SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 13001-13010
Open Access. Online & Print

DOI: 10.5281/zenodo.20374091

BIAS, FAIRNESS, AND ETHICAL CHALLENGES IN ARTIFICIAL INTELLIGENCE: A COMPREHENSIVE REVIEW OF CAUSES, IMPACTS, AND MITIGATION STRATEGIES

Ohmini Krishnamurthy Rajendran[1]*

[1]Consultant, MBBS, MD Radiodiagnosis, KIMS Hospital and Research Centre
Krishna Rajendra Road , Parvathipuram, Vishweshwarapura, Basavanagudi, Bengaluru, Karnataka 560004
Jsmnk4@gmail.com

Received: 01/03/2026
Accepted: 26/04/2026

Corresponding Author: Ohmini Krishnamurthy Rajendran (Jsmnk4@gmail.com)

ABSTRACT

Concerns about AI system fairness and bias have emerged as a result of the rapid application of artificial intelligence (AI) in a variety of fields, including healthcare decision making, diagnostics, and others. Health care, employment, the justice system, credit rating, and GenAI models that create synthetic media all depend on this. These systems may produce biases in the representation of people in synthetic data, which can lead to unfair outcomes and promote inequality. Concisely, this survey discusses AI fairness and bias's causes, effects, and mitigation strategies. There is a discussion of data, algorithm, and human decision bias, with an emphasis on generative AI bias, in which models may amplify and perpetuate social preconceptions. Since generative AI is increasingly used to create public perception material, we discuss the societal effects of biased AI systems on inequality and stereotypes. We look at the ethical issues associated with their adoption, ways to mitigate them, and the need for multidisciplinary collaboration. Using a comprehensive review of the existing literature in a wide range of academic fields, we identify and evaluate AI bias, including generative AI bias. This study discusses the current AI bias mitigation strategies, such as data pre-processing, model selection, and post-processing, as well as the negative effects of AI bias on people and society. We emphasize the need for specialized strategies and the inherent issues with generative AI models. To combat AI prejudice, a comprehensive plan is needed. The exploration of fair and ethical AI paradigms, more openness and accountability in AI systems, and diverse and representative datasets are all necessary for this. This study discusses AI bias's causes, effects, and mitigation measures, with a focus on generative AI.

KEYWORDS: Artificial Intelligence, Bias, Fairness, Discrimination, Mitigation Strategies.

1. INTRODUCTION

As AI systems are deployed increasingly, discussions concerning fairness and prejudice in AI have heated up as bias and discrimination become more likely. The poll discusses the causes, effects, and solutions of AI bias and fairness. AI systems like Buolamwini and Gebru's [1] face recognition systems and Dastin's employment algorithms have biases against particular populations. Communities in employment, lending, and the criminal justice system can be harmed by these prejudices, which may perpetuate systemic discrimination and inequality. Increasing data quality and implementing fair algorithms are two mitigating techniques suggested by researchers and practitioners.

This study covers AI bias sources and impacts, including data, algorithmic, and user bias and ethical concerns. It discusses multidisciplinary collaboration, obstacles, constraints, and research on mitigation strategies. Academics, politicians, and researchers acknowledge AI's bias and fairness. The complicated issues of AI fairness and prejudice, as well as their causes, effects, and methods for mitigating them, are the focus of this overview. In order to assist in the development of AI systems that are more responsible and ethical, the research reveals the causes, effects, and ways to reduce fairness and prejudice in AI [2].

2. SOURCES OF BIAS IN AI

Numerous industries and lives can be transformed by artificial intelligence (AI). However, prejudice is a significant obstacle to AI adoption and development. Systematic errors in decision-making that result in unequal outcomes are known as bias. AI may have bias in data collecting, algorithm design, and human interpretation. An AI system known as machine learning may learn and reproduce biases in the training data, leading to conclusions that are unfair or discriminatory. In this section, AI's data, algorithmic, and user bias will be examined. We shall also examine their influence in real life.

2.1. Definition of Bias in AI and Its Different Types

A systematic error in decision-making that results in imbalance is known as bias. The collection of data, construction of algorithms, and interpretation of outcomes all introduce bias into AI. An AI system known as machine learning may learn and repeat biases in the data they are trained on, leading to inaccurate or biased outcomes. We must address AI prejudice to create fair and just systems. We will discuss the causes, effects, and ways to reduce AI bias in the following sections [3].

