Mitigating Bias in Artificial Intelligence: Methods and Challenges

Saja Salim Mohammed[1], Israa Alsaadi[2] and Hind Ibrahim[1,2], Sarah Ali Abdulkareem[1] and Hasinah Maizan[3]
[1]*University of Diyala, 32001 Diyala, Iraq*
[2]*University of Baghdad, 10071 Baghdad, Iraq*
[3]*Universiti Kebangsaan Malaysia, 43600 Selangor, Malaysia*
saja.salim.edu.iq, israamh_comp@csw.uobaghdad.edu.iq, hindim@uodiyala.edu.iq, sarah.ali@uodiyala.edu.iq, P121110@siswa.ukm.edu.my

Keywords: Bias Mitigation In AI, Algorithmic Fairness, AI Ethics, Fairness Metrics, Data Bias.

Abstract: The extensive application of Artificial Intelligence (AI) across the core domains of society has brought forth massive challenges towards prejudice, embedding discrimination, feeding inequalities, and eroding trust among citizens. This report explores the multi-dimensioned aspect of AI systems' prejudice by understanding the causes of the phenomenon in terms of data, algorithms, and end-user interface and also exploring its social implications and normative concerns. We give a comprehensive overview of existing state-of-the-art bias detection methods, i.e., statistical approaches, explainability tools, and fairness measures, and discuss mitigation techniques in pre-processing, in-processing, and post-processing. Challenges persist, such as negative fairness-accuracy trade-offs, limited standardized benchmarks, and need for inter-disciplinary efforts. Through case studies and regulatory analysis, we determine best practices and novel frameworks that will propel fair AI. The paper concludes by offering the directions of future research, emphasizing the necessity of open, transparent, accountable, and inclusive approaches to prevent AI systems from deviating from moral principles and societal values.

1 INTRODUCTION

Artificial Intelligence (AI) is beginning to pervade all aspects of society. From chatbots that help answer questions, to cars that drive themselves, and to lucrative algorithms that decide one's credit score, AI technologies are becoming unavoidable. AI is both very good at solving very specific problems when fed the right data and also very complex. Some contemporary AI systems such as deep neural networks or certain recommender systems can be beyond anyone's full understanding, including their creators. Naturally, many of the consequences of rapidly adopting such complex technologies are unknown. One critical, known drawback is the surprise realization that AI is not quite as rational as previously believed, and that it is open to embedding various societal recognized biases as exhibited by phenomena such as the Facebook ad-stereotype scandal, COMPAS crime prediction tool, the YouTube-optimized recommendation system and many others [1], [2], [3], [4]. Therefore, this prejudice of baked-in biases in AI systems has urged for the development of a field of AI fairness to quantify and mitigate such biases, and thus guaranteeing the ethical deployment and development of AI technologies. Just the right balance of foreshadowing of societal responsibilities and new technological questions is what makes embedded bias in AI such a compelling issue of inquiry for any researcher or practitioner in the AI field [5], [6], [7].

Recent high-profile cases have made it necessary to address AI bias. For instance, gender bias has been observed to happen in AI hiring tools, racial bias in facial recognition systems and recidivism risk assessment tools used in the criminal justice system. [8], [9]. This essay is dedicated to the discussion of bias in artificial intelligence, thereby highlighting different ways in which this bias can manifest. First, breastfeeding machines will be used as an illustration of how training data can embed and perpetuate gendered, cultural or social biases The narrative will then shift to biases in the organization of AI fairness research effort, namely the lack of regulations to incentivize technological solutions not just to individual fairness, and the untackled problems that might be discounting the potential of fairness as a concept to reshape certain societal dynamics. The

paper will go ahead and cover modern-day mitigation actions, including debiasing of datasets [10], fair algorithms construction [11] , and bringing ethics into consideration when constructing AI [12]. However, mitigating bias in AI is an intricate issue. It calls for evading technical, ethical, and societal challenges [13]. This article will also observe the issues with bias reduction, such as not being able to define and measure fairness precisely in different scenarios, the risk of trade-offs between model performance and fairness [14], and the challenge in addressing intersectional biases [15]. By providing a broad survey of the methods and challenges of mitigating AI bias, this paper seeks to contribute to the ongoing discussion on the design of more equitable and accountable AI systems. In addition to advancing technical understanding of bias reduction, our aim is to emphasize the need for an interdisciplinary strategy to consider the broader social impacts of AI deployment. As AI continues to revolutionize our world, the imperative to confront and mitigate bias becomes increasingly urgent. This paper aims to be an information source for researchers, practitioners, and policymakers in the pursuit of developing just and unbiased AI systems that can benefit society at large.

