Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Survey of Bias Mitigation and Fairness Techniques in Responsible Artificial Intelligence Systems

Harish Kant Pathak
Independent Researcher
Corresponding Author: Harish Kant Pathak, **E-mail**: harish.pathak@gmail.com

| ABSTRACT

The use of artificial intelligence (AI) systems is steadily increasing across critical areas such as criminal justice, education, healthcare, and finance. The systems provide efficiency improvements which seem to deliver fairness yet they maintain the biases which exist in their design and training data. The rapid adoption of AI in important decision-making processes has raised serious concerns about bias and fairness. This paper examines bias in AI systems through its two distinct forms explicit and implicit, while showing how algorithms reproduce existing social inequalities. The research investigates essential fairness metrics together with multiple bias categories which develop during the different stages of the AI lifecycle starting from data acquisition and model building until user engagement. The paper examines how biased AI affects various real-world sectors including healthcare and finance, employment and the justice system. The document emphasizes how statistical analysis together with explainability techniques helps organizations find bias problems while building transparent systems which users can trust. Finally, the study reviews various mitigation approaches pre-processing, in-processing, and post-processing, along with broader fairness concepts such as accountability, intersectionality, group fairness, and individual fairness. Lastly, it identifies some of the most important challenges, such as conflicting fairness criteria, information constraints, scalability challenges, and ethical considerations, where context-aware AI development is necessary to guarantee equitable results of various populations.

| KEYWORDS

Bias Detection Mitigation, Fairness in AI, Fairness Metrics, Artificial Intelligence, social inequalities, AI Accountability, Explainable AI (XAI).

| ARTICLE INFORMATION

ACCEPTED: 01 June 2026 **PUBLISHED:** 22 July 2026 **DOI:** 10.32996/jcsts.2026.8.8.12

1. Introduction

Artificial Intelligence (AI) can be traced to 1950s when the first pioneers began to ask the question of whether machines were intelligent enough to behave in a smart way, and this was the first introduction of the term Artificial Intelligence, and the first AI programs were developed. Over the following decades, AI has evolved to become a fundamental technology in industries with progresses including neural networks with backpropagation, development of deep learning (DL) models, including data science, convolutional neural networks (CNN), and, most recently, generative AI and large language models (LLMs) [1][2]. Though such developments have brought about a high level of productivity and financial benefits, it has also generated serious issues about the model transparency, biasness, and lack of transparency, and the lack of a common standard [3]. As a result, both society and regulations are placing greater pressure to make AI systems reliable, ethical and human-value-centered. Responsible AI has consequently become an inclusive framework that is intended to control the risks and challenges inherent in deployment of AI, concentrating on elements of responsibility, openness, and justice, and adherence to the changes in legal and ethical rules that allow AI to benefit society [4]. AI has emerged as a prototype technology that is transforming the way important industries, including healthcare, banking, employment, law enforcement, and face recognition, are managed [5]. AI systems provide notable efficiency and scalability advantages through making it possible to automate, perform predictive analytics, and make data-driven insights [6]. Nevertheless, there is increased evidence that these systems may be biased and discriminatory in nature, and

commonly express historical inequalities inherent in their training data [7][8]. These prejudices can have unequal results, disadvantageously impact the underserved or underrepresented groups and cause severe ethical, social, and legal issues.

The biases of AI may be introduced at several points in machine learning (ML) cycle, such as gathering data, labeling, model architecture, and deployment. Initial beliefs that AI systems are objective have been refuted through research that reveals that algorithms can be trained and perpetuate the existing prejudices in society [9]. These risks have become even more significant with emergence of DL and large-scale data, as the intricate models can potentially learn to identify subtle patterns that represent discriminatory associations [10][11]. The presence of biased AI systems in lending and hiring and medical treatment and criminal justice processes at banks creates unfair outcomes which develop into trust issues among people who cannot access essential resources. The scientific community has created various techniques to identify and correct biases which safeguard the integrity and transparency and accountability of artificial intelligence systems [12][13]. The system includes four types of interventions which include data-level and algorithm-level and post-processing and governance system interventions [14]. The system contains multiple flaws which include the need to define fairness within specific contexts and the need to assess performance through fairness-performance trade-offs and the system's ability to generalize across different datasets and the system's failure to consider intersectional biases which affect people who have multiple protected characteristics.

