RESPONSIBLE AI, A PATHWAY TO FAIRNESS: A SYSTEMATIC LITERATURE REVIEW

Ibrahim Shaban Qabaqebo
University of London
15 July 2023

ABSTRACT

Artificial Intelligence (AI) promises profound societal changes, but concurrently poses significant challenges concerning fairness and responsible use. This systematic literature review explores four dimensions of responsible AI and fairness: data, algorithms, governance, and human-centric approach. Recent research is interrogated to uncover the state of the art in identifying and mitigating unfairness in AI systems and explored the prevailing challenges and opportunities. Issues of bias in large-scale training datasets, definitions of fairness, and the opaque nature of AI models were highlighted. Strategies to tackle these issues include dataset documentation, causal conceptions of fairness, and explainable AI (XAI). The need for robust governance mechanisms, including effective regulations and policies, was also underscored. A notable "governance gap" is identified and argued for the imperative of a dynamic, adaptable approach to AI governance that can keep pace with AI advancements. A human-centered approach to AI development is stressed to AI development, emphasizing the importance of diverse perspectives - from technical experts to policymakers - in shaping equitable AI systems. This review serves as an essential step towards understanding the landscape of fairness in AI, revealing pressing areas for future research.

Keywords: *Bias · Fairness · Responsible · Artificial Intelligence*

1. Introduction:

The rise of Artificial Intelligence (AI) has brought the field to the forefront of technological advancement and daily discourse. As noted by István Mezgár and Váncza (2022), AI is predicted to be the “biggest megatrend of the next decade.” This transformative power of AI, driven by big data, is further recognized by UNESCO (2018), which refers to it as the 'Fourth Industrial Revolution’.

Indeed, as stated by Yuval Noah Harari at the AI for Good Summit 2023, “this is the first tool in the human history that can make decisions by itself” (Summit 23 - AI for Good, 2023), AI systems have a wide range of applications in fields such as education, criminology, finance, and healthcare (Li and Zhang, 2023); However, the inherent complexity of AI and deep learning models poses a 'black box' problem, making the interpretation of their decisions quite challenging (Barredo Arrieta *et al.*, 2020; Ntoutsi *et al.*, 2020; István Mezgár and Váncza, 2022; Pagano *et al.*, 2023; Papagiannidis *et al.*, 2023) OpenAI (2023) also acknowledges the challenges surrounding the fairness of AI decisions, as they can exhibit biases related to gender, race, and age (Li and Zhang, 2023). These biases can evidently affect individuals and groups in their day-to-day lives.

Several research works have delved into Explainable Artificial Intelligence (XAI), responsible AI, fairness, bias, and ethics within AI systems. Yet, despite these considerable efforts, there remains a critical need for expanded research and work to ensure the deployment of trustworthy and fair AI systems. According to Barredo Arrieta *et al.*, (2020) there has been a significant rise in interest in XAI and responsible AI post 2018. In the recent open letter issued by the Future of Life Institute, technology leaders (including Elon Musk and Steve Wozniak), along with esteemed academics (like Yoshua Bengio and Stuart Russell) and others, advocate for a six-month "pause" on "giant AI experiments" (Baum *et al.*, 2023). This call-to-action comes from concerns over the potential adverse societal impacts of these experiments, including discrimination and unfair decision-making.

2. Method:

This systematic literature review revolves around four core dimensions fundamental to responsible AI and fairness: data, algorithms, governance, and human-centric (*Summit 23 - AI for Good*, 2023). The literature selected, primarily recent resources, including academic papers, preprints, blogs, books, podcasts, and key insights from the AI for Good Summit 2023.

This literature review delves into recent research on fairness within AI systems, aiming to address two pivotal inquiries:
Q1: How does recent research approach the identification and mitigation of unfairness in AI systems?
Q2: What challenges and opportunities exist in addressing unfairness within AI systems?