2 UNDERSTANDING BIAS IN ARTIFICIAL INTELLIGENCE

Bias in AI can be either intentional (explicit bias) or non-intentional (implicit bias) making it very difficult to, in many cases, differentiate bias from expected behavior [16], [17]. There are two primary considerations of the bias that is pertinent to understanding bias in an AI system. First, bias should be considered as a mirroring of a societal bias in an AI application, a techno-social mirror. In such a context, research on the bias of AI raises questions on how to understand and approach biases in society, and how to ask an AI application to address bias. Second, concerning the error of AI decision-making, it is important to recognize the origin of the bias as a feature in the design or data, and to take action to alleviate it systematically [18], [19]. To address bias in the design and use of AI applications, an understanding of the bias in a technological context is required. With this in mind, bias in an AI system would be framed such that it can provide useful conceptual insights for a community of researchers tormented by the implications of bias within this technology [20]. Table 1 shows the primary metrics of fairness [21].

Table 1: Primary metrics of fairness [21].

Fairness Metric	Summary
Statistical or demographic parity	Requires equal probability of positive predicted class across protected and unprotected groups
Disparate impact	Represents the ratio of the rate of positive prediction between protected and unprotected groups
Calibration	Requires similarity between probability prediction or risk scores and actual outcomes regardless of group
Predictive parity	Requires equal positive predictive values across protected and unprotected groups
Error rate	Represents the ratio of incorrect to total predictions compared among protected and unprotected groups
Equal opportunity	Requires that a preferred outcome is predicted equally across protected and unprotected groups
Equalized odds	Requires equal true-positive and false-positive rates between protected and unprotected groups

This is at the center of developing unbiased AI models. It is necessary to recognize the truth that bias may be unconsciously embedded in AI systems through training by way of the data on which it is trained. Data collection and data selection have much to do with how the presence of bias in AI systems can be determined. It is required to discover sources of bias and also potential impacts on AI decision-making in order to combat bias in AI systems [22]. Bias can be ingrained in various phases of the life cycle of AI development. As an example, there may be bias in training data that has been used for training AI systems. It will result in biased decision-making from AI systems whenever new data is fed into them. Overcoming and recognizing such biases is essential in the development of fair and ethical AI systems. The ways in which different sources of bias can impact AI systems need to be explored [23].

3 TYPES OF BIAS IN AI

Researchers have demonstrated the widespread and nuanced spread of technological bias. Biases inherently determine the quality of the predictions of all artificially intelligent systems, as long as they are based on learning algorithms and data sets [24]. The quality of the predictions of ML models depends strongly on the quality of the training data sets, so data must be of high quality and unbiased. The possibility of considering a system as unbiased is not tolerated, but a data set is considered unbiased if it trains models that predict the same quality, regardless of the input and output values included [25]. Table 2 enumerates the types of bias, providing a descriptive definition of each type together with an example.

Bias can have a wide variety of aspects and possible formations. In order to catch these different types, machine learning is described here for the types of bias that can occur with applications and best practices. Each of these types of bias is very different: they may have different causal effects and have different possibilities for countermeasures [26]. It shows that the quality of the data affects the performance of the model differently than the model fails to generalize in different ways and is only possible with specialized countermeasures. Categorized biases may have a direct bear model that does not match the quality predictions of the models on certain input values [27]. At the same time, the model that produced the data set on the other hand, the causal effects of the bias on the predictions of the submitted model may differ from the properties of the submitted model [28], [29].

In order to obtain models that make sound decisions based on their observations and adhere to moral norms, it will be necessary to obtain a model that predicts generalizes and is fair [30]. To that end, it is necessary to better understand the bias that the model makes predictable, and to learn countermeasures to prevent the successful general model of the bias modeling. Bias in AI and ML are beginning to find an increasing reassurance in a wide representation of the people and institutions [31].

Table 2: Common types of bias in ai systems.

	Type of Bias	Description	Example
1	Data Bias	Occurs when the training data is not representative of the population or use case	Facial recognition systems trained primarily on light-skinned faces performing poorly on dark-skinned faces
2	Algorithmic Bias	Stems from the choices made in designing and implementing AI algorithms	A resume screening algorithm favoring certain keywords associated with one gender over another
3	Interaction Bias	Arises from the way users interact with AI systems	Voice assistants struggling with accents or dialects not well-represented in training data
4	Historical Bias	Reflects past societal biases present in the data used to train AI	An AI-based hiring tool reflecting historical gender imbalances in certain professions
5	Sampling Bias	Results from non-random sampling of subgroups in a population	A medical diagnosis AI trained mostly on data from urban hospitals may perform poorly for rural patients
6	Confirmation Bias	The tendency to search for or interpret information in a way that confirms pre-existing beliefs	An AI content recommendation system amplifying users' existing views without presenting diverse perspectives
7	Automation Bias	The propensity to favor suggestions from automated decision-making systems	Over-reliance on AI-generated results without critical evaluation by human experts
8	Reporting Bias	Occurs when the frequency of events, properties, or outcomes in a dataset doesn't reflect real-world probabilities	An AI trained on news articles might overestimate the frequency of rare but newsworthy events

The possibility that widespread and understandable mechanisms are widely engaged in the data processing applications only increases this interest and the red flags continue to multiply. By and large, research and knowledge on bias in AI belong to a variety of fields and theoretical traditions: media scholarship, policy reports, sociology, artificial intelligence perspectives, computer science, commentary on algorithmic discrimination, etc. At the same time, the most extensive conversations on the topic tend to happen within disciplinary silos [32].

4 IMPACTS OF BIAS IN AI SYSTEMS

Public concern over AI bias and its societal and ethical consequences has grown in recent years. Many machine learning models inherently amplify discrimination, often reinforcing systemic biases, particularly in critical sectors like finance, healthcare, and law enforcement. The complexity of AI decision-making makes bias difficult to trace and mitigate, disproportionately impacting vulnerable groups [33], [34]. Moreover, the opacity of these systems erodes public trust, with fears of a global AI arms race compromising safety and accountability. Addressing these challenges requires interdisciplinary collaboration to ensure fairness, transparency, and ethical AI development [35].

5 ETHICAL IMPLICATIONS OF BIAS IN AI

Bias in AI systems has generated significant attention from the public and key stakeholders, as it may lead to various forms of discrimination. Indeed, there have been reported cases of AI bias in areas such as crime prediction, employment, and online advertising. A broadly defined bias is present whenever the modeled output deviates systematically from the desired output provided by experience, and one possible effect of bias is discrimination [36]. Despite the growing interest, managing bias in AI can be a challenging task. The causes of bias are diverse, and there is no universal way in which they can be examined in the AI systems [37]. Moreover, addressing bias and discrimination in AI systems is a cross-disciplinary challenge, and engaging with technical, legal, social, and ethical issues constitutes a complex endeavor. To address bias successfully, a better understanding of AI behavior and societal impact is needed. AI experts should work in unison with policy-makers, ethicists, and social scientists to raise public awareness, recognize the challenges faced by a variety of stakeholders, and develop special policies and programming ensuring fair AI [38].

6 METHODS FOR DETECTING BIAS IN AI

The growing consensus among policymakers, industry leaders, and the broader community highlights the urgency of addressing bias in AI. This has led to a surge in detection methods and fairness-enhancing techniques, though discussions on bias remain fragmented, particularly in evaluating these methods systematically. A rigorous assessment is essential to distinguish effective solutions from those needing refinement [39], [40]. Identifying the most important factors that contribute to bias is essential to understanding why bias occurs in AI systems, as illustrated in Figure 1.

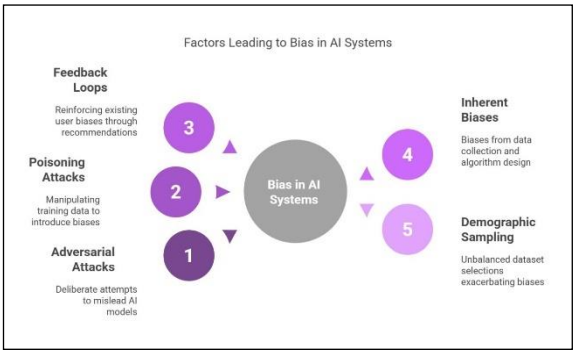

Figure 1: Factors leading to bias in AI systems.

6.1 Statistical Methods

This subsection explores statistical methods for detecting bias in AI, focusing on quantitative techniques to measure fairness. It examines how scholars and practitioners implement these methods, including metric selection, statistical modeling, and dataset choices. A framework of techniques is outlined, with case studies highlighting common pitfalls and the need for a more robust statistical approach. Visualizations are recommended to enhance clarity, along with guidelines for ethical and transparent statistical reporting [41], [42].

Understanding how bias manifests is the first step toward mitigation. Statistical methods, such as disparate impact analysis, help quantify fairness by

assessing whether protected groups face adverse outcomes. The severity of bias is often more relevant than its mere presence, requiring careful metric selection and comparison to baselines. Reporting should go beyond binary significance tests, incorporating intersectional analysis to avoid oversimplification. Combining visualizations with traditional statistical methods can improve communication and drive more informed decision-making [43], [44].

6.2 Explainability and Interpretability Techniques

AI systems are often opaque due to their complexity, making bias difficult to detect and correct. While this complexity has driven innovation, it also raises concerns about whether bias in AI reflects real-world patterns or stems from flawed training data and design choices. Without explainability, even experts struggle to assess an AI system's fairness and reliability [45].

Explainability and interpretability techniques are essential for understanding and controlling AI models. Research shows that transparency fosters trust and enables proactive bias detection. Various methods, such as permutation feature importance, local interpretable model-agnostic explanations, and Shapley values, summarize model behavior without directly replicating it. However, these explanations offer different levels of insight and may not fully capture a model's inner workings. Therefore, efforts should focus on building trust in AI decision-making rather than exhaustive analysis of every model detail [46], [47].

7 MITIGATION STRATEGIES FOR BIAS IN AI

The rise of AI has driven innovation across various fields, but its deployment in high-stakes areas like criminal justice and healthcare raises concerns about fairness and bias. AI systems can reinforce societal prejudices when trained on skewed datasets, leading to unfair or harmful outcomes, particularly for marginalized groups. This issue extends beyond faulty algorithms – bias can emerge even in properly functioning systems, amplifying existing inequalities [48].

Growing awareness of AI-induced harm has spurred research into bias detection and mitigation, particularly in models using electronic health record (EHR) data. This scrutiny extends beyond bias and discrimination to broader societal impacts, with increasing academic attention on AI's role in both public and private sectors. Understanding these consequences is essential as AI continues to shape critical decision-making processes [49].

7.1 Pre-Processing Techniques

Mitigating bias before training is often preferable to addressing it afterward. AI has the potential to drive innovation, but its deployment in social sectors demands rigorous fairness testing [50]. Bias arises when model predictions disproportionately benefit or harm certain groups, as seen in cases like gender bias in Newsela's quizzes. While eliminating bias is challenging, addressing it is essential to prevent harm [51]. Bias often originates in training data, where historically marginalized groups are underrepresented, increasing their risk of being affected by biased models. Datasets compiled from multiple sources and perspectives add complexity, making it difficult to quantify bias precisely. Research highlights the need to examine data, power dynamics, and bias to develop fairer AI systems [52].

7.2 In-Processing Approaches

Mitigating bias in AI across data collection, model development, deployment, and decision-making is crucial. In-processing approaches, particularly useful for off-the-shelf AI or when end-users are not developers, focus on real-time interventions such as filtering biased inputs and adjusting outputs. Two key strategies include incorporating fairness constraints and using regularization techniques during training to prevent data-driven biases from shaping decisions. Fairness constraints ensure that group differences do not influence model predictions, while regularization actively steers the training process toward fairer outcomes [53]. Various algorithms integrate fairness principles into AI models, though they differ in their conceptual approaches. Bias mitigation must extend beyond modeling to every stage of an AI system's lifecycle, requiring collaboration across stakeholders to understand how bias emerges and propagates. To effectively address bias, it is essential to rigorously define it, assess its impact, and establish monitoring frameworks. A decoupled model monitoring approach, centered on AI bias audits, is proposed to enhance transparency and accountability throughout an AI system's development and use [54].

7.3 Post-Processing Methods

Post-processing adjusts model outputs to enhance fairness without requiring retraining. Techniques such as re-weighting predictions, ensuring equal error rates, and calibration help balance outcomes across demographic groups. While these methods improve fairness, continuous post-deployment monitoring remains essential, as bias is inherent in many real-world problems, and even the best interventions involve trade-offs [55], [56]. Research on post-deployment fairness monitoring has grown, yet practical implementation remains underexplored. Figure 2 highlights the importance of addressing bias at all stages of machine learning, from problem definition through data collection and processing to model development, testing, and deployment. Ignoring these biases can lead to unfair or inaccurate models, which can impact the decisions based on them [57].

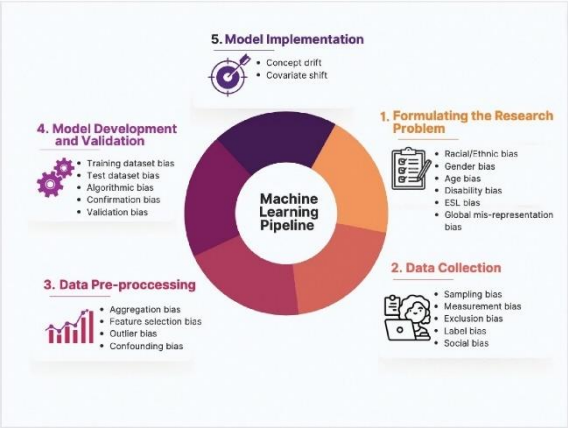

Figure 2: Sources of bias that may contribute to predictive variances of AI algorithms.

8 CHALLENGES IN IMPLEMENTING BIAS MITIGATION TECHNIQUES

Mitigating bias in AI and ML systems is a multi-pronged process that is plagued by several barriers at various stages of model development and deployment. These can be categorized under technical, ethical, organizational, and regulatory barriers [58], [59].

A) Technical Challenges:
1) Definition and Quantification of Bias: Bias is context-specific, and its meaning varies across domains and applications, and therefore difficult to standardize;
2) Data Limitations: bias generally results from imbalance or non-representative data, and it is difficult to get high-quality, unbiased data;
3) Algorithmic Complexity: mitigation of bias through fairness constraints or adversarial debiasing can increase the risk of model accuracy degradation or unintended side effects;
4) Scalability: bias mitigation is challenging and expensive on these large AI systems with dynamic, real-time data streams.

B) Ethical and Societal Challenges:
1) Trade-offs Between Fairness and Performance: balancing model performance and fairness may be contentious, as different stakeholders may have conflicting objectives;
2) Differing Notions of Fairness: different notions of fairness (e.g., demographic parity, equalized odds) may be incompatible with each other, so it is not sure which one would take priority;
3) Unintended Consequences: some mitigation techniques may introduce new biases or reduce transparency in AI decision-making.

C) Organizational Challenges
1) Lack of Incentives: businesses prefer performance and profitability to equity, deterring investment in attempting to combat bias;
2) Limited Expertise: companies do not possess the necessary expertise in fairness-aware ML techniques, and hence it is difficult to adopt them;
3) Resistance to Change: managers and programmers may resist altering existing AI systems since they perceive risks and costs involved in model changes.

D) Regulatory and Policy Challenges:
1) Lack of Uniform Regulations: regulatory frameworks for AI fairness are yet to be developed, leading to disparity in measures for bias evasion;
2) Challenging Enforcement: adherence to fairness standards is difficult to facilitate through the execution of robust auditing mechanisms, which is difficult for most organizations;
3) Variability Globally: cross-border AI systems experience complexity in bias mitigation with varying expectations on fairness across nations and sectors.

9 CASE STUDIES AND BEST PRACTICES

Best practices and case studies present information on how bias reduction in artificial intelligence is put into practice in the real world. Best practices and case studies are practical sources of information on how AI systems can be made transparent and fair by avoiding bias in them [60]. Case studies indicate successful strategies on bias mitigation in AI systems. They provide valuable information on the best practice of avoiding bias in AI. Case studies and real examples can offer realistic strategies and methods for identifying and minimizing bias in AI systems. For example, a case study of a facial over-representation system of some racial groups as errors can offer lessons about potential biases and how they can be resolved [61], [62]. In addition, the best approach to integrating ethical guidelines while designing and implementing AI can offer good lessons in minimizing bias. For example, observing how Google and Microsoft have managed bias in their AI software can demonstrate to one how to best keep bias at bay [63]. In the same way, observing case studies of how bias reduction in the application of AI in healthcare and finance can demonstrate to one lesson in how to implement best practice. For example, how bias has been addressed in AI algorithms for patient diagnosis and treatment recommendations in medicine can be utilized to establish useful bias mitigation methodologies [64]. By the introduction of an algorithmic-experimental combination of identifying, measuring, and suppressing the butterfly effect in AI systems the obstacles of bias are transcended to achieve fairness in outcomes and the emergence of responsible AI [65].

10 CASE STUDIES AND BEST PRACTICES

Regulatory frameworks and guidelines play an important role in preventing bias in artificial intelligence through standardization and specification of requirements and standards for transparency and fairness in AI systems. Regulatory frameworks and guidelines offer a context for detecting and solving potential biases in AI models and algorithms [66]. Regulatory frameworks and guidelines are important in upholding fairness and accountability in AI systems. Regulatory frameworks and guidelines offer a set of rules and standards to be abided by developers and users. These standards are central to the provision of the ethical and ethical use of AI technology. They offer best practices that corporations and organizations can adhere to in an effort to avoid causing harm to people or to society as a whole. These standards tend to entail recommendations for transparency, accountability, and fairness in AI systems [67]. Regulatory standards can also be used to apply the issues of data privacy and security to AI systems. For example, standards can include data encryption and storage protocol requirements [68].

11 FUTURE DIRECTIONS IN BIAS MITIGATION RESEARCH

There are numerous possibilities for future research on bias mitigation in the years to come. As new AI models are developed and applications change, it is imperative to continually innovate. New problems will arise, and existing issues will evolve. To meet these challenges, an interdisciplinary approach involving researchers, practitioners, and policymakers in a variety of areas will be necessary. Collaboration between technical experts, lawmakers, and domain professionals will be crucial in driving advancement. While some strategies can be broadly applied, each specific context may present its own unique challenges requiring bespoke solutions. Collaboration between computer scientists, social scientists, ethicists, and legal experts is well-suited to addressing these multifaceted difficulties [69].

There is an abundance of technologies which can be brought to bear on this complex issue. Natural language processing, deep learning, transfer learning, and reinforcement learning all present opportunities to develop improved methods of bias mitigation. However, it is important that the drive to utilize these advanced techniques does not come at the expense of interpretability and auditability. Other technologies could be used to explain model predictions and biases, thereby offering insight into how to generate more effective de-biasing strategies [70]. As the use of AI in decision-making processes becomes more prevalent, the availability of high quality, relevant data is increasing. There is potential to leverage this data to identify latent biases and develop strategies to address them. Changes in recruitment and hiring procedures, legislation regarding facial analysis technology, and concerns about profit-maximising businesses have generated momentum for using data-driven strategies to enhance the fairness of AI outcomes. Because of this, many novel methods have

been created over the past several years [71]. There are many analysis tools and strategies that use datasets to evaluate the fairness of a technique and decide how to alter it. It is anticipated that the use of data-driven innovation to address discrimination and bias in the years to come. Looking beyond the methodologies, it is crucial that AI researchers and developers keep an eye on emerging trends and prepare for future obstacles. Crucially, these future considerations must also include ethical concerns – the industry must ensure that these technologies have a fairness basis to be used in people's lives. With the development and use of AI having grown tremendously in recent years, the issue of bias and fairness is of ever-increasing importance. A wide range of innovations will therefore be necessary to move forward on this issue and help develop AI that is equitable [72].

12 CONCLUSIONS

AI technologies have the potential to replicate and exacerbate social inequalities present in the contexts from which they are developed and deployed. This presents significant societal issues in terms of the perpetuation of stereotyping, bias, prejudice, and ultimately discrimination and oppression. Beyond societal issues, the commercial bias of AI models could result in unfair treatment of agencies, sectors of the industry, or States. Therefore, it is necessary to reduce bias in AI models in order to fairly determine their impact on industrial and societal actors. Bias has disparate explanations depending on the context, and technical tools make the analysis and understanding of bias in data sets more straightforward. However, the clear understanding of the fairness implications of these biases and the solution of this problem is a challenge and an active area of research.