1.1 Organization of the paper

The rest of the paper is organized as follows. Section 2 explains the principles of bias and fairness in AI, including the main concepts, measures, indicators, sources, and effects. Section 3 focuses on the detection and mitigation of bias, covering statistical approaches, explainability tools, and lifecycle-based interventions. Section 4 discusses the key challenges and limitations in achieving fairness in real-world AI systems. Section 5 reviews recent research in the field, highlighting major contributions as well as existing gaps. Finally, Section 6 concludes the paper by summarizing the main findings and suggesting directions for future research on responsible AI.

2. Foundations of Bias and Fairness in AI

The principles of bias and fairness in AI include fundamental requirements for equitable system design as well as the causes, traits, and justifications of prejudice as impacted by data, algorithms, and the implementation context. Discrimination comes in various forms depending on the field, thus requiring area-specific interventions. Biased models may provide unequal access in crucial areas such as criminal justice, education, recruiting, healthcare, and finance, and computer vision or cause increased error rates in some groups [15][16][17][18]. To solve these problems, various datasets, explainable models, transparency, regulatory checks, and fairness-conscious assessment should be used in order to make AI deployment responsible and inclusive [19].

It is sometimes very difficult to distinguish bias from expected behavior in AI since bias can be either deliberate (explicit bias) or non-intentional (implicit bias). To understand bias within an AI system, must account for two dimensions of bias. First, bias in AI applications must be perceived as a techno-social mirror reflecting societal bias [20]. In this context, research on AI bias raises questions about how to understand and address societal biases, as well as what to ask an AI application regarding bias management. Second, when AI makes errors, it's critical to identify the source of bias, whether it comes from data or the design and take methodical action to lessen it. Understanding technical definition of prejudice is essential to eliminating bias in development and use of AI systems.

The primary fairness measures used to determine whether AI systems produce predictions and outcomes that are equitable for both protected and unprotected populations are displayed in Table 1.

Table 1. Shows the primary metrics of fairness

Fairness Metric	Summary
Statistical (Demographic) Parity	Requires that protected and unprotected groups have an equal chance of being assigned a positive projected class.
Disparate Impact	Represents proportion of positive predictions between groups that are protected and those that are not.
Calibration	Requires that all groups' anticipated probability or risk ratings match actual results in an equal manner.
Predictive Parity	Requires equal positive predictive values for protected and unprotected groups.
Error Rate Balance	Represents the ratio of inaccurate predictions to all forecasts for both protected and unprotected groups.
Equal Opportunity	Requires equal TPR (preferred outcomes) across protected and unprotected groups.
Equalized Odds	Requires equal TPR and FPR across protected and unprotected groups.

This is necessary to make AI models unbiased. It's critical to recognize that bias may inadvertently exist in data used to train AI systems. Data collection and selection play a major role in determining whether bias is present in AI systems. To counteract prejudice in AI systems, sources of bias and its effects on AI decision-making must be identified. At different phases of the AI development life cycle, bias may become deeply embedded.