2.2. Sources of Bias in AI, including Data Bias, Algorithmic Bias, and User Bias

Machine learning pipelines may be skewed by factors such as user interactions, algorithm design, and data collection. AI data, algorithmic bias, and user bias are shown in the poll. Biased machine learning algorithms are those trained on data that isn't representative. It is possible for the data to be inaccurate, biased, incomplete, or lacking crucial information. Machine learning models show algorithmic bias. Biassed algorithms or decision criteria may achieve this. AI is influenced by users' own preconceptions or biases, either by accident or intention. Users' biased training data or actions may cause this.

Using bias-aware algorithms, human input, and dataset augmentation, these biases can be reduced. Diverse data are added to training datasets to improve representation and remove bias. Algorithms that are aware of bias reduce system output biases. The program uses user feedback to find and correct biases. New approaches to reducing AI bias are being looked into. Research and refine these tactics to establish fair AI systems for all users.

2.3. Real-World Examples of Bias in AI

Biased AI systems abound in the criminal justice and healthcare sectors. US criminal courts' COMPAS predicts reoffending. ProPublica reported that even without prior convictions, black defendants posed a greater risk. Wisconsin exhibits a similar pattern, according to another study. Healthcare AI that predicted death disadvantaged African-Americans. Obermeyer et al. found that African-American patients had higher risk estimations despite similar age and health. Racism may result in poor healthcare for African-Americans[4]. Another example is police face recognition AI bias. Darker complexion face recognition produced more false positives, according to NIST. Prejudice may lead to improper arrests and convictions.

Lastly, generative AI may discriminate. OpenAI's DALL-E, Midjourney, and StableDiffusion all displayed racial and stereotype bias. Gender bias was evident in the models, which featured mostly male CEOs. Women's CEO under-representation mirrors this bias. Black terrorists or villains were typically depicted by models. Social prejudices may arise as a result of generative AI. Internet photos contained inequalities, therefore GenAI models trained on them may be biased. This demonstrates that in order to avoid biased and unrepresentative generative model

outputs, AI research requires extensive and balanced training datasets[5]. The various biases and the significant effects they have on AI systems are depicted in Table 1, highlighting the need for additional research and mitigation.

Table 1: Specific Types of AI Bias.

Type of Bias	Description	Examples
Sampling Bias	Occurs when training data do not properly represent the target population, leading to skewed predictions and poor performance for certain groups.	A face recognition system performs poorly on non-white faces because it was mainly trained on white faces.
Algorithmic Bias	Happens during algorithm design and implementation, where certain traits or characteristics are unfairly favored.	A hiring algorithm unfairly favors candidates based on age or gender.
Representation Bias	Arises when the dataset is not representative of the actual population being modeled.	Female patients are diagnosed less accurately due to a medical dataset dominated by male patient data.
Confirmation Bias	Occurs when AI systems reinforce the assumptions, beliefs, or expectations of developers or users.	A hiring system predicts job success according to the hiring manager's pre-existing beliefs.
Measurement Bias	Happens when methods of data collection or measurement over-represent or under-represent certain groups.	A survey collects more urban responses than rural ones, leading to under-representation of rural attitudes.
Interaction Bias	Occurs when AI systems interact unfairly with users or respond differently to different groups.	A chatbot shows gender bias during conversations.
Generative Bias	Found in generative AI systems where outputs reflect biases present in training data, causing imbalanced or culturally skewed content.	A text generation model trained mainly on Western literature over-represents Western culture and under-represents other cultures.
Image Generation Bias	Happens in AI image-generation models when training data lack diversity across ethnicities or nationalities.	An image-generation model struggles to generate people from multiple ethnic backgrounds because the training dataset mainly contains limited nationalities.

3. IMPACTS OF BIAS IN AI

The rapid development of AI has both benefits and risks. A major issue is the impact that AI prejudice has on society. Inequality may be made worse by AI bias. This may discriminate against the poor and restrict them access to essential services. It reinforces gender preconceptions and may lead to new skin-color, race-, and appearance-based discrimination[6]. For AI systems to be fair and beneficial to all users, bias must be identified and eliminated. Biased AI raises ethical questions about discrimination, developer and policymaker accountability, public trust in technology, and human agency and autonomy. Collaboration is required to address these ethical issues. Developing and deploying AI systems need ethical and legal frameworks for justice, transparency, and accountability.

3.1. Negative Impacts of Bias in AI on Individuals and Society, Including Discrimination and Perpetuation of Existing Inequalities

Humans and society may encounter AI prejudice. Inequality may rise as a result of biased AI systems highlighting prejudice. People of color are more likely to be wrongfully convicted, so it's possible that criminal justice algorithms will wrongfully convict or punish them. Healthcare and finance might be limited by AI bias. Biased credit scoring algorithms may under-represent low-income and minority borrowers, making loans and mortgages harder to get.

