Journal of Social Research

P-ISSN: 2827-9832
E-ISSN: 2828-335x
http://ijsr.internationaljournallabs.com/index.php/ijsr

ARTIFICIAL INTELLIGENCE AND BIAS: CHALLENGES, IMPLICATIONS, AND REMEDIES

Alfonso Min
Human Rights Commission of Malaysia
alfonso@suhakam.org.my

ABSTRACT

This paper investigates the multifaceted issue of algorithmic bias in artificial intelligence (AI) systems and explores its ethical and human rights implications. The study encompasses a comprehensive analysis of AI bias, its causes, and potential remedies, with a particular focus on its impact on individuals and marginalized communities. The primary objectives of this research are to examine the concept of algorithmic bias, assess its ethical and human rights implications, identify its causes and mechanisms, evaluate its societal impact, explore mitigation strategies, and examine regulatory and community-driven approaches to address this critical issue. The research employs a multidisciplinary approach, drawing from literature reviews, case studies, and ethical analyses. It synthesizes insights from academic papers, governmental reports, and industry guidelines to construct a comprehensive overview of algorithmic bias and its ramifications. This research paper underscores the urgency of addressing algorithmic bias, as it raises profound ethical and human rights concerns. It advocates for comprehensive approaches, spanning technical, ethical, regulatory, and community-driven dimensions, to ensure that AI technologies respect the rights and dignity of individuals and communities in our increasingly AI-driven world.

Keywords: *artificial intelligence and bias, challenges, implications, remedies*

INTRODUCTION

Artificial Intelligence (AI) and Its Ubiquity

Artificial Intelligence, often abbreviated as AI, is a transformative technology that simulates human intelligence in machines, enabling them to perform tasks typically requiring human intelligence, such as learning, reasoning, problem-solving, and decision-making. AI systems are designed to analyze vast amounts of data, recognize patterns, and make predictions or decisions autonomously.

Over the past few decades, AI has evolved from a theoretical concept to a pervasive and integral part of our daily lives. This evolution has been driven by advancements in machine learning, deep learning, and computational power. AI has found applications in various sectors, including healthcare, finance, transportation, education, and more. Its prevalence is evident in virtual personal assistants, autonomous vehicles, medical diagnosis, and recommendation systems.

AI's Role in Decision-Making Processes

AI's significance lies in its ability to enhance and streamline decision-making processes across diverse domains. It empowers organizations, institutions, and individuals to make data-driven decisions, improve efficiency, and optimize outcomes

Impact on Individuals' Lives

The integration of AI into decision-making processes has a profound impact on individuals' lives:

1. Accessibility and Convenience: AI-driven technologies, like voice-activated virtual assistants, make daily tasks more accessible and convenient, enhancing overall quality of life (Dixon et al., 2018)
2. Career Opportunities and Job Market: AI has created new career opportunities and transformed the job market. It demands a workforce skilled in AI-related fields, shaping educational and professional trajectories (McKinsey & Company, 2017).
3. Ethical Considerations: AI decisions can affect individuals' privacy, security, and rights. Addressing ethical considerations, such as bias and discrimination, becomes crucial (Jobin et al., 2019).
4. Healthcare Outcomes: AI's impact on healthcare can mean more accurate diagnoses and personalized treatments, potentially saving lives and improving the overall well-being of patients (Obermeyer et al., 2019).

In conclusion, AI's increasing prevalence and its role in decision-making processes have far-reaching implications for individuals and society at large. While it offers numerous benefits, it also presents challenges related to ethics, accountability, and the need for responsible AI development and deployment.

Algorithmic Bias and Discrimination in AI Systems

In recent years, the rapid proliferation of artificial intelligence (AI) technologies across various sectors has raised critical concerns regarding algorithmic bias and discrimination. Algorithmic bias refers to the presence of systematic and unfair disparities in the outcomes of AI systems, often affecting individuals from marginalized or underrepresented groups.

