iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** III **Month of publication:** March 2023

DOI: https://doi.org/10.22214/ijraset.2023.49990

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue III Mar 2023- Available at www.ijraset.com

From Bias to Fairness: A Review of Ethical Considerations and Mitigation Strategies in Artificial Intelligence

Saurabh Srivastava[1], Khushi Sinha[2]
Department of Engineering-Computer Science, Chandigarh University, Mohali, Punjab-140413

Abstract: ***Artificial intelligence (AI) has become increasingly popular in recent years and has been used in a range of industries to improve outcomes, streamline processes, and improve decision-making. But there are also moral questions raised by the employment of AI, particularly in light of potential bias and discrimination. In order to promote justice and reduce bias, this paper offers a thorough discussion of ethical issues and mitigation techniques in AI.***

The evolution of AI and its possible advantages and disadvantages are first covered in the paper. After that, it explores the different ethical issues surrounding AI, such as trust, accountability, fairness, and openness. The study emphasises the effects of bias and discrimination on AI systems as well as the possible outcomes of these problems. The study also discusses the various mitigation measures, such as algorithmic strategies, data pre-processing, and model validation, that have been suggested to mitigate bias and enhance justice in AI. In order to develop the subject of AI ethics, the study analyses the advantages and disadvantages of different frameworks and emphasises the necessity of continued interdisciplinary research and collaboration. The study's importance in advancing ethical concerns and fairness in AI is highlighted in the paper's conclusion. It offers information about the state of the field at the moment and points out potential directions for further study. Overall, the article is a useful tool for academics, professionals, and decision-makers who want to support ethical and responsible AI development and application.

Keyword: Artificial Intelligence ,Machine Learning, Ethics, Transparency, Mitigation Strategies, Fairness, Bias ,Framework.

I. INTRODUCTION

Artificial intelligence (AI) is an emerging technology that is revolutionising many different industries. Complex tasks can be carried out by AI systems more effectively and precisely than by conventional approaches, opening up possibilities for process improvement, better results, and improved decision-making. Yet, as AI is used more frequently, ethical issues have also emerged, particularly in light of the possibility of bias and discrimination in AI systems.

Biases can be incorporated into AI systems either consciously or unconsciously. The underlying data used to train the system and the methods used to process that data both have the potential to introduce bias. Biased AI systems may have serious repercussions, including mistreatment, discrimination, and even harm to specific people or groups.The potential for bias and discrimination in AI is one of the main worries. The data used to train the system, the methods used to interpret the data, and the environment in which the system is used are just a few of the variables that might cause biases to appear in AI systems. Biased AI systems may treat some people or groups unfairly and discriminatorily, perpetuating current social injustices and diminishing the promised advantages of AI. A rising number of people are interested in creating ethical frameworks and norms for the creation and use of AI in order to solve these issues. With openness, accountability, and justice at the forefront, these standards seek to ensure that AI is developed and deployed in a responsible and ethical manner.AI system development and implementation must abide by moral principles that safeguard people from harm and guarantee openness, responsibility, and justice. The complexity and size of AI systems, however, make it difficult to recognise and address ethical problems. For instance, biases could be challenging to identify and unintentionally reinforced by the system's design.

It's important to recognise and eliminate any potential bias and discrimination in order to ensure the ethical and responsible use of AI. These problems have been addressed using a variety of approaches, such as algorithmic techniques, data pre-processing, and model validation. The effectiveness of these tactics must be continually monitored and evaluated because they have limitations.Also, modern guidelines and standards for ethical AI have arisen, such as the GDPR and the FAT framework, to direct the creation and application of AI systems.

Although these frameworks strive to assure openness, responsibility, and equity in the use of AI, they also have limitations and implementation difficulties. It is vital for AI to take ethics into account given the growing influence of AI systems on people and society. This work seeks to contribute to the expanding body of research on AI ethics by offering a thorough analysis of ethical use of AI.