2.1 Types of Bias and Fairness in AI

Table 2 enumerates types of bias and fairness based on AI, providing a descriptive definition of each, along with real-world examples discussed below:

Table 2. Common Types of Bias and Fairness in AI Systems

Category	Type	Description	Example(s)
Bias	Data Bias	Occurs when training data is not typical of the intended audience or use case.	Dark-skinned faces are difficult for facial recognition algorithms that were predominantly trained on light-skinned faces.
	Algorithmic Bias	Stems from decisions made when creating and applying AI systems.	The algorithm used to scan resumes favors phrases related to a certain gender.
	Interaction Bias	Arises from way users interact with AI systems.	Accents and dialects that are underrepresented in training data are a challenge for voice assistants.
	Historical Bias	Reflects cultural prejudices from the past in the training data.	A hiring tool that reflects the historical disparity between genders in occupations.
	Sampling Bias	Results from demographic subgroups sampled non-randomly.	Medical diagnosis AI trained on urban hospital data is performing poorly for rural patients.
	Confirmation Bias	Propensity to favor data that supports preconceived notions.	Recommendation algorithms magnify users' preexisting opinions without considering other viewpoints.
	Automation Bias	Over-reliance on automated decision outputs.	Humans are accepting AI results without critical evaluation.
	Availability Bias	Overestimation of the importance of frequently reported or memorable events.	News-trained AI overestimating frequency of rare but newsworthy events.
Fairness	Group Fairness	Guarantees various demographic groups receive proportionate or equal treatment. includes equitable opportunity, disproportionate maltreatment, and demographic parity.	Equal misclassification rates, equal true positive rates, and equal outcome distribution among demographic groupings.
	Individual Fairness	Make sure that comparable people, regardless of group membership, receive comparable results.	Using metrics based on resemblance or distance to treat similar people fairly.
	Counterfactual Fairness	Confirms equity even in hypothetical scenarios involving different protected attributes.	Even if characteristics such gender or race were different, an individual's decision was the same.
	Procedural Fairness	The decision-making process is transparent and equitable.	Transparent and explainable AI decision workflows.
	Causal Fairness	Make sure system doesn't reinforce prejudices or previous injustices.	Designing models that remove causal pathways linked to discriminatory outcomes.

2.2 Impacts of Bias in AI Systems

The rapid development of AI has provided substantial benefits across various industries, yet it has also brought serious dangers related to algorithmic bias. Discrimination of oppressed groups of people and limitations in access to vital services like healthcare, finance, employment and justice can be the results of bias AI systems, which tend to enhance and multiply the existing social disparities [21][22][23]. This type of bias can further gender stereotypes, fail to represent minorities, and deliver unfair results depending on race, ethnicity, skin tone, socioeconomic class, or physical attractiveness. As an illustration, unfair criminal justice systems can lead to a more severe outcome for many communities, whereas biased credit rating or hiring software can restrain the chances of the disadvantaged ones. Generative and computer vision models can also perpetuate stereotypes by correlating leadership with men or making criminals black individuals. These consequences hurt a person by withholding services, diminishing chances, or making false judgments, as well as perpetuating non-discriminatory discourse and undermining the trust in technology of the population. In order to ensure that everyone is treated equally and receives the same

outcomes, bias management must be taken seriously while developing and utilizing AI systems, as they have the potential to become a part of society.

2.3 Sources of Bias in AI

The different phases of ML process, AI bias may arise from a variety of causes, including data collecting, algorithm development, and human interactions. Data bias, algorithmic bias, and user bias are only a few of the many biases in AI covered, along with examples of each. Due to poorly collected or non-representative data, ML models yield skewed outcomes during training. This is referred to as data bias. This could be because material has been received via biased channels, is partial, omits crucial information or has errors. On other hand, algorithmic bias occurs when outputs of ML algorithms exhibit intrinsic biases. When algorithms are created using prejudiced standards or assumptions, it might occur. Users can inadvertently or deliberately implant their preconceptions or biases into AI systems. This is referred to as user bias. Figure 1 depicts the ML pipeline, focusing on the key steps of problem formulation and data collection, preprocessing, development, and implementation, as well as potential points of bias at each step.