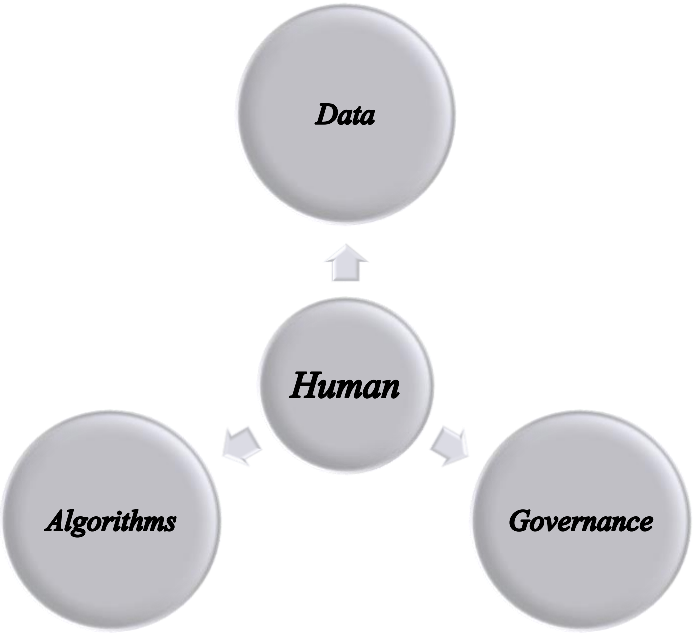

Figure 1 Four dimensions for Fairness in AI

To analyze and synthesize findings, a mapping and tagging system were employed. The themes tagged include the studies' problem, decision-making domain, datasets used, data issues, fairness definitions, fairness measures, ethical concerns, risks, biases, strategies for responsible AI, AI techniques, statistical approaches for bias avoidance and fairness, governance concept and recommendations, required regulations. This method provides a comprehensive and updated perspective on fairness in AI systems.

3. Results:

The fairness of data in AI and machine learning (ML) has become a significant area of concern in recent years. Many AI and ML models are trained on large-scale datasets, which may inherently carry various forms of biases, such as gender, racial, or socioeconomic biases (István Mezgár and Váncza, 2022; Lobel, 2022; Bias in AI and Machine Learning: Sources and Solutions - Lexalytics, 2022; Summit 23 - AI for Good, 2023; Li and Zhang, 2023; Pagano et al., 2023).

A case in point is the paper "Fairness of ChatGPT" by Li and Zhang (2023) presents an in-depth exploration of fairness and efficacy of Large Language Models (LLMs), using ChatGPT as the focus; Li and Zhang (2023) utilized four distinctive datasets known for their relevance in these areas: PISA (education), COMPAS (criminology), German Credit (finance), and heart disease (healthcare). Each of these datasets carry inherent biases, for instance, the COMPAS dataset has been criticized for its racial bias, often leading to higher risk scores for African American individuals, which can unfairly influence decisions on sentencing or parole in the criminal justice system. The study systematically explored group-level and individual-level fairness in the model's outputs and discovered significant biases. To combat these issues, the paper suggests an approach that includes data preprocessing, fair training objectives, and debiasing techniques such as adversarial training. This view on the importance of data preprocessing resonates with the findings from multiple other studies (Gebru *et al.*, 2021; Pushkarna, Zaldivar and Kjartansson, 2022; Weerts *et al.*, 2023; *Summit 23 - AI for Good*, 2023).

This view on the importance of data preprocessing echoes findings from other studies. Two research studies, Datasheets for Datasets (Gebru *et al.*, 2021) and Google Data Cards Playbook (Pushkarna, Zaldivar and Kjartansson, 2022), propose different methods for a comprehensive dataset documentation to enhance transparency, accountability, and fairness in ML models, ultimately promoting responsible AI deployment. Both are underpinned by the shared objective of preventing discriminatory outcomes, fostering diversity, and facilitating human oversight in AI development. Datasheets is a structured document that elicits information about the motivation, composition, collection process, preprocessing, labeling, distribution, and maintenance of a dataset. The datasheets aim to mitigate unwanted societal biases, facilitate greater reproducibility of ML results, and help researchers and practitioners select more appropriate datasets for their tasks (Gebru *et al.*, 2021). On the other hand, Google Data Cards Playbook aims to provide a clear and thorough understanding of a dataset's origins, development, intent, ethical considerations, and evolution, especially in people-facing contexts and high-risk domains (Pushkarna, Zaldivar and Kjartansson, 2022).