Ethical and Human Rights Implications of Biased AI

The ethical and human rights implications of biased AI are multifaceted and require careful consideration:

1. Right to Non-Discrimination:
   - AI systems that discriminate against certain racial, gender or socioeconomic groups violate the right to non-discrimination, a fundamental human right (European Commission, 2020).
   - Discriminatory AI can reinforce and perpetuate societal biases, leading to unequal treatment (Crawford & Schultz, 2013).
2. Privacy Violations:
   - Biased AI systems may invade individuals' privacy by making decisions based on sensitive personal attributes, such as race or gender (European Commission, 2020).
   - This compromises the right to privacy, necessitating robust data protection measures.
3. Transparency and Accountability:
   - The lack of transparency in AI decision-making processes can hinder accountability (Crawford et al., 2016).
   - Individuals have the right to know how and why decisions are made about them, especially when these decisions have significant consequences.
4. Impact on Vulnerable Communities:

- Vulnerable and marginalized communities often bear the brunt of biased AI decisions, exacerbating existing disparities (Law, 2018).
- This highlights the need to safeguard the rights of these communities through ethical AI practices.

In conclusion, the issue of algorithmic bias and discrimination in AI systems has profound ethical and human rights implications. Addressing this problem requires a concerted effort to develop fair, transparent, and accountable AI technologies that respect the rights and dignity of all individuals.

METHOD

The research methodology employed in this study is firmly rooted in a multidisciplinary approach, strategically combining various research techniques and data sources to offer a comprehensive understanding of algorithmic bias and its extensive implications. This multifaceted approach draws upon three fundamental methodologies: literature reviews, case studies, and ethical analyses.

Firstly, literature reviews serve as the cornerstone of this research. These reviews entail a thorough examination and synthesis of academic papers, reports, and studies sourced from reputable outlets across diverse fields. By harnessing the collective knowledge of experts in AI ethics, bias mitigation, and algorithmic fairness, the research builds a robust foundation grounded in the most recent developments and insights in the field.

Secondly, case studies play a pivotal role in this research methodology. In-depth scrutiny of real-world cases and examples serves to illustrate and substantiate the existence and impact of algorithmic bias. These case studies offer concrete evidence of how biased AI systems can manifest across various domains, including finance, healthcare, and criminal justice. They provide compelling narratives that underscore the tangible implications of algorithmic bias on individuals and communities.

Lastly, ethical analyses represent an integral component of the research methodology. Ethical considerations are seamlessly interwoven throughout the study to assess the moral implications of algorithmic bias. This encompasses an examination of human rights principles, such as non-discrimination and privacy, within the context of AI development. Ethical guidelines and frameworks are systematically applied to evaluate the alignment of AI systems with these fundamental principles.

In conclusion, the research methodology is meticulously designed to facilitate a comprehensive and well-rounded exploration of algorithmic bias. By seamlessly integrating literature reviews, case studies, and ethical analyses, this approach ensures that the study not only acknowledges the existence of bias but also delves into its origins, repercussions, and potential remedies. This multidisciplinary methodology fosters a deeper understanding of the intricate issue of algorithmic bias, ultimately paving the way for informed discussions and policy recommendations within the realm of AI ethics.

RESULTS AND DISCUSSION

Impact on Individuals

Real-World Consequences of Biased AI

Hiring and Employment: Biased AI used in hiring processes can perpetuate discrimination. For instance, if an AI-based applicant screening system is biased against certain demographics, qualified candidates may be unfairly excluded from job opportunities. This can lead to decreased job prospects, economic disparities, and emotional distress among affected individuals (Dastin, 2022).

Lending and Financial Services: Biased AI algorithms in lending can result in unequal access to financial resources. If AI-driven lending decisions favor specific groups, others may face difficulties obtaining loans or credit, hindering their financial stability and potential for economic growth (Crawford & Schultz, 2013).

Criminal Justice: AI systems used in criminal justice, such as risk assessment algorithms, have been found to exhibit racial bias. These biases can lead to disproportionately harsh sentencing and parole decisions for marginalized communities, perpetuating inequalities in the criminal justice system (Kleinberg et al., 2017).