Gender stereotypes may be encouraged by AI bias. Security systems may continue to exhibit gender bias because they are trained on data that is primarily composed of males. GenAI models show mostly male CEOs [7].

AI bias has the potential to engender new racial, color, and attractiveness biases as well as perpetuate injustices. As would be expected, GenAI models with gender bias portray terrorists and criminals as minorities.

Service denials, job losses, and false convictions may result from public use of these technologies. Prospects and relationships are influenced by how people perceive themselves and others. False AI systems may spread prejudice and exclude people. Culture and civilization may suffer as a result of more AI in everyday life. These biases must be addressed during AI system development to prevent damage..

3.2. Discussion of the Ethical Implications of Biased AI

Prejudiced AI raises ethical concerns. Discrimination on the basis of disability, ageism, sexism, and race is severe. Inequality may rise and

individuals may be marginalized by biased AI. Because healthcare is delicate, erroneous AI systems may harm patients and delay treatment[8]. AI systems must be built and operated in an open and ethical manner by developers, organizations, and governments.

AI system designers and developers are responsible for biases. However, AI system developers and implementers must be held accountable for prejudice by ethical and regulatory standards. AI systems that are biased may damage people's faith in technology, making it harder for people to use new technology or even rejecting it. If people don't trust AI or think it discriminates against them, it might not be beneficial to society and the economy.

AI that is biased reduces human agency. Prejudice toward AI may restrict freedom and maintain control. Recruitment AI systems have the potential to unilaterally exclude marginalized individuals from employment and social involvement[9]. Biased AI's ethical concerns need developer, legislator, and social engagement.

Ethics and regulation are required for AI system development and use to be fair, transparent, and accountable. It is necessary to investigate the social function of AI and involve people in its ethical development.

4. MITIGATION STRATEGIES FOR BIAS IN AI

Academics and practitioners alike have generated numerous mitigation strategies for AI bias. Data pre-processing, model selection, and post-processing are all options. Each technique has drawbacks, such as the lack of varied and representative training data, the difficulty of recognising and evaluating bias, and the trade-offs between fairness and accuracy. Ethical issues arise when bias categories and groups are ranked for the purpose of bias prevention. In fact, developing AI systems that are just and fair for society requires resolving these issues. We need to keep looking into these problems and coming up with solutions, and we need to make sure that AI systems work for everyone[10].

4.1. Overview of Current Approaches to Mitigate Bias in AI, Including Pre-Processing Data, Model Selection, and Post-Processing Decisions

AI bias reduction is challenging. There are many options. It is common practice to preprocess AI model data so that it accurately represents the entire population, particularly marginalized groups. Oversampling, undersampling, or false data are all possible.

Table 2: Diverse AI Bias Reduction Challenges and Methods.

Methodology	Explanation	Instances / Examples	Challenges and Limitations	Ethical Issues
Pre-processing Data	Identifies and mitigates data biases before model training. Data are balanced through oversampling, undersampling, or synthetic data generation to ensure demographic representation, including historically marginalised populations.	1. Oversampling darker-skinned individuals in face recognition systems. 2. Data enhancement for under-represented populations. 3. Adversarial debiasing to train bias-resistant models.	1. The process is time-consuming. 2. May be ineffective when training data are highly skewed.	1. Over- or under-representation of certain groups may create or reinforce bias. 2. Privacy concerns related to collecting and using data from historically marginalised communities.
Fair Model Selection Approaches	Focuses on selecting fair AI models using group fairness and individual fairness strategies. Regularisation penalises discriminatory predictions, while ensemble methods reduce bias by combining multiple models.	1. Selecting demographic parity classifiers. 2. Model selection based on group or individual fairness. 3. Regularisation methods to penalise discriminatory predictions. 4. Ensemble approaches to minimise bias.	1. No universal agreement on fairness definitions and thresholds. 2. Fairness may conflict with other performance measures such as accuracy or efficiency.	1. Models may still reinforce social prejudices if fairness criteria are not properly met.
Post-processing Techniques	Adjusts AI model outputs after prediction to reduce bias and ensure fairness. Researchers propose methods that equalise odds by maintaining equal false-positive and false-negative rates across demographic groups.	1. Post-processing methods that equalise odds.	1. Techniques can be complex and require large datasets.	1. Trade-offs may occur between different types of bias during calibration. 2. Unequal distribution of outcomes across groups may still appear unexpectedly.