Transitioning to the subject of fairness definitions, Nilforoshan et al., (2022) ICML 2022 award-winning paper, "Causal Conceptions of Fairness and their Consequences", suggests that fairness definitions may need to differ across contexts. The authors introduce a two-part taxonomy for causal conceptions of fairness. The first set of definitions addresses counterfactual disparities, for example, ensuring that an algorithm recommends equal proportions of candidates from all race groups for college admission, considering those who would succeed if admitted. The second set seeks to limit the influences of one's group membership on decisions, implying that decisions should be impervious to factors such as race.

Amid these complications, the Fairlearn project offers practical solutions. They developed a comprehensive set of tools and resources aimed at helping practitioners assess and mitigate fairness-related harms in AI systems. Fairlearn's approach involves disaggregated evaluation and several methods for mitigating fairness-related harms, including pre-processing, in-training, and post-processing algorithms (Weerts et al., 2023). The project emphasizes that fairness is a sociotechnical challenge that demands careful evaluation in each unique context, thus reiterating the broader theme of the necessity for a nuanced understanding of fairness in machine learning.

Recognizing these complexities underlines the pivotal role of Explainable Artificial Intelligence (XAI). As Barredo Arrieta et al. (2020) indicate, there exists an intimate connection between fairness and XAI. XAI explainable artificial intelligence is important for fairness because it allows for the identification and mitigation of bias in ML models. As ML models are trained on historical data, often embedded with biases that might reflect in the model's predictions. These biases can result in unfair outcomes, such as discrimination against certain groups of people. For instance, in hiring scenarios, an XAI model could discern any biases in the process, such as a preference for candidates with certain demographics or backgrounds. By explaining how the model makes its decisions, XAI can help hiring managers ensure that their selection process is fair and unbiased Barredo Arrieta et al. (2020).

Nevertheless, the "black box" nature of neural networks and deep learning models often results in a lack of interpretability (Baum et al., 2023) These models, with their sheer complexity and potentially millions of parameters, make understanding their decision-making process a daunting task. This opacity can give rise to numerous issues, such as embedded bias in the data, challenges in identifying errors or biases within the model, and difficulties in pinpointing the origin of these errors or biases (Barredo Arrieta et al., 2020).

In her groundbreaking book, "The Equality Machine." Lobel, (2022) introduces several original concepts that force us to reconsider how we view and interact with machines and how to ensure equity in these systems “The quest for equality is a microcosm of all the struggles of humanity”. Additionally, in her paper “The Law of AI for Good” She proposes the idea of "automation right," suggesting that humans may have the right to access AI systems that can perform tasks more safely than people, for example, when self-driving technology will become statistically safer than human drivers (Lobel, 2023).

Lobel, (2023) also brings attention to the "Double Standard fallacy," This fallacy reflects the unrealistic expectation of 100 % reliability from AI, whereas we should instead consider if AI outperforms a human in the same role. Moreover, she highlights the "Ban/Freefall fallacy" indicating that an all-or-nothing approach towards AI governance is insufficient. Instead, a more nuanced strategy is needed to balance the benefits and potential risks posed by AI technology.

These insights underline the importance of AI governance, a recurring theme found across the resources analyzed for this review (Barredo Arrieta et al., 2020; Ntoutsi et al., 2020; Gebru et al., 2021; István Mezgár and Váncza, 2022; Lobel, 2022; Pushkarna, Zaldivar and Kjartansson, 2022; Baum et al., 2023; Pagano et al., 2023, 2023; Papagiannidis et al., 2023; Weerts et al., 2023; ‘Pause Giant AI Experiments: An Open Letter’, 2023). The necessity for effective AI governance is widely recognized and has become a key topic of discussion in global forums, such as the AI for Good Global Summit 2023 (Summit 23 - AI for Good, 2023).