Healthcare: In healthcare, biased AI can have life-threatening consequences. For example, if medical diagnostic AI systems exhibit racial bias, patients from certain racial backgrounds may receive delayed or incorrect diagnoses, impacting their health outcomes and quality of life (Obermeyer et al., 2019).

Disproportionate Effects on Marginalized Communities

Racial Disparities: Marginalized racial groups, such as Black and Hispanic communities, are often disproportionately affected by biased AI. Discriminatory AI in various domains can exacerbate existing racial disparities, from education to law enforcement (Obermeyer et al., 2019).

Gender Disparities: Gender bias in AI can harm women and gender-diverse individuals. For example, biased AI used in hiring processes may favor male applicants, limiting career opportunities for women and perpetuating gender wage gaps (Dastin, 2022).

Socioeconomic Inequities: Biased AI can reinforce socioeconomic disparities. When AI systems discriminate against individuals from lower-income backgrounds, it can hinder their access to opportunities, resources, and services, further entrenching inequality (Crawford & Schultz, 2013).

Vulnerable Populations: Vulnerable populations, including the elderly and individuals with disabilities, may face unique challenges due to biased AI. For example, if AI-driven healthcare systems exhibit age bias, elderly patients may receive suboptimal care or be denied access to certain medical treatments (Obermeyer et al., 2019).

Privacy and Data Exploitation: Marginalized communities are often disproportionately impacted by privacy violations resulting from biased AI. The exploitation of personal data through biased AI systems can disproportionately affect individuals with limited resources to protect their privacy (Law, 2018).

The impact of biased AI on individuals and marginalized communities underscores the urgent need for ethical AI development and the application of fairness-aware algorithms. Addressing these issues is not only a matter of technical responsibility but also a crucial step toward upholding human rights principles and ensuring equitable opportunities and access to resources for all.

Amazon's Gender-Biased Recruitment Algorithm

Amazon's recruitment algorithm was trained on a decade's worth of resumes submitted to the company. As a result, most of the data came from male applicants due to historical hiring patterns at Amazon. The algorithm learned to prioritize resumes that resembled those of existing Amazon employees, who were predominantly male, reflecting an implicit bias toward male candidates. The algorithm consistently downgraded resumes containing terms or experiences associated with women. It led to systematic discrimination against female candidates, making it more challenging for qualified women to advance in the recruitment process (Dastin, 2022).

This case study illustrates how biased training data and implicit biases among those involved in algorithm development can lead to discriminatory AI outcomes. Amazon's recruitment algorithm demonstrated gender bias by favoring male candidates, which highlights the importance of fairness and diversity in AI training data.

ProPublica's Analysis of COMPAS Risk Assessment Tool

The COMPAS risk assessment tool, widely used in the U.S. criminal justice system, was trained on historical criminal records data. This data contained racial disparities in arrests and convictions. The proprietary nature of the COMPAS algorithm and its complexity made it difficult to discern how it made decisions, potentially hiding bias. ProPublica's analysis revealed that COMPAS showed racial bias in its risk assessments. It tended to overpredict the risk of recidivism for black defendants while underpredicting it for white defendants(Räz, 2022).

The COMPAS case study highlights how biased training data and the opacity of complex algorithms can perpetuate racial disparities in the criminal justice system. It underscores the need for transparency and fairness in AI systems used for critical decision-making.

Gender and Racial Bias in Facial Recognition Technology

Facial recognition systems were often trained on imbalanced datasets with a majority of lighter-skinned and male faces. This biased training data led to algorithms that performed poorly on darker-skinned individuals and women. Some algorithms prioritized certain facial features that were more prevalent in the training data, further exacerbating bias. Facial recognition systems exhibited gender and racial bias, misclassifying individuals with darker skin tones more frequently and underrepresenting women. This had real-world consequences, including wrongful arrests and misidentifications (Buolamwini & Gebru, 2018).