When oversampled, Buolamwini and Gebru found that face recognition algorithms work better for darker skin. Prior to model training, bias is reduced by data preprocessing. Underrepresented groups may benefit from adversarial debiasing and data enhancement. Document augmentation and biases in the dataset [11]. AI bias can be avoided by carefully selecting data analysis models. Fairness-based model selection is recommended by researchers for both groups and individuals. Kamiran and Calders' demographic parity classifiers distribute positive and negative outcomes equally across demographic groups. Fairness and bias-reduction models are also options. Regularization penalizes discriminating predictions, whereas ensemble methods reduce bias by combining models[12]. AI bias is reduced by post-processing choices. It is necessary to adjust the output of AI models for fairness and bias. Researchers match model predictions with the same likelihood after processing, requiring the same false positive and false negative rates for each demographic group. These strategies lessen AI bias but have drawbacks. When training models, biased data pre-processing may be inefficient and time-consuming. Post-processing may require a lot of data and unfairness may restrict model selection. Reducing AI bias requires more research and development. Generative AI bias requires a comprehensive strategy to overcome. Preprocessing of diversity and representativeness data should begin. To avoid over-representing a single group in training datasets, it requires actively finding and utilizing diverse data sources. Clear bias detection is necessary for model selection. Bias models can be evaluated with adversarial training. AI-generated biases are corrected by content post-processing [13]. Transfer learning or filters might help models. Audits, monitoring, and feedback are necessary for fair generative AI. Ethical AI principles, diverse AI research teams, and multidisciplinary cooperation to find and reduce AI bias should support such endeavors. Think about how these practices will affect society and ethics. Fairer model projections may change group results and include bias trade-offs. Table 2 lists solutions and drawbacks.

4.2. Discussion of the Limitations and Challenges of These Approaches

There are a number of AI bias mitigation methods, but each has drawbacks. Lack of varied and representative training data is a major issue. AI systems may provide skewed results due to data bias. It is difficult to collect diverse and representative data for sensitive or uncommon events. When gathering medical or financial data, privacy problems may arise. As a mitigation strategy, dataset augmentation may be hindered by such obstacles. Another challenge is measuring and detecting AI bias. Algorithm bias is difficult to identify and quantify because of its opaqueness or complexity. It can be difficult to identify bias because it can originate from users, algorithms, and data. Bias-aware algorithms and user feedback channels may be less effective as mitigation strategies. Fairness and accuracy may be compromised by mitigation strategies. Changing the algorithm to treat all groups equally reduces algorithmic bias. However, this may reduce accuracy for specific groups or environments. To strike a balance between accuracy and fairness, deliberate compromises are required. Finally, ethical difficulties arise when ranking prejudice categories and prioritising groups for bias prevention. Should we investigate all biases or should we concentrate more on prejudice that affects historically marginalized groups? Development and implementation of bias mitigation may be complicated by these ethical considerations. Creating fair and equitable systems requires eliminating AI bias, which is tough. To deploy AI systems for the good of society, continuing research and mitigating strategies are needed.

5. FAIRNESS IN AI

AI fairness is hot in academia and industry. AI fairness implies no bias in AI systems. This is difficult because of the various biases in these systems. Literature promotes counterfactual, societal, and individual justice. Prejudice and fairness are similar but distinct because bias is not intentional while fairness is. For AI fairness, consider the context and stakeholders. AI fairness has advantages in real life[14].

5.1. Definition of Fairness in AI and Its Different Types

It is difficult to debate AI fairness among academics and industry. Fair AI requires no bias. Know how to get rid of the many biases in AI to keep things fair. Literature promotes counterfactual, group, and individual justice. Fair AI systems treat groups in an equitable or fair manner. Due to demographic parity, unfairness, differential maltreatment characterized by misclassification rates, and equal opportunity, the true positive rate (sensitivity) and the false positive rate (specificity) are identical across demographic groups. In order to be fair to individuals, AI systems must treat people

who are in the same group equally. Similarity-based or distance-based criteria may ensure the AI system treats similar people similarly.

New idea counterfactual fairness tries to make AI systems fair in hypothetical scenarios. An AI system would have made the same decision on an individual regardless of group membership if their traits were different, according to counterfactual fairness.