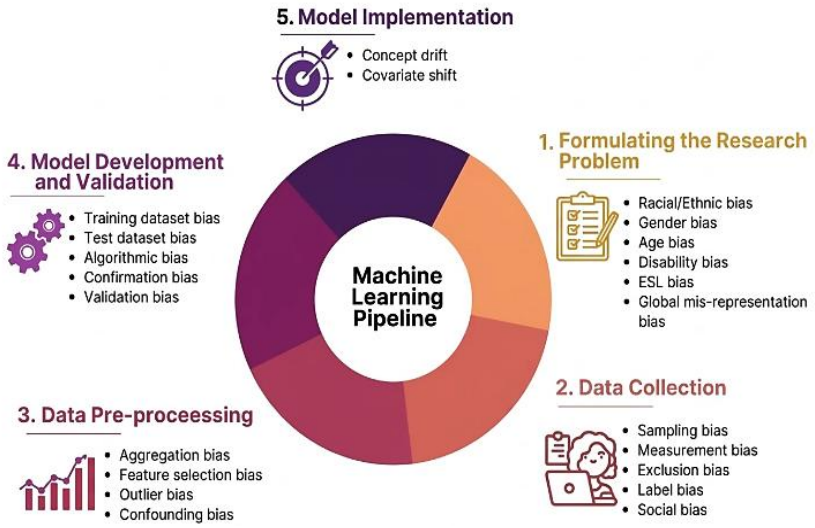

Figure 1. Sources of bias that may contribute to predictive variances of the AI

To overcome these biases, a number of approaches have been proposed, which include augmenting the datasets, bias-aware algorithms, and user feedback systems [24]. Including additional and different types of data in training datasets can make them more representative and less biased. Bias - aware algorithms attempt to lessen the influence of bias on system outputs by accounting for various types of bias. One method to find and correct biases in the system is to ask users for input. To reduce bias in AI systems, new strategies and approaches are being developed, and studies are underway in this field. It is important to continue researching and developing such strategies to come up with more just and equal AI systems to all users. Bias Mitigation and Fairness Techniques in Responsible AI Systems.

2.4 Fairness in AI System

The issue of fairness in AI is a contentious one that has been extensively explored in both the academic and business realms. The essential tenet of AI fairness is that AI systems are devoid of prejudice and discrimination, which is not always easy, since various forms of bias can emerge in AI systems. The literature suggests a number of different forms of fairness: Individual, collective, and counterfactual fairness [25]. Despite the close relationship between prejudice and fairness, there are significant differences between them, such as the fact that prejudice can be inadvertent, whereas fairness is inherently a purposeful, intentional objective. Achieving fairness in AI requires researchers to study existing situations together with their associated groups who have a stake in the results. The real-world applications of AI systems show that implementing fairness measures brings multiple advantages to their operations.

The principle of individual fairness demands equal treatment for comparable individuals from different groups. Group fairness aims to achieve equitable results for multiple demographic groups. Statistical parity requires decision-making processes to remain unaffected by protected characteristics which include gender and race. The two concepts of equal opportunity and equalized odds require all groups to show matching error rates which includes TPR while equalized odds require matching both TP and FP. Counterfactual fairness ensures that the outcome for a person remains unchanged when their protected characteristic is altered.

3. Bias Detection and Mitigation Techniques

AI systems use statistical methods and explainability tools as their primary methods for bias detection and reduction. The field standard methods are pre-processing, in-processing, and post-processing interventions. The discussion addresses essential fairness principles which include individual fairness and group fairness together with transparency and accountability and explainability and intersectionality and their practical applications and the obstacles researchers face while developing ethical AI technologies.

3.1 Methods for Detecting Bias in AI

The growing agreement between lawmakers and business executives and the general public shows that they all want to decrease bias present in AI systems. It has resulted in an explosion of detection techniques and techniques to increase fairness, but debate about bias continues to be divided, especially in the systematic assessment of these techniques [26]. A stringent diagnostic is required to identify between the good solutions and the ones that need improvement. To understand reasons for bias in AI systems, it is essential to identify most important factors causing it, as shown in Figure 2.