Gender and racial bias in facial recognition technology showcase how biased training data and algorithmic design choices can lead to unjust outcomes, including misidentification and potential harm to marginalized groups. It underscores the ethical imperative of addressing these biases. These case studies exemplify the real-world consequences of AI bias and discrimination. They underscore the critical need for fairness, transparency, and ethical considerations in AI development and deployment to prevent unjust outcomes and uphold human rights principles

Data Collection and Preprocessing

Strategies for Collecting Diverse and Representative Training Data

Utilize a wide range of data sources to ensure diversity. This includes sources from different geographical regions, cultural backgrounds, and socioeconomic contexts. Incorporating data from various contexts helps reduce bias and ensures a more comprehensive representation. Ensure that data collection methods are inclusive and considerate of all demographics. Outreach efforts, surveys, and feedback mechanisms can be employed to actively engage underrepresented groups in data collection (Gebru et al., 2021).

Augment training data by introducing variations and synthetic examples. This can help balance the representation of different groups and increase the diversity of the dataset (Shorten & Khoshgoftaar, 2019). Implement fair sampling techniques to address imbalances in the data. Oversampling underrepresented groups and undersampling overrepresented groups can help achieve a more equitable distribution (Chawla et al., 2002).

Techniques for Data Preprocessing to Reduce Bias

Conduct rigorous data cleaning to identify and rectify biased or erroneous data points. This process involves detecting and addressing data anomalies that can introduce bias into the dataset (Doshi-Velez & Kim, 2017). Carefully engineer features to reduce bias. Feature selection and transformation can help remove or mitigate the impact of sensitive attributes that might contribute to bias (Kamiran & Calders, 2012).

Apply techniques such as oversampling and undersampling to balance the representation of different groups in the dataset. This helps prevent the model from favoring majority groups (Chawla et al., 2002). Implement ethical and responsible data labeling practices. Human annotators should be provided with clear guidelines on avoiding bias and stereotypes when llabelingdata (Gebru et al., 2021). Employ bias detection algorithms to identify bias in the data and subsequent bias mitigation techniques. Fairness-aware machine learning models can be used to mitigate bias during model training (Hardt et al., 2016).

Implement privacy-preserving techniques such as differential privacy to protect sensitive attributes in the data. This ensures that the privacy of individuals is maintained while reducing the risk of bias (Dwork & Roth, 2013). Choose interpretable machine learning models that allow for better understanding and identification of bias. Transparency in model behavior aids in addressing and mitigating bias (Rudin, 2019).

By employing these strategies for collecting diverse and representative training data and implementing data preprocessing techniques to reduce bias, developers and data scientists can work towards creating AI systems that are fair, unbiased, and aligned with ethical and human rights principles.

Algorithmic Fairness

Approaches for Designing Fair and Accountable AI Algorithms

Incorporate fairness constraints into the model's objective function during training. These constraints ensure that the model's predictions are statistically similar across different demographic groups, such as race or gender. Common fairness metrics include equal opportunity and demographic parity (Hardt, 2016). Apply regularization methods to penalize models for making biased predictions. By adding fairness-related terms to the objective function, the model is encouraged to produce equitable outcomes while optimizing for accuracy (Hardt et al., 2016).

Assign different weights to different data points to give more importance to underrepresented or disadvantaged groups. This technique helps the model learn from all groups equally, reducing bias (Chen et al., 2018). Implement adversarial networks within the model architecture. An adversarial network attempts to detect and counteract bias in the model's predictions, promoting fair outcomes (Zhang et al., 2022). After model training, apply post-processing interventions to adjust model predictions to achieve fairness goals. These interventions can correct biased outputs and make predictions more equitable (Hardt, 2016).

Trade-offs between Fairness and Accuracy

Achieving both fairness and accuracy can be challenging as there is often a trade-off between the two. Fairness constraints may restrict the model's ability to make accurate predictions, and vice versa. Striking the right balance is crucial (Chouldechova, 2017). Focusing too much on fairness might result in group disparities where certain groups consistently receive favorable or uunfavorabletreatment from the model, regardless of their actual characteristics (Crawford et al., 2016).