Decisions of AI models have an influence on the real world, and data about the real world is used to train new AI models. Thus, more biased decisions will lead to data more biased to train and thus a more biased generation of AI models. Therefore, it is necessary to mitigate bias in AI models, but this is challenging due to the large variety of possible biases and because of the difficulty of finding/predicting them. Moreover, this problem is particularly complicated taking into account the commercial bias of AI models and the growing associated societal concerns about fairness. For a fair determination of the bias/Fairness implications of decision J of agency A, the intended scope $\sim S\ A$ of decision J should be compared to the predicted scope $\sim S\ \mu$ of the reaction of rival M.

The reaction of the crop is the decision J of the proposed AI model. A model is trained to distort the competitive landscape in a profitable way. It models the competitive interactions between two agencies but is general enough to be applied to any actors. A simpler version can be solved numerically. Experiments show this simpler version can somewhat predict the competition for market share. More intermediates, such as faked explanations or mode-neutral perturbations, could help agencies currently out of direct competition with AI to better estimate the reactions of the AI models.

REFERENCES

[1] X. Ferrer, T. Van Nuenen, J. M. Such, M. Coté, and N. Criado, "Bias and discrimination in AI: a cross-disciplinary perspective," IEEE Technol. Soc. Mag., vol. 40, no. 2, pp. 72-80, 2021.

[2] A. W. Fazil, M. Hakimi, and A. K. Shahidzay, "A comprehensive review of bias in ai algorithms," Nusant. Hasana J., vol. 3, no. 8, pp. 1-11, 2023.

[3] S. O'Connor and H. Liu, "Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities," AI Soc., vol. 39, no. 4, pp. 2045-2057, 2024.

[4] P. S. Varsha, "How can we manage biases in artificial intelligence systems–A systematic literature review," Int. J. Inf. Manag. Data Insights, vol. 3, no. 1, p. 100165, 2023.

[5] D. Varona and J. L. Suárez, "Discrimination, bias, fairness, and trustworthy AI," Appl. Sci., vol. 12, no. 12, p. 5826, 2022.

[6] T. B. Modi, "Artificial Intelligence Ethics and Fairness: A study to address bias and fairness issues in AI systems, and the ethical implications of AI applications," Rev. Rev. Index J. Multidiscip., vol. 3, no. 2, pp. 24-35, 2023.

[7] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," in Ethics of Data and Analytics, Auerbach Publications, 2022, pp. 296-299.

[8] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in Conference on Fairness, Accountability and Transparency, PMLR, 2018, pp. 77-91.

[9] M. Karimi-Haghighi and C. Castillo, "Enhancing a recidivism prediction tool with machine learning: effectiveness and algorithmic fairness," in Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, 2021, pp. 210-214.

[10] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM Comput. Surv., vol. 54, no. 6, pp. 1-35, 2021.

[11] R. K. E. Bellamy et al., "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," IBM J. Res. Dev., vol. 63, no. 4/5, pp. 1-4, 2019.

[12] S. S. Mohammed and J. M. Al-Tuwaijari, "Skin disease classification system based on metaheuristic algorithms," in AIP Conference Proceedings, AIP Publishing, 2023.

[13] D. Pessach and E. Shmueli, "Algorithmic fairness," in Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook, Springer, 2023, pp. 867-886.

[14] V. Vakkuri, K.-K. Kemell, and P. Abrahamsson, "Ethically aligned design: an empirical evaluation of the resolvedd-strategy in software and systems development context," in 2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), IEEE, 2019, pp. 46-50.

[15] J. Whittlestone, R. Nyrup, A. Alexandrova, K. Dihal, and S. Cave, "Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research," London Nuff. Found., 2019.

[16] E. Ntoutsi et al., "Bias in data-driven artificial intelligence systems—An introductory survey," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 10, no. 3, p. e1356, 2020.

[17] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," arXiv Prepr. arXiv1609.05807, 2016, [Online]. Available: https://arxiv.org/abs/1609.05807.

[18] K. Crenshaw, "Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics," Droit et société, vol. 108, p. 465, 2021.

[19] M. Sadek, E. Kallina, T. Bohné, C. Mougenot, R. A. Calvo, and S. Cave, "Challenges of responsible AI in practice: scoping review and recommended actions," AI Soc., pp. 1-17, 2024.

[20] W. S. Nsaif et al., "Chatbot development: Framework, platform, and assessment metrics," Eurasia Proc. Sci. Technol. Eng. Math., vol. 27, pp. 50-62, 2024.

[21] A. S. Tejani, Y. S. Ng, Y. Xi, and J. C. Rayan, "Understanding and mitigating bias in imaging artificial intelligence," RadioGraphics, vol. 44, no. 5, p. e230067, 2024.

[22] S. Akter et al., "Algorithmic bias in data-driven innovation in the age of AI," Int. J. Inf. Manag., vol. 60, p. 102387, 2021.

[23] F. Osasona, O. O. Amoo, A. Atadoga, T. O. Abrahams, O. A. Farayola, and B. S. Ayinla, "Reviewing the ethical implications of AI in decision making processes," Int. J. Manag. Entrep. Res., vol. 6, no. 2, pp. 322-335, 2024.

[24] E. Ferrara, "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," Sci, vol. 6, no. 1, p. 3, 2023.

[25] G. M. Johnson, "Algorithmic bias: on the implicit biases of social technology," Synthese, vol. 198, no. 10, pp. 9941-9961, 2021.

[26] R. Schwartz, L. Down, A. Jonas, and E. Tabassi, "A proposal for identifying and managing bias in artificial intelligence," Draft NIST Spec. Publ., vol. 1270, 2021.

[27] P. Chen, L. Wu, and L. Wang, "AI fairness in data management and analytics: A review on challenges, methodologies and applications," Appl. Sci., vol. 13, no. 18, p. 10258, 2023.

[28] W. S. Nsaif et al., "Conversational agents: An exploration into Chatbot evolution, architecture, and important techniques," Eurasia Proc. Sci. Technol. Eng. Math., vol. 27, pp. 246-262, 2024.

[29] R. Agarwal et al., "Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework," Health Policy Technol., vol. 12, no. 1, p. 100702, 2023.

[30] T. P. Pagano et al., "Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods," Big Data Cogn. Comput., vol. 7, no. 1, p. 15, 2023.

[31] A. Fabris, A. Esuli, A. Moreo, and F. Sebastiani, "Measuring fairness under unawareness of sensitive attributes: A quantification-based approach," J. Artif. Intell. Res., vol. 76, pp. 1117-1180, 2023.

[32] A. Agarwal, H. Agarwal, and N. Agarwal, "Fairness Score and process standardization: framework for fairness certification in artificial intelligence systems," AI Ethics, vol. 3, no. 1, pp. 267-279, 2023.

[33] M. S. Farahani and G. Ghasemi, "Artificial intelligence and inequality: challenges and opportunities," Qeios, vol. 7, pp. 1-14, 2024.

[34] M. Zajko, "Conservative AI and social inequality: conceptualizing alternatives to bias through social theory," AI Soc., vol. 36, no. 3, pp. 1047-1056, 2021.

[35] K. S. Chadha, "Bias and Fairness in Artificial Intelligence: Methods and Mitigation Strategies," International Journal for Research Publication and Seminar, pp. 36-49, 2024.

[36] M. N. Khreisat, D. Khilani, M. A. Rusho, E. A. Karkkulainen, A. C. Tabuena, and A. D. Uberas, "Ethical implications of AI integration in educational decision making: Systematic review," Educ. Adm. Theory Pract., vol. 30, no. 5, pp. 8521-8527, 2024.

[37] N. Naik et al., "Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility?," Front. Surg., vol. 9, p. 862322, 2022.

[38] K. Patel, "Ethical reflections on data-centric AI: balancing benefits and risks," Int. J. Artif. Intell. Res. Dev., vol. 2, no. 1, pp. 1-17, 2024.

[39] R. Bahta, "Investigating the Complexities and Interdependencies of Algorithmic Biases in Healthcare Artificial Intelligence," 2024.

[40] A. Hasan, S. Brown, J. Davidovic, B. Lange, and M. Regan, "Algorithmic bias and risk assessments: Lessons from practice," Digit. Soc., vol. 1, no. 2, p. 14, 2022.

[41] J. Huang, G. Galal, M. Etemadi, and M. Vaidyanathan, "Evaluation and mitigation of racial bias in clinical machine learning models: scoping review," JMIR Med. Inform., vol. 10, no. 5, p. e36388, 2022.

[42] I. Mishkhal, N. Abdullah, H. H. Saleh, N. I. R. Ruhaiyem, and F. H. Hassan, "Facial Swap Detection Based on Deep Learning: Comprehensive Analysis and Evaluation," Iraqi Journal for Computer Science and Mathematics, vol. 6, no. 1, article 8, 2025, [Online]. Available: https://doi.org/10.52866/2788-7421.1229.

[43] R. R. Fletcher, A. Nakeshimana, and O. Olubeko, "Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health," Front. Artif. Intell., vol. 3, p. 561802, 2021.

[44] V. Shah and S. R. Konda, "Neural Networks and Explainable AI: Bridging the Gap between Models and Interpretability," Int. J. Comput. Sci. Technol., vol. 5, no. 2, pp. 163-176, 2021.

[45] P. Thunki, S. R. B. Reddy, M. Raparthi, S. Maruthi, S. B. Dodda, and P. Ravichandran, "Explainable AI in Data Science-Enhancing Model Interpretability and Transparency," African J. Artif. Intell. Sustain. Dev., vol. 1, no. 1, pp. 1-8, 2021.

[46] A. S. Albahri et al., "A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion," Inf. Fusion, vol. 96, pp. 156-191, 2023.

[47] M. Suffian and A. Bogliolo, "Investigation and Mitigation of Bias in Explainable AI," in CEUR Workshop Proceedings, 2022, pp. 89-94.