II. BACKGROUND

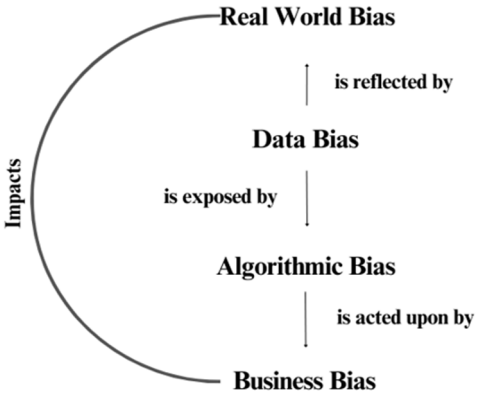
The emergence of AI systems can be dated to the middle of the 20th century, when computer scientists started investigating the possibility of building tools that could mimic human intellect. With the creation of the first AI programmes and the founding of the Dartmouth Conference, widely regarded as the birthplace of AI, the field of AI research made major strides in the 1950s and 1960s.
With early work in expert systems and symbolic reasoning, the field of artificial intelligence may be traced back to the 1950s. Unfortunately, advancement was modest until machine learning emerged in the 1980s, allowing computers to gradually improve their performance by learning from data. As a result, tremendous progress was made in fields including robotics, computer vision, and natural language processing.While AI research made advancement over the ensuing decades, it was frequently slow. Significant advances in AI research, especially in the area of deep learning, didn't occur until the 2010s, when they paved the way for the creation of systems capable of performing challenging tasks like speech and image recognition with an accuracy that could compete with that of humans.
Convolutional neural networks and recurrent neural networks, two examples of deep learning algorithms, have demonstrated impressive performance in recent years on a variety of tasks, including speech and picture recognition, game playing, and language translation. These developments have rekindled interest in artificial intelligence and its potential to revolutionise many facets of society. Although though AI may have certain advantages, there are worries about how it may affect society, especially in terms of bias and discrimination. For instance, it has been discovered that language models reinforce racial and gender prejudices and that facial recognition algorithms are less accurate for people with darker skin tones. These problems may have detrimental effects, such as maintaining current socioeconomic inequities or depriving particular groups of chances.As a result, there is a rising understanding of the necessity of addressing ethical issues in the creation and application of AI systems. This entails making sure AI systems are developed and implemented in a just and transparent manner, with measures in place for accountability and compensation in the event of harm.

III. ETHICAL CONSIDERATION IN AI

The creation and use of AI systems must take ethical considerations into account. It is crucial to make sure that AI technologies are created and deployed in ways that are fair, transparent, and accountable as they become more widespread in our daily lives. This section will examine several ethical issues relating to AI systems and give illustrations of practical uses where these issues are particularly pertinent.Bias is one of the most important ethical issues in AI. Because AI systems can only learn from data that is as neutral as the data itself, biassed data will lead to biassed AI systems. For instance, if an AI system is taught to make decisions using historical data that is biassed against particular groups, it may learn to do so. Serious repercussions may result from this, such as prejudice in lending, hiring, or criminal justice choices. It is crucial to make sure that the data used to train AI systems is varied and reflective of the population it is intended to serve in order to combat bias in AI.In addition, methods like bias testing and algorithmic auditing can be used to find and eliminate the causes of bias in AI systems. Fairness is another essential AI ethical factor. No one or any group shall be subjected to discrimination by AI systems because of their race, gender, ethnicity, or any other protected attribute. It is crucial to take into account how AI systems affect various populations and to keep an eye on their performance for inconsistent effects in order to ensure justice. AI ethics must also take explainability and transparency into account. AI systems frequently function as "black boxes," making it challenging to comprehend how they make judgements.This can raise issues with accountability and bias as well as raise concerns about how to make sure AI systems are working ethically and equitably. Another crucial ethical issue in AI is privacy, especially in light of the growing usage of AI systems to gather and analyse vast amounts of personal data. It is crucial to make sure AI systems are built to respect people's privacy and that they aren't employed to infringe on that right or to continue discriminatory or spying practises. Accountability is an important ethical factor in AI, to sum up.It is crucial to make sure that AI systems are held responsible for their choices and that procedures are in place to look into and correct any possible harm these systems may have caused. The significance of ethical concerns in AI is highlighted by real-world applications of AI, such as face recognition, autonomous vehicles, and predictive policing. For instance, face recognition technology has come under fire for having the potential to support prejudice and discrimination against disadvantaged groups.