Overly aggressive fairness constraints can lead to the loss of information about individual instances, making it challenging to provide tailored and accurate recommendations or decisions (Hardt et al., 2016). Fairness interventions can increase the complexity of models, making them less interpretable. It becomes crucial to balance fairness with transparency and interpretability (Rudin, 2019). Striving for fairness may involve making ethical trade-offs, such as deciding how to allocate limited resources fairly. These decisions can be complex and require careful consideration (Kleinberg et al., 2017).

In conclusion, designing fair and accountable AI algorithms is a complex task that involves trade-offs between fairness and accuracy. Striking the right balance is essential to ensure that AI systems produce equitable outcomes without compromising their overall effectiveness. Ethical considerations, transparency, and ongoing monitoring are vital components of achieving algorithmic fairness.

Regulation and Ethical Guidelines

Government Regulations

Existing anti-discrimination laws, such as the U.S. Civil Rights Act and the European Union's Equal Treatment Directive, apply to AI systems. They prohibit discrimination based on protected characteristics like race, gender, and age, imposing legal obligations on organizations to ensure fairness in AI (Calo, 2018). Governments are increasingly considering AI impact assessments as part of their regulatory efforts. These assessments require organizations to evaluate the potential impact of AI systems on human rights, including the risk of bias and discrimination (European Commission, 2021). Various sectors, such as finance and healthcare, have industry-specific regulations that pertain to AI fairness. For example, the U.S. Equal Credit Opportunity Act regulates fairness in lending decisions, including those made by AI algorithms.

Industry-Specific Guidelines

Industry organizations and consortia, such as the Partnership on AI and IEEE, have developed ethical AI principles and guidelines. These frameworks emphasize fairness, transparency, accountability, and human rights considerations in AI. Many companies have established AI ethics committees or advisory boards to guide responsible AI development. These committees often include external experts who assess AI systems for bias and ethical concerns. International standards organizations, including ISO and IEEE, have developed standards related to AI ethics and fairness. These standards provide best practices for organizations to ensure fairness in AI systems (ISO, 2019).

Impact of Frameworks like the European Union's AI Act

The European Union's AI Act, proposed in April 2021, represents a significant development in the regulation of AI and AI bias. Key aspects include:

The AI Act introduces a risk-based approach to AI regulation. High-risk AI applications, such as those used in critical infrastructure or healthcare, are subject to strict requirements, including impact assessments and data quality checks. The AI Act explicitly prohibits AI systems that manipulate human behavior, use biometric data for surveillance, or exploit vulnerable groups. These prohibitions aim to prevent biased and harmful AI practices. Organizations deploying high-risk AI systems must undergo a conformity assessment to ensure compliance with AI Act requirements. This includes addressing potential bias and discrimination in AI systems (European Commission, 2021). The AI Act emphasizes transparency and accountability, requiring organizations to provide clear information about AI systems' functionality and decision-making processes. This transparency helps address bias concerns (European Commission, 2021).

The impact of the European Union's AI Act extends beyond the EU, as it is likely to influence global AI regulations and best practices. It reflects the growing recognition of the need for legal and ethical frameworks to address AI bias and uphold human rights in AI development. Government regulations and industry-specific guidelines play a pivotal role in addressing AI bias by setting legal obligations and ethical standards for AI developers. Frameworks like the European Union's AI Act mark significant progress toward ensuring fairness and accountability in AI systems.

CONCLUSION

Bias is a widespread issue in AI systems across domains like finance, healthcare, and criminal justice. Unfortunately, marginalized communities, including racial minorities, tend to bear the brunt of biased AI decisions. One significant source of AI bias lies in the biased training data, which often leads to skewed predictions. To address these issues, human rights principles must guide AI development, ensuring fairness and privacy are upheld. Furthermore, it's crucial to acknowledge that poor algorithm design can significantly contribute to AI bias, underscoring the need for fairness-aware algorithms.