At the summit, the United Nations (UN) Secretary-General António Guterres emphasized the significance of AI governance and regulation in his welcome speech (UN Secretary-General Welcome Speech for AI for GOOD Global Summit 2023, 2023). His address reinforced the idea that creating robust AI governance systems is a collaborative effort that should involve all stakeholders, further highlighting the collective responsibility we hold in steering the development and deployment of AI technologies.

The urgency and importance of AI governance is acknowledged across the bulk of the literature reviewed, but a dearth of concrete regulations and policies remains conspicuous. Papagiannidis et al., (2023) , in their paper "Toward AI Governance: Identifying Best Practices and Potential Barriers and Outcomes," broach this gap by offering some general recommendations. These encompass the need for companies to weave governance mechanisms into AI development processes, the assurance of transparency, explainability, and accountability of AI systems, and the drafting of explicit policies and guidelines for AI development and deployment. Furthermore, they emphasize the adherence to legal and ethical standards for AI applications and suggest the provision of training and awareness programs to inform and equip employees and stakeholders adequately.

While these recommendations serve as a good starting point, the role of policymakers in the furtherance of this field cannot be understated as emphasized by Nilforoshan et al., (2022). Similarly, as shown by the EU's first AI regulations efforts (EU AI Act: first regulation on artificial intelligence | News | European Parliament, 2023). However, This notion resonates with Gary Marcus's keynote speech at the AI for Good Summit 2023, wherein he introduces the term "governance gap." Marcus points out the considerable disconnect between our acknowledged need for AI governance and our understanding of the precise nature and components of such governance.

The governance gap also refers to the divergence between the relatively slow pace at which governments operate and the breakneck speed of AI developments and their adoption. As such, the challenge lies in cultivating an adaptable, dynamic, and expedient approach to governance that can keep pace with the rapidly evolving landscape of AI while ensuring that ethical, legal, and societal concerns are adequately addressed. The challenge is immense, but so too are the potential rewards if it can be effectively met. AI governance is not merely a constraint on AI's potential; it's an essential foundation for achieving that potential in a sustainable and beneficial manner, with a focus on human values.

The human-centric aspect of AI is emphasized by Timnit Gebru in her discussion on fairness in AI (AI Rewind 2019: Trends in Fairness and AI Ethics with Timnit Gebru | The TWIML AI Podcast, 2022) . Gebru draws attention to the rising communities like Black in AI, {Dis}Ability in AI, LatinX in AI (LXAI), Queer in AI, and Women in AI (#WAI) (Black in AI; {Dis}Ability in AI; LXAI; Queer in AI; Women in AI (#WAI)), which work to ensure the inclusion and representation of diverse individuals in AI development and policy-making. Such initiatives reiterate the necessity of embedding human values and diverse perspectives in AI, thereby making the systems equitable and beneficial for all.

Most of the papers advise that the development of responsible AI requires a human-centered approach that considers the perspectives and human values of diverse stakeholders.

In conclusion, the objective is to achieve balance creating AI systems that are high performing yet ethically sound, effectively mirroring the diversity and complexity of the human experience. This endeavor calls for an intricate blend of technical acuity, ethical considerations, and active inclusion of diverse perspectives in AI development and policymaking. By meeting this challenge, we can steer the development and deployment of AI technologies in a direction that is beneficial and equitable for all.

4. Conclusion and Discussion:

In conclusion, the reviewed literature provides insightful exploration into the critical topics of fairness, bias, and governance in the realm of artificial intelligence. A common thread running through these studies is the recognition of the complex challenges posed by these issues, particularly in an era marked by rapid technological advances.

Taken together, these studies represent a snapshot of a field in evolution, grappling with complex ethical and governance issues. While the papers highlight the existing limitations and challenges, they also underscore the scope and potential for future work in creating a more responsible and equitable AI landscape.