Overall, ethical issues must be addressed thoughtfully and deliberately to ensure that AI systems are used for the benefit of society as a whole. Ethical issues are a critical component of the development and implementation of AI systems.

4 Stages of Ethical AI



IV. MITIGATION STRATEGIES

AIn order to reduce bias and promote fairness in AI systems, a multifaceted strategy that takes into account different development and deployment phases is necessary. The following techniques can be applied to encourage justice and reduce bias in AI systems:

1) *Algorithmic Transparency:* It is the capacity to comprehend how an AI system arrives at a conclusion or a suggestion. It is simpler to spot any biases or unfairness in the system when the decision-making process is more open. Techniques like explainable AI (XAI) and model interpretability can be used to accomplish this. In contrast to model interpretability, which involves using techniques to understand how a model comes to its decisions, XAI involves creating AI systems with built-in explanations that can be understood by humans.
2) *Data Collection And Curation:* The data used to teach AI systems may be biassed. Making ensuring the data used to train an AI system is representative and diverse is crucial. This can be accomplished by carefully gathering and curating data, including the use of data augmentation methods to broaden the data's diversity.
3) *Promoting Diversity And Inclusion In The AI:* Staff can aid in reducing bias in the systems' algorithms. In order to develop and implement AI systems fairly and with less bias, it is beneficial to have a diverse workforce. A diverse workforce can contribute a variety of perspectives and experiences to the process. Targeted recruitment efforts, training and development initiatives, and other measures can be used to accomplish this and the establishment of codes of conduct that prioritize diversity and inclusion.
4) *Regular Monitoring and Evaluation:* Monitoring and evaluating AI systems on a regular basis can help to spot and correct flaws that may develop over time. This entails examining how AI systems work and finding any patterns or biases that may manifest. It becomes possible to recognise and correct any biases or unfairness in the system before they have a major impact by routinely monitoring and evaluating AI systems.
5) Designing AI systems with ethical concerns in mind from the beginning is known as ethical design. Fairness, accountability, openness, and privacy must be given top priority throughout the development process. It is possible to create AI systems that are more resistant to prejudice and unfairness by designing them with these factors in mind.
6) *Regulatory Frameworks:* Creating legal guidelines for AI systems can aid in advancing justice and reducing bias. Regulatory frameworks can create methods for accountability and oversight as well as standards and guidelines for the creation and use of AI systems. It is feasible to guarantee that AI systems are created and implemented in a fair and ethical manner by creating regulatory frameworks for them.
7) *Ethical Principles and Codes of Conduct:* Creating ethical principles and codes of conduct for AI systems can aid in advancing justice and reducing prejudice. These standards and codes of conduct can create moral principles and best practises for the creation and use of AI systems, as well as support accountability and openness.

In conclusion, reducing bias and advancing fairness in AI systems calls for a multifaceted strategy that includes algorithmic transparency, data collection and curation, diversity and inclusion in the AI workforce, routine monitoring and evaluation, ethical design, and collaboration and engagement with stakeholders.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue III Mar 2023- Available at www.ijraset.com

V. CURRENT FREAMEWORKS AND STANDARDS

Many frameworks and standards are now in use to encourage fairness and reduce bias in AI systems. These frameworks and standards are made to address ethical issues including bias, fairness, transparency, explainability, privacy, and accountability that are connected to the creation and application of AI systems.

1) One such framework is the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, which was launched in 2016. This framework is designed to promote the ethical design and deployment of AI systems and consists of a set of guidelines and principles for developers and users of AI systems. The guidelines include provisions for transparency, accountability, and the protection of privacy and human rights.

2) Another framework is the Algorithmic Impact Assessment (AIA) framework, which was developed by the AI Now Institute in 2018. This framework is designed to help organizations assess the potential impacts of AI systems on various stakeholders and identify potential sources of bias or discrimination. The AIA framework consists of a set of tools and methods for conducting impact assessments and identifying potential sources of bias or discrimination.

3) The European Union's General Data Protection Regulation (GDPR) is another important framework for promoting fairness and mitigating bias in AI systems. The GDPR is designed to protect the privacy and personal data of EU citizens and requires organizations to obtain explicit consent before collecting and using personal data. The GDPR also includes provisions for transparency, accountability, and the right to be forgotten.

