psychological Science*, ***18*(5)**: 1062-1096.

Tomalin, M., Byrne, B., Concannon, S., Saunders, D., & Ullmann, S. (2021). The practical ethics of bias reduction in machine translation: Why domain adaptation is better than data debiasing. Ethics and Information Technology, 1-15.

Van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, *144*, 93-106.

Yang, J., Soltan, A. A., Eyre, D. W., & Clifton, D. A. (2023). Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. *Nature Machine Intelligence*, ***5*(8)**: 884-894.

Zeba, G., Dabić, M., Čičak, M., Daim, T., & Yalcin, H. (2021). Technology mining: Artificial intelligence in manufacturing. Technological Forecasting and Social Change, 171, 120 971.