The consequences of biased AI are tangible and far-reaching, affecting various aspects of life such as hiring, lending, criminal justice, and healthcare, ultimately leading to unfair treatment. Specific examples presented in the paper vividly illustrate the real-world impact of AI bias.

Mitigation strategies are available to combat AI bias, including data preprocessing techniques and integrating ethical considerations throughout the AI development process.

Additionally, government regulations and industry-specific guidelines, like the EU's AI Act, play a vital role in ensuring accountability and fairness.

Engaging affected communities in AI development is essential to identify and address bias effectively. Diverse accountability mechanisms, such as independent audits, ethical guidelines, regulatory oversight, transparency reports, and public feedback channels, collectively contribute to tackling AI bias.

In conclusion, addressing AI bias is essential not only for the advancement of technology but also for upholding human rights and ensuring a fair and just society. Future research and policies must consider the far-reaching implications of bias while striving for transparency, fairness, and accountability in AI systems.

REFERENCES

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research, 81*.

Calo, R. (2018). Artificial intelligence policy: A primer and roadmap. In *University of Bologna Law Review* (Vol. 3, Issue 2). https://doi.org/10.6092/issn.2531-6133/8670

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*. https://doi.org/10.1613/jair.953

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, *5*(2). https://doi.org/10.1089/big.2016.0047

Crawford, K., & Schultz, J. (2013). Big data and due processes: Toward a framework to redress predictive privacy harms. In *Public Law & Legal Theory Research Paper Series*.

Crawford, K., Whittaker, M., Elish, M. C., Barocas, S., Plasek, A., & Ferryman, K. (2016). The AI Now Report. *The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term*.

Dastin, J. (2022). Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women *. In *Ethics of Data and Analytics*. https://doi.org/10.1201/9781003278290-44

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and Mitigating Unintended Bias in Text Classification. *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3278721.3278729

Doshi-Velez, F., & Kim, B. (2017). A Roadmap for a Rigorous Science of Interpretability. *ArXiv Preprint ArXiv:1702.08608v1*.

Dwork, C., & Roth, A. (2013). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, *9*(3–4). https://doi.org/10.1561/0400000042

European Commission. (2020). White Paper On Artificial Intelligence - A European approach to excellence and trust. *Journal of Chemical Information and Modeling, June*.

European Commission. (2021). Proposal for a Regulation on a European Approach for Artificial Intelligence. *Journal of Chemical Information and Modeling, December 2022*.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. In *Communications of the ACM* (Vol. 64, Issue 12). https://doi.org/10.1145/3458723

Hardt, M. (2016). *Equality of Opportunity in Machine Learning*. Google Research Blog.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9). https://doi.org/10.1038/s42256-019-0088-2

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1). https://doi.org/10.1007/s10115-011-0463-8

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *Leibniz International Proceedings in Informatics, LIPIcs*, *67*. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Law, T. (2018). Review of Eubanks' Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. *Surveillance & Society*, *16*(3). https://doi.org/10.24908/ss.v16i3.12612

McKinsey & Company. (2017). Jobs lost, jobs gained: Workforce transitions in a time of automation. *Донну*, *December*.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464). https://doi.org/10.1126/science.aax2342

Räz, T. (2022). COMPAS: on a pathbreaking debate on algorithmic risk assessment. *Forensische Psychiatrie, Psychologie, Kriminologie*, *16*(4). https://doi.org/10.1007/s11757-022-00741-9

Rudin, C. (2019). Stop explaining black-box machine learning models for high-stakes decisions and use interpretable models instead. In *Nature Machine Intelligence* (Vol. 1, Issue 5). https://doi.org/10.1038/s42256-019-0048-x

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0197-0

Zhang, K., Khosravi, B., Vahdati, S., Faghani, S., Nugen, F., Rassoulinejad-Mousavi, S. M., Moassefi, M., Jagtap, J. M. M., Singh, Y., Rouzrokh, P., & Erickson, B. J. (2022). Mitigating Bias in Radiology Machine Learning: 2. Model Development. *Radiology: Artificial Intelligence*, *4*(5). https://doi.org/10.1148/ryai.220010