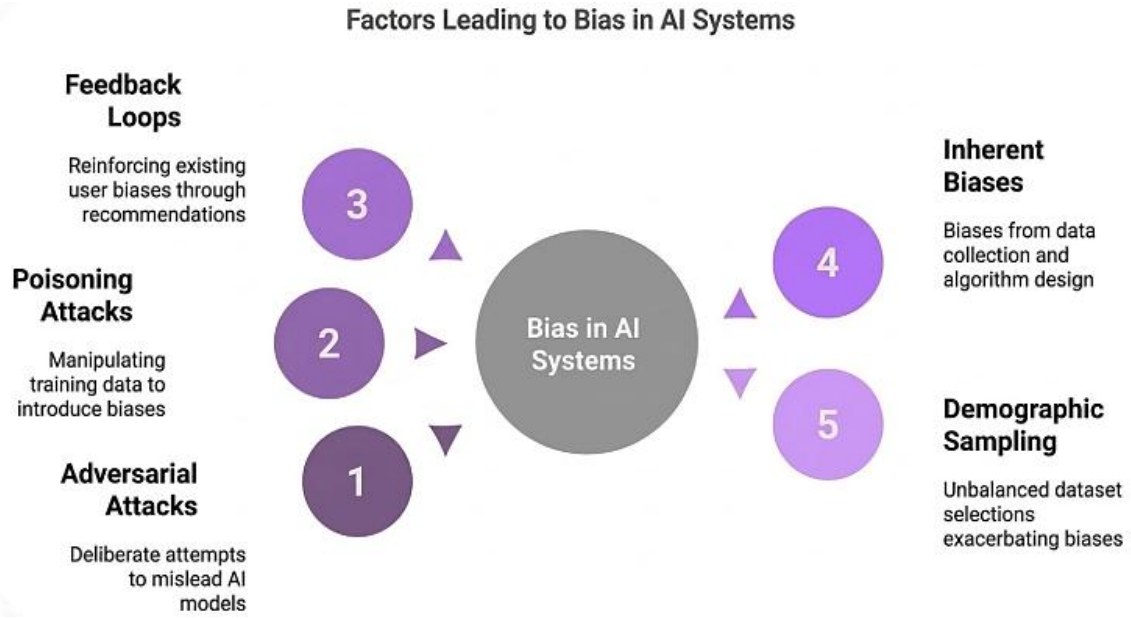

Figure 2. Factors leading to bias in AI systems

3.1.1 Statistical Method

This considers statistical techniques, particularly quantitative techniques for gauging fairness, for identifying bias in AI. The decisions of metrics, statistical modelling, and data selection are only a few examples of how academics and professionals use these methods. There is a description of a system of methods, using case studies, indicating potential problems and the requirement for a more statistical approach [27]. To enhance transparency and ensure ethical, transparent reporting of the statistics, it is proposed to use visualizations. The first step in mitigation is understanding the manifestation of the bias. The evaluation of the fact that the groups of protection are negatively impacted can be quantified using statistical tools such as disparate impact analysis, which was carried out by 96 Proceedings of 13th International Conference on Applied-on-Applied Innovations in IT (ICAIIT), June 2025. The degree of bias shows greater significance than the existence of bias itself, which requires careful selection of metrics that need proper baseline comparisons. The reporting needs to extend beyond standard significance tests to include intersectional analysis, which helps avoid oversimplifying the results. The use of visual displays together with conventional statistical techniques enables organizations to enhance their communication capabilities, which leads to improved decision-making results.