Both procedural and causal fairness guarantee that biases and inequities from the past are not perpetuated in the decision-making process. In point of fact, fairness types might overlap. Trade-offs may be necessary to achieve justice due to the fact that numerous fairnesses conflict. AI fairness necessitates consideration of the context and stakeholders. For AI justice, it is essential to comprehend fairness categories and how to balance and prioritize them in various contexts.

5.2. Comparison of Fairness and Bias in AI

It is difficult to debate AI fairness among academics and industry. Fair AI requires no bias. Know how to get rid of the many biases in AI to keep things fair. Literature promotes counterfactual, group, and individual justice. Fair AI systems treat groups in an equitable or fair manner. Due to demographic parity, unfairness, differential maltreatment characterized by misclassification rates, and equal opportunity, the true positive rate (sensitivity) and the false positive rate (specificity) are identical across demographic groups. In order to be fair to individuals, AI systems must treat people who are in the same group equally. Similarity-based or distance-based criteria may ensure the AI system treats similar people similarly.

New idea counterfactual fairness tries to make AI systems fair in hypothetical scenarios. An AI system would have made the same decision on an individual regardless of group membership if their traits were different, according to counterfactual fairness.

Both procedural and causal fairness guarantee that biases and inequities from the past are not perpetuated in the decision-making process. In point of fact, fairness types might overlap. Trade-offs may be necessary to achieve justice due to the fact that numerous fairnesses conflict. AI fairness necessitates consideration of the context and stakeholders. For AI justice, it is essential to comprehend fairness categories and how to balance and prioritize them in various contexts.

Table 3: Defining Fairness for AI in Terms of Categories.

Fairness Type	Description	Examples
Group Fairness	Ensures AI systems treat demographic groups equally or equitably. It includes concepts such as demographic parity, differential treatment, and equal opportunity.	1. **Demographic Parity:** Equal positive and negative outcomes across demographic groups [31]. 2. **Disparate Treatment:** Misclassification rates used to identify unfair treatment [30]. 3. Equal true positive (sensitivity) and false positive (1-specificity) rates across demographic groups [11].
Fairness to Individuals	Ensures that similar individuals receive similar treatment regardless of group membership. Similarity- or distance-based measures are commonly used.	AI systems use similarity-based or distance-based criteria to treat comparable individuals equally [25].
Counterfactual Fairness	Ensures fairness in hypothetical scenarios by requiring AI systems to make the same decision for a person even if their demographic attributes were different.	An AI system gives the same judgement for a person even when attributes such as gender or race are altered hypothetically [35].
Procedural Justice	Focuses on ensuring fair, transparent, and accountable decision-making processes in AI systems.	AI systems with transparent and explainable decision-making processes.
Fair Cause	Ensures that AI systems do not reinforce historical discrimination, social inequalities, or past injustices.	Developing AI systems that avoid historical biases and injustices [4–6].

5.3. Real-World Examples of Fairness in AI

The benefits of incorporating fairness are demonstrated by real-world AI fairness examples. Criminal recidivism can be predicted using COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). According to research, the system incorrectly predicted higher reoffending rates for African-American defendants than for white defendants. The Northpointe COMPAS was changed to a more accurate but less racially biased "race-neutral" algorithm to fix this. Another use of AI in recruiting is AI recruiting techniques may disadvantage women in male-dominated fields, according to research. "Gender decoder" software is used by some employers to look for language in job ads that might discourage women. The final example is healthcare AI. AI algorithms that anticipate health care outcomes may discriminate against African Americans, according to studies. Subgroup analysis is used by researchers to find and reduce bias in AI model training data. Let's examine some real-world instances of how fairness

might benefit AI systems. Promoting social justice and equality, addressing prejudice and fairness may improve AI accuracy, ethics, and impartiality.

6. MITIGATION STRATEGIES FOR FAIRNESS IN AI

Fairness in AI decisions becomes increasingly important as it becomes more widely used. AI in healthcare, finance, and law must make unbiased decisions since it may impact lives. Either individual or collective justice may solve this issue. Fairness trade-offs and difficulties defining fairness are among the issues with these approaches. This section discusses AI fairness mitigation strategies, issues, and solutions. Understanding these mitigation strategies may assist us in developing AI systems that are impartial, fair, and equitable [16].