These areas are as follows:

- **Data Bias:** A major gap lies in the development of more comprehensive methods for identifying and mitigating bias in AI systems. The complexity and contextual variability of bias demand robust solutions, which will require advanced preprocessing techniques and fairness-centric training strategies.
- **Algorithmic Fairness:** The field lacks a clear, universally agreed-upon definition of fairness, which leads to varied implications across different contexts. Future research should focus on conceptual and theoretical work to establish robust fairness metrics and mitigation strategies for AI and ML.
- **Governance:** A significant "governance gap" exists in AI. Despite widespread acknowledgment of its importance, concrete regulations and policies are yet to be established. There is a pressing need to develop practical, adaptable governance frameworks that can align with AI's rapid progress while safeguarding societal, ethical, and legal values.
- **Diversity and Inclusion:** While several human-centric initiatives have been launched to foster diversity in AI, the field needs to be more proactive in integrating diverse perspectives in AI development and policymaking. Future work should focus on ensuring AI technologies reflect human diversity and promote equity in their benefits.
- **Awareness and training:** There's a pronounced knowledge gap concerning AI's ethical aspects and fairness, not only among developers, but also among public and policymakers. This highlights the pressing need for increased AI literacy with a specific emphasis on fairness issues.

As AI continues to evolve and permeate various aspects of society, these areas represent critical points of attention to ensure that its development and deployment are fair, ethical, and beneficial to all.

Overall word count: 2772 words
Count word: 2430 without parentheses ()

5. Acknowledgements:

I express my deepest appreciation to Professor Larisa Soldatova, our module leader, for her guidance. Equally, I'm thankful to the dedicated review team for their insightful critiques that enriched this project. My friends, Fadi Ibrahim, Dr. Muhamad Saloum, Sally Al-Edilbi, and Antonio Parente, deserve special mention for their steadfast support, feedback, and endless patience. Similarly, my mentor, Dr. Andrea Orazio Spinello, PhD, has been instrumental, guiding me with wisdom and motivation throughout this journey.

Finally, my profound thanks to my supervisor, Thomas Georgi. His unwavering belief in my mission for responsible AI has fueled my passion and commitment to this project.

In sum, your collective influence has been invaluable and I'm immensely grateful for your contributions to my growth and this research.

6. References:

AI Rewind 2019: Trends in Fairness and AI Ethics with Timnit Gebru | The TWIML AI Podcast (2022) *TWIML*. Available at: https://twimlai.com/podcast/twimlai/ai-rewind-2019-trends-fairness-ai-ethics-timnit-gebru/ (Accessed: 16 July 2023).

Barredo Arrieta, A. *et al.* (2020) ‘Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI’, *Information Fusion*, 58, pp. 82–115. Available at: https://doi.org/10.1016/j.inffus.2019.12.012.

Baum, K. *et al.* (2023) ‘From fear to action: AI governance and opportunities for all’, *Frontiers in Computer Science*, 5. Available at: https://www.frontiersin.org/articles/10.3389/fcomp.2023.1210421 (Accessed: 13 July 2023).

Bias in AI and Machine Learning: Sources and Solutions - Lexalytics (2022). Available at: https://www.lexalytics.com/blog/bias-in-ai-machine-learning/ (Accessed: 12 July 2023).

Black in AI (no date). Available at: https://blackinai.github.io/#/ (Accessed: 16 July 2023).

{Dis}Ability in AI (no date). Available at: https://elesa.github.io/ability_in_AI/ (Accessed: 16 July 2023).

EU AI Act: first regulation on artificial intelligence | News | European Parliament (2023). Available at: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (Accessed: 16 July 2023).

Gebru, T. *et al.* (2021) ‘Datasheets for Datasets’. arXiv. Available at: https://doi.org/10.48550/arXiv.1803.09010.

‘gpt-4.pdf’ (no date). Available at: https://cdn.openai.com/papers/gpt-4.pdf (Accessed: 13 July 2023).

https://plus.google.com/+UNESCO (2018) *The Fourth Revolution*, *UNESCO*. Available at: https://en.unesco.org/courier/2018-3/fourth-revolution (Accessed: 13 July 2023).

István Mezgár and Váncza, J. (2022) ‘From ethics to standards – A path via responsible AI to cyber-physical production systems’, *Annual Reviews in Control* [Preprint]. Available at: https://doi.org/10.1016/j.arcontrol.2022.04.002.