[48] M. DeCamp and C. Lindvall, "Latent bias and the implementation of artificial intelligence in medicine," J. Am. Med. Inform. Assoc., vol. 27, no. 12, pp. 2020-2023, 2020.

[49] L. Belenguer, "AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry," AI Ethics, vol. 2, no. 4, pp. 771-787, 2022.

[50] H. Saleh and I. Hussein, "Enabling Smart Mobility with Connected and Intelligent Vehicles: The E-VANET Framework," in Proceedings of International Conference on Applied Innovation in IT, vol. 12, no. 2, Anhalt University of Applied Sciences, 2024.

[51] P. Rouzrokh et al., "Mitigating bias in radiology machine learning: 1. Data handling," Radiol. Artif. Intell., vol. 4, no. 5, p. e210290, 2022.

[52] J. W. Gichoya et al., "AI pitfalls and what not to do: mitigating bias in AI," Br. J. Radiol., vol. 96, no. 1150, p. 20230023, 2023.

[53] M. Wan, D. Zha, N. Liu, and N. Zou, "In-processing modeling techniques for machine learning fairness: A survey," ACM Trans. Knowl. Discov. Data, vol. 17, no. 3, pp. 1-27, 2023.

[54] S. Siddique, M. A. Haque, R. George, K. D. Gupta, D. Gupta, and M. J. H. Faruk, "Survey on machine learning biases and mitigation techniques," Digital, vol. 4, no. 1, pp. 1-68, 2023.

[55] B. Koçak et al., "Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects," Diagn. Interv. Radiol., 2024.

[56] I. Banerjee, "Bias in radiology artificial intelligence: causes, evaluation and mitigation," in Medical Imaging 2024: Image Processing, SPIE, 2024, p. 129260O.

[57] L. H. Nazer et al., "Bias in artificial intelligence algorithms and recommendations for mitigation," PLOS Digit. Health, vol. 2, no. 6, p. e0000278, 2023.

[58] M. Nauta et al., "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai," ACM Comput. Surv., vol. 55, no. 13s, pp. 1-42, 2023.

[59] G. Curto and F. Comim, "SAF: Stakeholders' Agreement on Fairness in the Practice of Machine Learning Development," Sci. Eng. Ethics, vol. 29, no. 4, p. 29, 2023.

[60] B. Richardson and J. E. Gilbert, "A framework for fairness: A systematic review of existing fair AI solutions," arXiv Prepr. arXiv2112.05700, 2021, [Online]. Available: https://arxiv.org/abs/2112.05700.

[61] S. Shrestha and S. Das, "Exploring gender biases in ML and AI academic research through systematic literature review," Front. Artif. Intell., vol. 5, p. 976838, 2022.

[62] A. Limantė, "Bias in Facial Recognition Technologies Used by Law Enforcement: Understanding the Causes and Searching for a Way Out," Nord. J. Hum. Rights, vol. 42, no. 2, pp. 115-134, 2024.

[63] T. Santiago, "AI bias: How does AI influence the executive function of business leaders?," Muma Bus. Rev., vol. 3, no. 16, pp. 181-192, 2019.

[64] A. C. Timmons et al., "A call to action on assessing and mitigating bias in artificial intelligence applications for mental health," Perspect. Psychol. Sci., vol. 18, no. 5, pp. 1062-1096, 2023.

[65] E. Ferrara, "The butterfly effect in artificial intelligence systems: Implications for AI bias and fairness," Mach. Learn. with Appl., vol. 15, p. 100525, 2024.

[66] J. M. Alvarez et al., "Policy advice and best practices on bias and fairness in AI," Ethics Inf. Technol., vol. 26, no. 2, p. 31, 2024.

[67] M. Soleimani, A. Intezari, and D. J. Pauleen, "Mitigating cognitive biases in developing AI-assisted recruitment systems: A knowledge-sharing approach," Int. J. Knowl. Manag., vol. 18, no. 1, pp. 1-18, 2022.

[68] D. Dhinakaran, S. M. Sankar, D. Selvaraj, and S. E. Raja, "Privacy-Preserving Data in IoT-based Cloud Systems: A Comprehensive Survey with AI Integration," arXiv Prepr. arXiv2401.00794, 2024.

[69] S. Leavy, B. O'Sullivan, and E. Siapera, "Data, power and bias in artificial intelligence," arXiv Prepr. arXiv2008.07341, 2020, [Online]. Available: https://arxiv.org/abs/2008.07341.

[70] B. Li et al., "Trustworthy AI: From principles to practices," ACM Comput. Surv., vol. 55, no. 9, pp. 1-46, 2023.

[71] O. Akinrinola, C. C. Okoye, O. C. Ofodile, and C. E. Ugochukwu, "Navigating and reviewing ethical dilemmas in AI development: Strategies for transparency, fairness, and accountability," GSC Adv. Res. Rev., vol. 18, no. 3, pp. 50-58, 2024.

[72] E. Ferrara, "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," Sci, vol. 6, no. 1, p. 3, 2023.