6.1. Overview of Current Approaches to Ensure Fairness in AI, including Group Fairness and Individual Fairness

Fairness in AI is challenging, hence new methods are being investigated. Individual justice and communal justice are the two approaches. Genders, nationalities, and ethnicities ought to be treated equally by AI systems. AI cannot discriminate because of group fairness. Resampling, pre-processing, or post-processing AI model training data can accomplish this. Re-sampling biased data-trained AI models can ensure that each group is accurately represented. AI model output groups may be eliminated through post- or pre-processing. In order to lessen inequality, Corbett-Davies and her colleagues advocated for risk minimization. AI systems that are fair treat everyone equally, regardless of membership status. Because of individual fairness, AI does not harm people. Counterfactual or causal fairness allows for individual fairness. Counterfactual fairness is one example. This suggests that regardless of race or gender, the AI model would have reached the same conclusion. Beyond AI fairness for groups and individuals, other factors are important. Other attributes are accountability, openness, and explanation.

Accountability holds programmers accountable for technological harm, and transparency demonstrates to customers how the AI system makes decisions. AI system judgments are a part of explainability. Fairness in AI is difficult and requires social scientists, lawyers, ethicists, and computer scientists to work together. Utilize alternative fairness strategies to construct AI systems that are accountable, transparent, and fair.

6.2. Discussion of the Limitations and Challenges of These Approaches

These methods might make AI more fair, but they have drawbacks. Problematic compromises between fairness types exist. Group fairness techniques may unjustly penalise group members, while individual fairness methods may not address structural biases affecting entire groups. Also, it might be hard to figure out how to strike a balance and be fair in the context. The definition of fairness is different. Fairness is interpreted differently by individuals and communities, and it changes over time. Stakeholder-fair AI system development may be hindered as a result[17]. Statistics and assumptions used in many AI fairness approaches may not accurately reflect human behavior and decision-making. Intersectionality and how race, gender, and socioeconomic status affect findings may be overlooked in group fairness evaluations.

Last but not least, AI fairness might have unintended effects. Racial disparities in arrest rates may be exacerbated by decreasing bias in predictive police algorithms, according to some research. Table 4 summarizes the strategy.

Despite these obstacles, fair and egalitarian AI research continues to be important and active. New approaches that take into account justice and equality in a variety of contexts must be developed in subsequent research to address these issues.

Table 4: AI Fairness Approaches and Problems.

Approach	Description	Examples	Limitations and Challenges
Group Fairness	AI systems should treat genders, races, and ethnic groups equally. The goal is to prevent discrimination against any demographic group through methods such as resampling, pre-processing, or post-processing.	1. Dataset resampling to achieve balance. 2. Adjusting AI model outputs through pre-processing or post-processing techniques.	1. May unintentionally treat some group members unfairly. 2. May struggle to overcome deep structural or historical prejudices. 3. Group fairness measures may ignore intersectionality among individuals.
Individual Fairness	Ensures AI systems treat individuals equally regardless of group membership. Prevents biased decisions against particular	1. Counterfactual fairness: ensuring the same decision regardless of race or gender.	1. May fail to address broader group-level biases. 2. Difficult to determine the most

	individuals and may use counterfactual or causal fairness methods.		appropriate definition of fairness and balance trade-offs between fairness measures.
Transparency	Allows users to understand how AI systems make decisions through transparent and user-friendly methods.	Transparent AI decision-making processes and interpretable systems.	1. Fairness definitions may vary for individuals and groups. 2. Concepts of fairness may change over time and across contexts.
Accountability	Holds developers and organisations responsible for harm caused by unfair AI systems and ensures mechanisms for damage repair.	AI developers being responsible for biased or harmful AI outcomes.	1. Establishing responsibility and liability can be complex. 2. Difficulties in enforcing accountability across organisations and jurisdictions.
Explainability	Refers to the ability of AI systems to explain their decisions and outputs to users in understandable terms.	Interpretable AI decisions that users can understand and verify.	1. Managing the complexity of human behaviour and decision-making is difficult.
Intersectionality *(consideration rather than a strategy)*	Examines how overlapping factors such as race, gender, and socioeconomic status jointly influence outcomes in AI systems.	Designing AI systems that evaluate multiple identity dimensions together.	1. Managing the complexity of multiple intersecting identities is challenging. 2. Ensuring fairness across all identity combinations can be difficult.

7. CONCLUSIONS

The study concludes with a discussion of the numerous types of bias found in AI and ML systems as well as the significant social effects they have, including the growing concern regarding generative AI bias. If they are not constructed and examined with care, these sophisticated computer tools may perpetuate and exacerbate prejudices, including those regarding race, gender, and other social dimensions. Numerous biased AI systems, particularly generative AI, have demonstrated the need for comprehensive tools to identify and eliminate biases during AI development. Counterfactual fairness, robust data augmentation, diverse representative datasets, and unbiased data collection were all examined in this study to reduce bias. We discussed AI's ethical implications for privacy and the necessity for openness, oversight, and constant review of AI systems.