Li, Y. and Zhang, Y. (2023) ‘Fairness of ChatGPT’. arXiv. Available at: http://arxiv.org/abs/2305.18569 (Accessed: 10 July 2023).

Lobel, O. (2022) *The Equality Machine: Harnessing Digital Technology for a Brighter, More Inclusive Future*. New York: PublicAffairs.

Lobel, O. (2023) ‘The Law of AI for Good’. Rochester, NY. Available at: https://doi.org/10.2139/ssrn.4338862.

LXAI (no date) *LXAI*. Available at: https://www.latinxinai.org (Accessed: 16 July 2023).

Nilforoshan, H. *et al.* (2022) ‘Causal Conceptions of Fairness and their Consequences’. arXiv. Available at: https://doi.org/10.48550/arXiv.2207.05302.

Ntoutsi, E. *et al.* (2020) ‘Bias in data-driven artificial intelligence systems—An introductory survey’, *WIREs Data Mining and Knowledge Discovery*, 10(3), p. e1356. Available at: https://doi.org/10.1002/widm.1356.

Pagano, T.P. *et al.* (2023) ‘Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods’, *Big Data and Cognitive Computing*, 7(1), p. 15. Available at: https://doi.org/10.3390/bdcc7010015.

Papagiannidis, E. *et al.* (2023) ‘Toward AI Governance: Identifying Best Practices and Potential Barriers and Outcomes’, *Information Systems Frontiers*, 25(1), pp. 123–141. Available at: https://doi.org/10.1007/s10796-022-10251-y.

‘Pause Giant AI Experiments: An Open Letter’ (no date) *Future of Life Institute*. Available at: https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (Accessed: 13 July 2023).

Pushkarna, M., Zaldivar, A. and Kjartansson, O. (2022) ‘Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI’, in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery (FAccT ’22), pp. 1776–1826. Available at: https://doi.org/10.1145/3531146.3533231.

Queer in AI (no date) *Queer in AI*. Available at: https://www.queerinai.com (Accessed: 16 July 2023).

Summit 23 - AI for Good (2023). Available at: https://aiforgood.itu.int/summit23/ (Accessed: 12 July 2023).

UN Secretary-General Welcome Speech for AI for GOOD Global Summit 2023 (2023). Available at: https://www.youtube.com/watch?v=jfnNq2Q2BGQ (Accessed: 15 July 2023).

Weerts, H. *et al.* (2023) ‘Fairlearn: Assessing and Improving Fairness of AI Systems’. arXiv. Available at: http://arxiv.org/abs/2303.16626 (Accessed: 13 July 2023).

Women in AI (#WAI) (no date) *Women in AI (WAI)*. Available at: https://www.womeninai.co (Accessed: 16 July 2023).

Annex:

Glossary of Technical Terms:

1. **Artificial Intelligence (AI):** A branch of computer science that aims to create systems capable of performing tasks that usually require human intelligence.
2. **Algorithm:** A set of rules or instructions given to an AI, or a computer, to help it learn on its own.
3. **Bias:** Systematic error introduced by the design of the model or the data used in training the model.
4. **Fairness:** In the context of AI, fairness refers to how equally or justly a model treats different groups of individuals. There are several definitions of fairness, and it often needs to be defined in the context of a specific task.
5. **Data:** The information used to train AI models. The quality and quantity of the data can significantly impact the performance of the model.
6. **Governance: In** the context of AI, governance refers to the rules, policies, and regulations that guide how AI systems are developed, used, and monitored.
7. **Explainable AI (XAI):** An area of AI focused on creating systems that provide clear, understandable explanations for their actions and decisions.
8. **Human-Centric Approach:** An approach to AI development that emphasizes the needs, safety, and values of humans.
9. **Training Datasets:** Large sets of data used to teach AI systems how to perform certain tasks or make predictions.
10. **Dataset Documentation:** Detailed information about a dataset's creation, composition, and use to help understand and interpret the output from an AI model.
11. **Causal Conceptions of Fairness:** An approach to fairness that takes into consideration the underlying causal relationships between variables.
12. **AI Advancements:** Refers to the ongoing development and innovation in AI technologies.