4) A multi-stakeholder organisation called The Partnership on AI intends to research and develop best practises for AI technologies. Fairness, openness, responsibility, and privacy are just a few of the ethical guidelines that the group has developed for AI.A methodology for auditing AI systems has been established by the AI Now Institute, a research centre based at New York University, and it focuses on evaluating potential problems connected to bias, discrimination, and other ethical issues.

5) Guidelines for creating and deploying trustworthy AI systems have been created by the World Economic Forum's Centre for the Fourth Industrial Revolution. These guidelines emphasise making sure the systems are open, understandable, and in line with ethical standards.

6) A set of moral principles known as the Montreal Declaration for Responsible AI aims to guarantee that AI technologies are created and applied in a way that respects human rights and advances societal welfare. The declaration places a strong emphasis on the value of openness, responsibility, and inclusivity in the creation of AI.

7) Governments and business stakeholders are brought together by the Global Partnership on Artificial Intelligence (GPAI), a global initiative, to support the ethical development and application of AI technologies. Transparency, responsibility, and respect for human rights are just a few of the top objectives the partnership has set for ethical AI.

8) The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems is a collection of ethical guidelines for AI that was created by the Institute of Electrical and Electronics Engineers (IEEE). The standards emphasize the importance of transparency, accountability, and inclusivity in AI development, and provide guidance on ethical decision-making related to AI technologies.

Each of these frameworks has advantages and disadvantages when it comes to tackling ethical issues with AI. For instance, the IEEE Global Initiative offers a thorough set of principles and guidelines for the ethical creation and use of AI systems, but it might be challenging to put these principles into practise. The ability of the AIA framework to address systemic bias may be constrained, but it offers a collection of tools and techniques for locating potential sources of bias or discrimination in AI systems. The GDPR offers robust private and personal data protections, but it might not handle other ethical issues like fairness and transparency.