Diverse training data and subtle bias issues in generative models, particularly content development and synthetic data, should be the focus of future research. Comprehensive frameworks and norms for responsible AI and ML, including openness in training data, model choices, and generating processes, are critically required. Diverse AI development and evaluation teams bring perspectives that aid in identifying and addressing biases. Finally, strong ethical and legal frameworks for AI and ML systems are needed to integrate privacy, transparency, and responsibility into the AI development lifecycle. In order to safeguard against minor biases that could harm society as we build increasingly complex synthetic worlds, research must also investigate the consequences of generative AI.

REFERENCES

[1]. Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: Focus on clinicians. *Journal of Medical Internet Research, 22*, e15154.

[2]. Yan, S., Kao, H. T., & Ferrara, E. (2020). Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 1715–1724.

[3]. Ferrara, E. (2023). Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday, 28*.

[4]. Dr. Latha Kiran Krishna Rajendran (Author). (2025). Theranostics: Integrating diagnostic imaging agents and therapeutic drugs into a single multifunctional nano-platform for real-time monitoring of treatment. *Power System Protection and Control, 53*(2). https://pspac.info/index.php/dlbh/article/view/305 https://doi.org/10.46121/pspc.53.2.31

[5]. Rajendran, O. K. (2025). Digital twin frameworks for personalized cancer progression modeling using longitudinal data. *Power System Protection and Control, 53*(4), 486–501.

https://doi.org/10.46121/pspc.53.4.33
[6]. Le, N., Rathour, V. S., Yamazaki, K., Luu, K., & Savvides, M. (2022). Deep reinforcement learning in computer vision: A comprehensive survey. *Artificial Intelligence Review*, 1–87.
[7]. Hemanth Kumar, R. M. (2026). Integrated transcriptomic and machine learning framework identifies a blood-based biomarker signature for anthracycline-induced cardiotoxicity in juvenile cancer survivors. *International Journal of Drug Delivery Technology, 16*(40s), 219–230. https://doi.org/10.25258/ijddt.16.40s.24
[8]. Rajendran, O. K. (2025). Deep learning for cross-modality mapping between histopathology and radiological imaging. *Power System Protection and Control, 53*(3), 313–328. https://doi.org/10.46121/pspc.53.3.21
[9]. Dulac-Arnold, G., Levine, N., Mankowitz, D. J., Li, J., Paduraru, C., Gowal, S., & Hester, T. (2021). Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. *Machine Learning, 110*(9), 2419–2468.
[10]. Dr. Latha Kiran Krishna Rajendran (Author). (2023). Immunotherapy and cell therapy: Developing CAR-T cell therapies and other immune-based treatments for cancer and autoimmune diseases. *Power System Protection and Control, 51*(2). https://pspac.info/index.php/dlbh/article/view/304 https://doi.org/10.46121/pspc.51.2.7
[11]. Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A., Prince, J. L., Rueckert, D., & Summers, R. M. (2021). A review of deep learning in medical imaging. *Proceedings of the IEEE, 109*(5), 820–838.
[12]. Dr. Latha Kiran Krishna Rajendran (Author). (2024). Strict liability or fault-based regimes for AI-caused harm? A doctrinal analysis across common law and civil law systems. *Power System Protection and Control, 52*(4). https://pspac.info/index.php/dlbh/article/view/312 https://doi.org/10.46121/pspc.52.4.13
[13]. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020. *CA: A Cancer Journal for Clinicians, 71*(3), 209–249.
[14]. Dr. Latha Kiran Krishna Rajendran (Author). (2024). Cancer nanomedicine: Utilizing the enhanced permeability and retention (EPR) effect. *Power System Protection and Control, 52*(2). https://pspac.info/index.php/dlbh/article/view/311 https://doi.org/10.46121/pspc.52.2.12
[15]. Dr. Latha Kiran Krishna Rajendran (Author). (2024). Mechanisms driving immunotherapy resistance in colorectal cancer liver metastases. *Power System Protection and Control, 52*(1). https://pspac.info/index.php/dlbh/article/view/303 https://doi.org/10.46121/pspc.52.1.5
[16]. Rajendran, O. K. (2023). Federated radiology AI models for multi-institutional cancer diagnosis without data sharing. *Power System Protection and Control, 51*(4), 38–54. https://doi.org/10.46121/pspc.51.4.5
[17]. Rajendran, O. K. (2023). AI-based radiogenomic models for predicting immunotherapy response in solid tumors. *Power System Protection and Control, 51*(4), 24–37. https://doi.org/10.46121/pspc.51.4.4
[18]. Rajendran, L. K. K. (2026). Integrative pharmacogenomic analysis of drug response heterogeneity across cancer cell lines. *Scientific Culture, 12*(4), 7537–7546. https://doi.org/10.5281/zenodo.12426762
[19]. Van De Schoot, R., et al. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence, 3*(2), 125–133.
[20]. Rajendran, O. K. (2024). Foundation model–driven precision oncology. *Power System Protection and Control, 52*(2), 154–163. https://doi.org/10.46121/pspc.52.2.14
[21]. Rajendran, O. K. (2024). Self-supervised multimodal learning for early cancer detection across imaging and genomics. *Power System Protection and Control, 52*(4), 167–178. https://doi.org/10.46121/pspc.52.4.14
[22]. Rajendran, L. K. K. (2026). Evaluating cancer-related risk factors across multisystem health. *Scientific Culture, 12*(4), 7520–7527. https://doi.org/10.5281/zenodo.12426760
[23]. Rajendran, L. K. K. (2026). Impact of treatment modalities on fertility, sexual function, and psychological outcomes in testicular cancer survivors. *International Journal of Drug Delivery Technology, 16*(30s), 447–453. https://doi.org/10.25258/ijddt.16.30s.43
[24]. Rajendran, L. K. K. (2026). From prediction to practice: Clinical decision support for bevacizumab risk stratification. *International Journal of Drug Delivery Technology, 16*(30s), 414–429. https://doi.org/10.25258/ijddt.16.30s.40
[25]. Rajendran, L. K. K. (2026). Survival and therapy recommendation system for non-small cell lung cancer. *International Journal of Drug Delivery Technology, 16*(30), 430–438.

https://doi.org/10.25258/ijddt.16.30.41
[26]. Rajendran, L. K. K. (2026). Interpretable machine learning for early mortality prediction in AML. *International Journal of Drug Delivery Technology, 16*(40s), 231–241. https://doi.org/10.25258/ijddt.16.40s.25
[27]. Rajendran, L. K. K. (2026). Machine learning–driven cancer risk stratification: Systematic review. *International Journal of Drug Delivery Technology, 16*(40s), 242–253. https://doi.org/10.25258/ijddt.16.40s.26
[28]. Zhou, S. K., et al. (2021). Deep reinforcement learning in medical imaging: A literature review. *Medical Image Analysis, 73*, 102193.
[29]. Schwartz, R., et al. (2022). Towards a standard for identifying and managing bias in artificial intelligence. *NIST Special Publication 1270*.
[30]. Ferrara, E. (2023). GenAI against humanity. *arXiv:2310.00737*.
[31]. Ferrara, E. (2023). The butterfly effect in artificial intelligence systems. *arXiv:2307.05842*.
[32]. Gebru, T., et al. (2021). Datasheets for datasets. *Communications of the ACM, 64*, 86–92.
[33]. Crawford, K., & Paglen, T. (2021). Excavating AI. *AI & Society, 36*, 1105–1116.
[34]. Ezzeldin, Y. H., et al. (2023). FairFed: Enabling group fairness in federated learning. *AAAI Conference on Artificial Intelligence*.
[35]. Nicoletti, L., & Bass, D. (2023). Humans are biased: Generative AI is even worse. *Bloomberg Technology + Equality*.
[36]. Cirillo, D., et al. (2020). Sex and gender differences and biases in AI for biomedicine. *npj Digital Medicine, 3*, 81.
[37]. Huang, J., et al. (2022). Evaluation and mitigation of racial bias in clinical machine learning models. *JMIR Medical Informatics, 10*, e36388.
[38]. Park, J., et al. (2022). Fairness in mobile phone-based mental health algorithms. *JMIR Formative Research, 6*, e34366.
[39]. Ricci Lara, M. A., et al. (2022). Addressing fairness in AI for medical imaging. *Nature Communications, 13*, 4581.
[40]. Yan, S., Huang, D., & Soleymani, M. (2020). Mitigating biases in multimodal personality assessment. In *Proceedings of the International Conference on Multimodal Interaction*.
[41]. Mehrabi, N., et al. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*, 1–35.
[42]. Chauhan, P. S., & Kshetri, N. (2022). The role of data and AI in driving diversity, equity, and inclusion. *Computer, 55*, 88–93.