3.1.2 Explainability and Interpretability Techniques

The hidden nature of AI systems prevents people from understanding them which creates challenges for finding and fixing bias issues. The existing biases in technology emerge from two sources because complex systems generate new solutions but they create doubts about their ability to reflect actual world conditions. Explainability assessment becomes challenging because AI system assessment requires expert evaluation of both fairness and system accuracy. AI model understanding and control development requires organizations to implement valid explainability and interpretability methods. Studies suggest that greater transparency helps build trust and enables earlier detection of bias. Shapley values, permutation feature importance, and locally

interpretable model-agnostic explanations serve as techniques for describing model behavior, yet their output does not guarantee consistent results. The methods provide valuable insights into model operation, yet they fail to show all aspects of internal model functioning. The complete technical explanation of AI systems should remain secondary to establishing straightforward methods that show how AI systems reach their decisions.

3.2 Mitigation Strategies for Bias in AI

AI development has brought new technologies to different industries yet the criminal justice system and the healthcare sector still struggle with problems that stem from biased systems which create unfair outcomes. AI systems developed with biased data training create systems that perpetuate existing societal biases which results in discriminatory outcomes that particularly impact marginalized communities [28]. Researchers have increased their efforts to study AI bias detection and reduction methods because people now recognize the risks that AI systems can create especially for electronic health record (EHR) data-based models [29]. Researchers examine how AI impacts society because its application in both public and private sectors has become a significant research topic. The understanding of these implications becomes critical because (AI) systems now play a major role in essential decision-making activities. Figure 3 shows an AI lifecycle which describes feedback-based systems by displaying all steps from data collection through model construction and evaluation to system deployment while focusing on reducing bias and achieving system clarity and continuous monitoring.