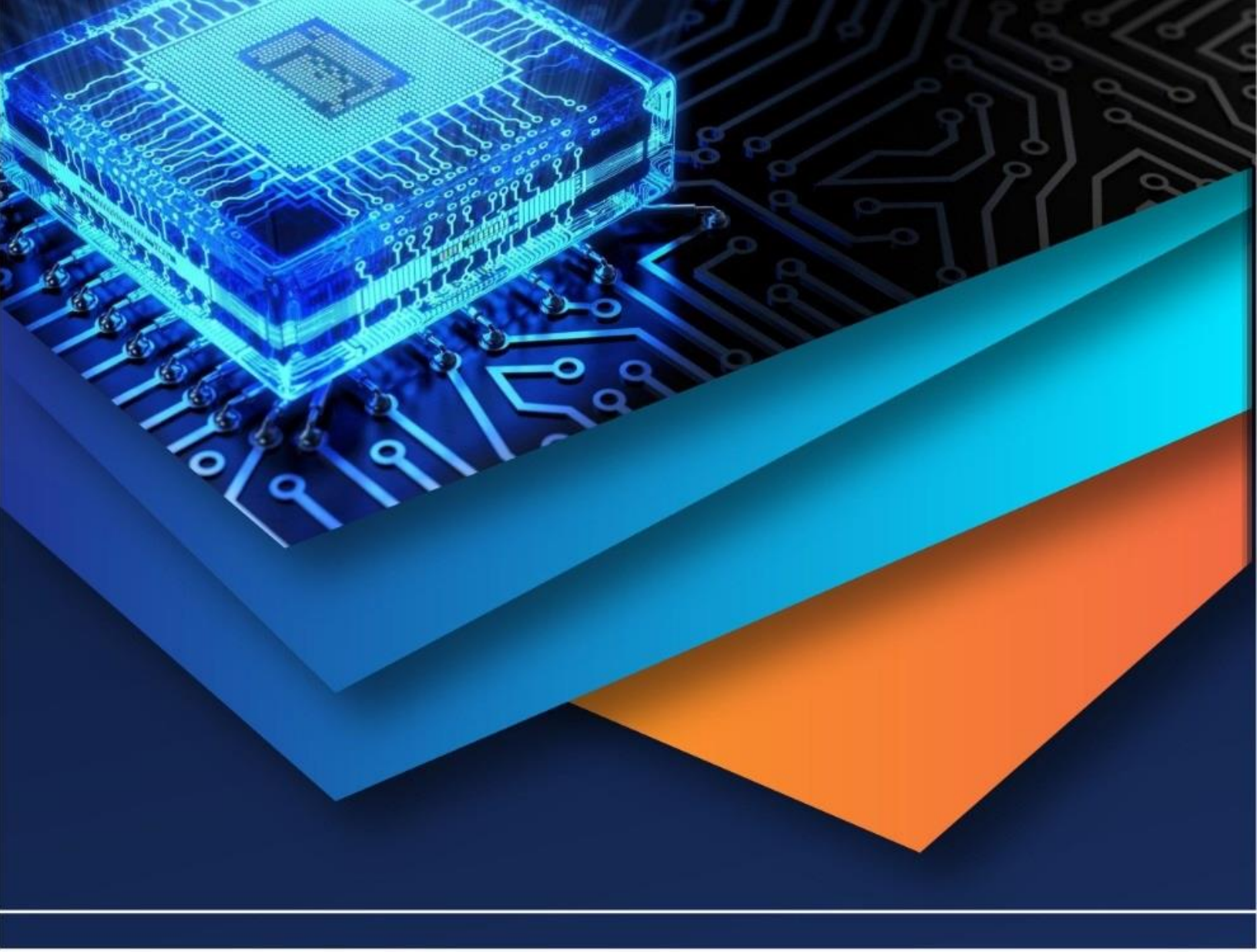
VI. CONCLUSION AND FUTURE WORK

Artificial intelligence's ethical issues and mitigation techniques, with a focus on addressing bias and encouraging justice in AI systems. This analysis emphasises the intricate character of AI's ethical problems, which need for a multidisciplinary strategy and continuous assessment. The AI pipeline's various phases, from data collecting and processing to algorithm design and implementation, are susceptible to prejudice and discrimination. As a result, solving these problems calls for an all-encompassing strategy that takes into account the system as a whole and its larger social context. Pre-processing methods, algorithmic approaches, model validation, and interpretability are just a few of the measures for reducing bias and fostering fairness in AI systems. These approaches have potential, but their usefulness depends on the situation and calls for constant review and improvement.The conclusion is that while these frameworks serve as helpful beginning points, constant evaluation and adaptation are necessary to make sure they stay applicable and efficient.

In conclusion, the ethical issues and mitigation techniques in artificial intelligence are intricate and varied, necessitating continuous analysis, review, and modification. We can advance the creation of ethical and just AI systems that benefit society as a whole by encouraging a more thorough and diverse approach to AI ethics.

REFERENCES

[1] G. Narayanan, A., & Lohia, P. (2020)"On the ethics of AI applications in healthcare. Journal of Chemical Information and Modeling", 60(5), 2140-2147.
[2] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). "Model cards for model reporting".. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).
[3] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé III, H., & Crawford, K. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010.
[4] Buolamwini, J., & Gebru, T. (2018)"Gender shades: Intersectional accuracy disparities in commercial gender classification".In Conference on Fairness, Accountability and Transparency (pp. 77-91).
[5] Sweeney, L. (2013)" Discrimination in online ad delivery. Communications of the ACM", 56(5), 44-54.
[6] Narayanan, A., & Pentland, A. (2017)" A.I. as personhood? How we can think about intelligent machines and moral considerability". In Proceedings of the 26th International Joint Conference on Artificial Intelligence (pp. 3884-3891). AAAI Press.
[7] Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). "Semantics derived automatically from language corpora contain human-like biases. Science", 356(6334), 183-186.
[8] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016)"The ethics of algorithms: Mapping the debate. Big Data & Society", 3(2), 2053951716679679.
[9] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). "Mitigating unwanted biases with adversarial learning". In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335-340). ACM.
[10] Buolamwini, J., & Gebru, T. (2018). "Gender shades: Intersectional accuracy disparities in commercial gender classification". In Conference on fairness, accountability and transparency (pp. 77-91).
[11] Suresh, H., & Guttag, J. V. (2019)" A framework for understanding unintended consequences of machine learning".arXiv preprint arXiv:1901.10002.
[12] European Union Agency for Fundamental Rights. (2018). Handbook on European data protection law. Publications Office of the European Union.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY