Reducing Power Inequality in the Age of AI and the Internet: A Practical Summary of Experience Using Toolkits

Sixuan Liu*

Department of culture media&creative industries, King's college London, London, United Kingdom

*Corresponding author: vivianxuan77@gmail.com

Abstract. This article discusses the potential power inequalities caused by algorithmic bias in the age of artificial intelligence, aiming to guide technology companies in identifying biases throughout the AI lifecycle and taking effective measures to reduce discrimination arising from technology and algorithms. It emphasizes the importance of algorithmic process transparency, advocates for increased diversity within AI development teams, and promotes designing frameworks based on the dataset to ensure that harmful biases do not permeate AI system functionality. These measures help improve the ethical application of AI technology and enhance the company's credibility and moral standing. Furthermore, the toolkit will assist stakeholders in the technology field in making informed decisions and guide decision-makers in building a sustainable AI ecosystem from an ethical standpoint. In this ecosystem, the advancements in AI enable all segments of society to benefit without discrimination.

Keywords: Toolkit, Technology Ethics, Machine Learning, and AI.

1. Toolkit Scope

Unfairness, bias of gender, inequality in race… In the fast-changing field of artificial intelligence, these are not optional considerations. These are all necessary elements for the equitable development of technology in the service of humanity. As application of AI have being extended into almost every sector—health, finance, and law enforcement—problems of biased algorithms in these systems will largely extend and further perpetuate inequalities and lead to the most adverse outcomes for marginalized groups. This toolkit is designed to help enable technology companies or organizations to develop and deploy AI systems based on fairness and ethical integrity.

The toolkit will guide technology companies in bias identification throughout the AI lifecycle and take robust steps to reduce discrimination issues brought about by technology and algorithms. In fact, focusing on algorithmic process transparency, increasing the inclusion of teams involved in developing AI systems, and designing frameworks based on this design dataset will help ensure that potentially harmful biases do not seep into AI system functionality. These elements can help improve the ethical application of AI technology and further enhance the credibility and moral standing of the company. But more importantly, these initiatives greatly improve the overall effectiveness of AI solutions in all global markets where AI is applied. The toolkit will also help stakeholders in technological field make informed decisions and guide decision makers to build a sustainable AI ecosystem from the level of ethical constraints. In this ecosystem, advances in AI enable all segments of society to benefit without discrimination.

2. Annotated Glossary

Algorithmic bias: systematic and repeatable errors of an algorithm in a computer system that gives unfair outcomes, such as privileging one arbitrary group of users over another. It will majorly emanate from misrepresentation incomplete training data, or flawed assumptions in the algorithmic model.

Data ethics: this is an area of ethics that is concerned with the responsibility of the governance of data, in particular considering ways in which data might be collected, shared, and used. In the light

of AI, data ethics would guide the handling of data to respect privacy and be characterized by transparency, fairness, and accountability.

Ethical AI: AI systems designed to operate based on principles in which considerations of fairness, accountability, transparency, and ethical behavior stand at the center of their functioning. Ethical AI is the design, development, and deployment of AI systems with a view toward the morals involved and the societal impacts of the technology.

Fairness in Machine Learning: An algorithm development process that does not create or reinforce unfair societal bias. Fairness relates to the equity of AI systems and their outputs across different groups with different characteristics, especially between protected and vulnerable groups.

T**airness in Machine Learning:** An algorithm development process that does not create or reinforce unfair societal bias. Fairness relates to the equity of AI systems and their outputs across different groups with different characteristics, especially between protected and vulnerable groups.

3. Annotated Bibliography

Robert, L. P., Pierce, C., Morris, L., Kim, S., Alahmad, R. (2020). "Designing Fair AI for Managing Employees in Organizations: A Review, Critique, and Design Agenda," *Human–Computer Interaction*. Accepted to the Special Issue on Unifying Human Computer Interaction and Artificial Intelligence.

This paper focuses on the theory of organizational equity and three types of equity: distributive equity, procedural equity and interactive equity. This design literature reviews existing design insights and proposes a structured design agenda to improve the fairness of AI systems. They further discussed other effects of AI inequity, such as reduced employee effort and high turnover, but also the need for AI systems to support some form of equity and help correct unfair situations. This paper is essential for understanding the concept of equity in AI-driven human resource management from a theoretical and practical perspective.

Bateni, A., Chan, M. C., Eitel-Porter, R. (2022). "AI Fairness: from Principles to Practice," published by Accenture.

This wide-ranging paper examines various methodologies and practical guidelines to bring about fairness in AI systems. Authors reviewed the most common definitions, measures, and mitigation techniques evaluatively. They raised concerns about the universal application of principles of fairness, which can be complex in their diversity of AI applications. They went on offering in-depth reviews of these limitations in over-simplistic bias assessment methods and advanced methods that are more effective.

Liu, M., Ning, Y., Teixayavong, S., Mertens, M., Xu, J., Ting, D. S. W., Cheng, L. T. E., Ong, J. C. L., Teo, Z. L., Tan, T. F., RaviChandran, N., Wang, F., Celi, L. A., Ong, M. E. H., & Liu, N. (2023). "A Translational Perspective Towards Clinical AI Fairness," npj Digital Medicine, 6:172.

This paper is addressed to AI practitioners, organizational leaders, and policymakers. It will be of direct interest to the general readers as well, as it includes, in equal measure, both struggles and solutions related to AI fairness. Our discussion of trade-offs spans contending with balancing fairness and accuracy in AI to setting the targets for fairness with human judgment. This paper, therefore, delves deep into the issues and considerations arising when fairness principles are attributed to clinical AI systems. The authors argue that the definition of fairness in AI has to be rethought in a clinical setting with its unique ethical, clinical, and social considerations. What is more important is to bring out clinical insights during AI development, which would assist in ensuring that developed fairness metrics are indeed health care scenarios applicable. The article emphasizes that translating AI fairness into workable clinical practice requires multidisciplinary collaboration between clinicians, AI researchers, and ethicists.

Cai, F., Zhang, J., & Zhang, L. (2024). "The Impact of Artificial Intelligence Replacing Humans in Making Human Resource Management Decisions on Fairness: A Case of Resume Screening." ***Sustainability*****, 16, 3840.**

This paper discusses how substituting AI for human judgment in human resources, particularly in curating resumes, affects applicants' fairness perceptions. A dual-scenario experimental design was used to compare human and AI procedural and distributive fairness. It found that AI decisions were perceived to be less fair when the decisions were adverse. The authors suggested a hybrid approach of AI and human decision-makers to moderate fairness concerns. This essential reference helps in understanding the effects of AI on fairness in HR practices. Also, it provides a model for studying and enhancing the perceptions of fairness in AI implementations.

Xivuri, K., & Twinomurinzi, H. (2023). "How AI Developers Can Assure Algorithmic Fairness," ***Discover Artificial Intelligence*****, 3:27.**

This paper illuminates the role that AI developers could play in encouraging algorithmic fairness, which is mainly through team diversity and the ethical integrity of the developers. According to the, the hurry in the delivery of results and the lack of diverse ideas in the team contribute to a significant part of the bias in the model. It identifies practical recommendations in the areas of governance, social, technical, and training, with special mention of what could a responsible organization do and the importance of transparency and ethics in the practices concerning the development of AI. This source will be invaluable for one to grasp the proactive measures that have to be taken to develop fair AI systems, underlining the fact that ethical considerations have to be integrated into the development process always.

I**slam, R.; Keya, K.N.; Pan, S.; Sarwate, A.D.; Foulds, J.R. (2023). "Differential Fairness: An Intersectional Framework for Fair AI,"** ***Entropy*****, 25:660.**

This paper contributes to the novel concept of "Differential Fairness," providing an intersectional lens for fairness in AI by marrying the aspects of differential valuable privacy for managing fairness across multiple protected attributes. Subsequently, the authors move on to further develop a solid theoretical framework and practical algorithms for assessing and implementing fairness within machine learning systems. They bring attention to the fact that these different systems of oppression cut across various dimensions of a person's identity, such as gender, race, and disability. The framework, illustrated by the many case studies and examples from its health care studies to its application in the criminal justice system, would detail how this can be achieved to reduce bias and engender equity in the AI system effectively.

4. Examples of Good Practice

IBM's AI Fairness 360 Toolkit: IBM has developed an open-source toolkit designed to help developers and researchers detect and mitigate bias in machine learning models throughout the AI application lifecycle. The AI Fairness 360 Toolkit contain the state-of-the-art latest algorithms for bias-existence checking in the models and data, along with its mitigation for a broad range of data types and models. The tool is compatible with several fairness definitions and explanations about the assessments with the realization that it is a must that firms in the AI industry have tool for looking towards implementing ethical AI practices.

Google's Pair Guidebook: The Google People + AI Research (PAIR) recently came out with this guidebook, which offers practical advice on how to design and use AI responsibly and humanely. Among other advice, it underscores the importance of developing diverse teams to test rigorously across multiple demographic groups as users of AI, as well as the need for perpetual, post-deployment monitoring on fairness and bias.

Accenture's Fairness Tool: Quantitatively assesses bias in AI algorithms and gives suggestions for adjustments to promote improved equity in the decision-making process—part of Accenture's broader commitment to trust and ethics in digital technology. This tool has already received

successful applications in finance and healthcare toward certain assurances that AI applications perform fairly across different populations.

Microsoft's Responsible AI Standard: Microsoft has devised an all-encompassing structure christened the Responsible AI Standard, which is meant to steer the development and use of AI systems compatible with ethical doctrines. The program consists of specific and detailed nature requirements for different dimensions, such as fairness, reliability, safety, privacy and security, inclusiveness and transparency, and accountability. Microsoft will audit regularly, document the life cycle of AI systems, and include many perspectives in developing AI. The standard will act as a benchmark, which drives the other companies towards implementing such comprehensive efforts.

Debias AI in Hiring at Unilever: Unilever has, through AI providers, put in place machine-learning solutions that eliminate human bias in the recruitment process. The company uses AI-powered games, video interviews, and more technologies to verify candidates based on their potential but not paper qualifications. The system is updated periodically to avoid reinforcing the status quo bias and further contributing to making the recruitment process more inclusive. This has made their workforce more diverse and reduced bias in hiring, representing strong illustrations of how AI can improve fairness in essential business operations.

AI Now Institute's Research and Recommendations: The AI Now Institute is an affiliate research center at New York University, and it focuses on the social effects that derive from the use of artificial intelligence. The group does interdisciplinary research to look at how artificial intelligence impacts society and, through that, develops detailed recommendations for policymakers, researchers, and companies to grapple with social problems, including bias in artificial intelligence. It emphasizes bringing ethical considerations into the development process of the AI system very early on and the need for robust regulatory frameworks on the use of AI.

Moral Machine Experiment at MIT: The Moral Machine is a website tool developed by research groups from MIT to collect human perspectives on decisions made by machine intelligence on moral dilemmas about autonomous driving. Evidence from such experiments informs the development of culturally relevant trends of ethical decision-making toward the realization of responsible and morally intelligent AI systems. These insights will help guide the behavior programming in artificial intelligence when it involves moral judgment, by which technologies are aligned to accord with human values and ethics.

Ada Lovelace Institute: A UK-based organization predicated on the belief that artificial intelligence—including AI—and other data technologies should operate in service of such a society, which is fair, just, and democratic. They work to conduct in-depth research and with policymakers to translate these learnings into workable solutions for the effective handling of bias in AI. These inclinations of expected ethical standards, which are predetermined and public, understanding, and responsive regulation, will be able to steer the ethical deployment of AI technologies.

5. Practical Recommendations

5.1. Enhance Data Diversity

One of the most essential steps in ensuring that bias is reduced to its minimum level or even being eliminated within AI systems is to guarantee that the training datasets are as diverse as possible. Data diversity should involve the inclusion of a wide range of data that is representative of the various demographic groups and scenarios that AI systems are likely to come across while being used in practical applications. Specific ways to enhance data diversity are as follows:

1）**Comprehensive Data Collection**: Companies should aim to collect data from as broad a source as possible to ensure that the dataset is as diverse as the whole population it wishes to cater to. This means that the collection will also contain data points from all the different demographic groups, including those from what are considered "minorities" or "under-represented," including women, older people, and disabled people.

2) **Assessment of Dataset Diversity**: Test the representational balance of datasets regularly to recognize significant imbalances or gaps that might exist. For instance, AI Fairness 360 by IBM is used for such purposes as the analysis of datasets related to representational fairness and bias. The test further helps understand whether there is an underrepresentation of certain groups or AI bias within the system.

3) **Synthetic Data Augmentation**: In cases when collecting real-world, diverse data is very challenging, synthetic data generation can prove to be a good solution. Techniques can be applied for creating balanced datasets, like the SMOTE method (Synthetic Minority Over-sampling Technique) or other advanced data simulation tools that allow for enhancing underrepresented categories in the training data and hence provide a more balanced view for AI to learn.

5.2. Promoting AI Ethical Audits

A proper ethical AI audit framework will help any organization identify, in a systematic manner, the underlying biases of an AI system and mitigate them appropriately. Such audits would fairly ensure the operation of AI —free of biases that result in discrimination. How companies can effectively implement such audits is described below:

1) **Develop an Audit Protocol**: Define clear protocols for executing an AI audit. This should include criteria for selection—such as whether the audit should consider all new deployments of AI or periodic reviews of deployed AI systems—specific fairness metrics that would be reviewed. Furthermore, stress test should be performed across different demographic groups.

2) **Audit Frequency**: Frequency for the audit shall be determined based on the application area of the highest impact and sensitivity. High-impacting systems such as hiring, health, or law enforcement likely require a higher audit frequency, maybe even bi-annual or annual, to ensure continued fairness.

3) **Use of Independent Auditors**: Consider delegating auditing processes to third-party auditors, who are entirely independent and have no conflict of interest in the outputs of the AI system. In this way, independent auditing may be regarded as giving an unbiased judgment on the AI systems, and thereby, the same process keeps the stakeholders and public informed of the extent of transparency maintained.

4) **Document and Act on Findings**: Ensure the documented audit findings are reported to relevant stakeholders, including how to deal with any noticed wrong or not in compliance with the required needs. Have a clear action plan for incorporating the findings of the audit back into the development and deployment process of the AI system.

5) **Regular Updates to Audit Practices**: AI is a dynamic field, as are the associated ethical standards. Update the auditing protocols regularly to incorporate new best practices and technologies to increase audit effectiveness in identifying and mitigating biases.

Implementing a comprehensive ethical AI audit framework will not only help in identifying biases but also build trust among users and regulatory bodies that the organization is committed to ethical AI practices. This proactive approach ensures that AI systems remain fair and effective throughout their lifecycle.

5.3. Continuous Monitoring and Updating

Further monitoring and updating of the AI systems is necessary over time. Societal norms and populations are evolving; therefore, the AI systems must be adjusted to remain unbiased and practical. The past two paragraphs have highlighted how ongoing vigilance and responsiveness can be structured as follows:

1) **Establish Monitoring Frameworks**: Establish a monitoring framework that continuously assesses the outputs of AI systems. This has to encompass real-time analytics, the capability to detect any deviation from the expected fairness benchmarks, and any emergent biases as these emerge. For example, this platform can use TensorFlow Model Analysis to test the performance of models on different demographics and in real-time.

2) **Feedback Loops for Updates**: Create feedback loops that provide end-users, as well as stakeholders, with the opportunity to comment on issues related to fairness or bias, the results of which are to be used in the updating of AI models and algorithms on a regular cadence. There should be a straightforward process for addressing the feedback on a timeline that includes when— if needed—a plan for model retraining or algorithmic updates should be made.

3) **Automated Retraining Policies**: Introduce automatic retraining policies whereby AI systems are retrained at regular intervals with new data that reflect changes in the demographics and society so that the models evolve in light of ongoing real-world distribution changes while maintaining relevance and fairness.

Subsequently, AI is actively monitored and updated by a dynamic organizational approach to avert the creation of AI systems within organizations that become obsolete in altered circumstances. A dynamic approach to managing AI spurs innovation; AI is ethical, and AI solutions continue to be fair and beneficial to all users over time.

6. Conclusion

This article provides an overview of the challenges and opportunities in ensuring technological fairness in the age of artificial intelligence (AI). As AI technology permeates various industries, ensuring the fairness of algorithms is not only a technical issue but also a critical matter concerning social justice and ethical values. The toolkit proposed in this article aims to guide technology companies in identifying and mitigating bias during the development and deployment of AI systems. It emphasizes the importance of algorithmic transparency, team diversity, and frameworks designed around fairness principles.

Through a review of existing literature, we recognize that the sources of algorithmic bias often stem from incomplete data and flawed assumptions in the models. Therefore, we advocate for enhancing data diversity, promoting AI ethics audits, and implementing continuous monitoring and updating of AI systems to create a fairer AI ecosystem. These measures will help companies improve the ethical application of their AI technology, enhance their credibility and moral standing, and increase the effectiveness of AI solutions in various markets.

Only by continuously optimizing AI systems to ensure they can adapt to the ever-changing social demands and ethical standards can we ensure that technological advancements continue to benefit societal development. This requires a dynamic management approach that includes ongoing monitoring, feedback, and update mechanisms to maintain the fairness and effectiveness of AI systems. The practical recommendations and toolkit proposed in this article provide a solid starting point for achieving this goal.

At the same time, we must recognize that AI fairness relies not only on technological improvements but also on the joint efforts of policymakers, researchers, AI developers, and ethicists. Through these collaborations, we can ensure that AI develops on a foundation of respect for human rights and social justice, achieving truly inclusive technological progress.

References

[1] Accenture. (n.d.). AI Fairness Tool. Responsible AI Governance Consulting & Solutions | Accenture.Retrieved form https://www.accenture.com/us-en/services/data-ai/responsible-ai.

[2] AI Now Institute. (n.d.). Research and Initiatives. Home - AI Now Institute. Retrieved from https://ainowinstitute.org.

[3] Bateni, A., Chan, M. C., & Eitel-Porter, R. (2022). AI fairness: from principles to practice. Accenture.

[4] Cai, F., Zhang, J., & Zhang, L. (2024). The impact of artificial intelligence replacing humans in making human resource management decisions on fairness: A case of resume screening. Sustainability, 16, 3840. https://doi.org/10.3390/su16093840.

[5] Google AI. (n.d.). People + AI Research (PAIR). Retrieved from https://pair.withgoogle.com.

[6] IBM Corporation. (n.d.). AI Fairness 360. Retrieved from https://aif360.res.ibm.com.

[7] Islam, R., Keya, K.N., Pan, S., Sarwate, A.D., & Foulds, J.R. (2023). Differential fairness: An intersectional framework for fair AI. Entropy, 25, 660. https://doi.org/10.3390/e25040660.

[8] Liu, M., Ning, Y., Teixayavong, S., Mertens, M., Xu, J., Ting, D. S. W., Cheng, L. T. E., Ong, J. C. L., Teo, Z. L., Tan, T. F., RaviChandran, N., Wang, F., Celi, L. A., Ong, M. E. H., & Liu, N. (2023). A translational perspective towards clinical AI fairness. npj Digital Medicine, 6:172. https://doi.org/10.1038/s41746-023-00918-4.

[9] Microsoft. (n.d.). Responsible AI Standard. Empowering responsible AI practices | Microsoft AI. Retrieved from https://www.microsoft.com/en-us/ai/responsible-ai.

[10] MIT Media Lab. (n.d.). Moral Machine. Retrieved from https://www.moralmachine.net.

[11] Robert, L. P., Pierce, C., Morris, L., Kim, S., & Alahmad, R. (2020). Designing fair AI for managing employees in organizations: A review, critique, and design agenda. Human–Computer Interaction. Accepted to the Special Issue on Unifying Human Computer Interaction and Artificial Intelligence.

[12] The Ada Lovelace Institute. (n.d.). About Us. Retrieved from https://www.adalovelaceinstitute.org/policy-briefing/ai-safety/.

[13] Unilever Global. (n.d.). Enhancing Livelihoods, Advancing Human Rights. Retrieved from https://www.unilever.com/sustainability/.

[14] Xivuri, K., & Twinomurinzi, H. (2023). How AI developers can assure algorithmic fairness. Discover Artificial Intelligence, 3:27. https://doi.org/10.1007/s44163-023-00074-4.