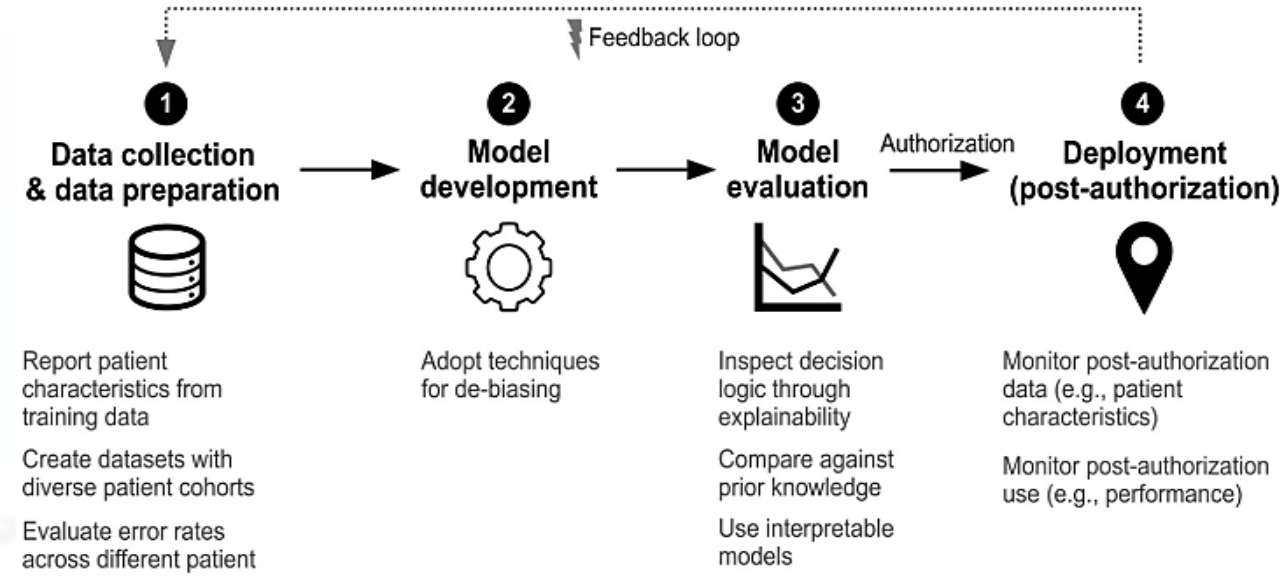

Figure 3. Bias mitigation strategies

3.2.1 Pre-Processing Techniques

The most effective method for addressing machine learning model bias requires organizations to implement preventive solutions during their initial development stage instead of waiting until after they complete model development. AI technology introduces innovative solutions, yet its use in social sectors requires organizations to establish rigorous testing standards for demonstrating equitable treatment [30]. The Newsela quizzes contain gender bias, which generates unfair outcomes because the model predictions provide different levels of support to distinct population segments. Organizations need to create methods for managing bias because complete bias elimination proves unachievable, and bias presents operational risks. AI systems develop bias because their training data fails to adequately represent historically marginalized groups, which results in discriminatory outcomes during AI model predictions. The fact that datasets that were assembled based on several sources and views complicates the quantification of bias because it is hard to measure bias on an exact scale. Studies have emphasized the requirement to examine prejudice, power, and data in order to develop more moral AI systems.

3.2.2 In-Processing Approaches

It is essential to mitigate bias in AI throughout data collection, model development, implementation, and decision-making. Interventions in real time, such as altering outputs and filtering biased inputs are the focus of in-processing techniques, which are specifically used when end users are not developers or for off-the-shelf AI. Two important measures are to add fairness constraints and employ regularization methods in training that allow decisions to be made without being influenced by data-based biases. While the training process is actively guided toward more fair results through regularization, fairness restrictions ensure that group inequalities do not impact model predictions [31][32]. The different algorithms incorporate fairness in AI

models, but they vary in terms of their conceptualizations. Mitigation of bias should not only be confined to modeling but also to all phases of the lifecycle of an AI system, However, understanding how prejudice is produced and maintained requires collaboration amongst parties. It is essential to precisely identify bias, assess its impacts, and establish monitoring systems to combat it effectively. Decoupled model monitoring is a proposed framework based on AI bias audits that would improve accountability and openness throughout the development and application of an AI system.

3.2.3 Post-Processing Methods

Post-processing modifies model outputs with the aim of increasing fairness without retraining. Demographic balancing techniques like re-weighting of predictions, that make all errors equal, and calibration are used to balance the results across demographic lines. Although these approaches enhance fairness, post-deployment follow-up is crucial as bias is a component of most real-life issues, and even the most effective interventions can be associated with trade-offs. Studies on post-deployment fairness tracking have increased, but there is little research on how this can be put into practice.

3.3 Mitigation Strategies for Fairness in AI

Making sure AI makes fair decisions is becoming more crucial as its use grows. Because use of AI in vital fields such as healthcare, economics, and law can profoundly affect people's lives, it is essential that these systems make impartial and fair decisions. Many strategies have been created to solve this issue, such as individual and collective fairness [33]. These methods do, however, have certain drawbacks and difficulties, such as trade-offs between various forms of fairness and issue of defining fairness itself [34][35]. The section presents two methods which enhance fairness in AI systems but they create difficulties and need further study. A better understanding of these approaches can help in developing AI systems that are more equitable, unbiased, and fair for all users.

Table 3 presents the main approaches to achieving fairness in AI systems which include group fairness and individual fairness together with transparency and accountability and explainability and intersectionality. The study reports the primary restrictions and difficulties which each method brings forth to establish responsible AI systems that maintain fairness between their dual requirements.

Table 3. Approaches to Fairness in AI Systems

Approach	Description	Examples
Group Fairness	Ensures that AI systems treat different demographic groups (e.g., gender, race, ethnicity) fairly and do not systematically discriminate against any group. Achieved through data or model adjustments.	1. Re-sampling techniques to balance datasets. 2. Techniques for pre-processing or post-processing to modify model results.
Individual Fairness	Ensures that comparable people, regardless of group membership, obtain comparable results. emphasizes individual fairness.	Counterfactual fairness guarantees that decisions are made regardless of characteristics like gender or ethnicity.
Transparency	Enables consumers to see and comprehend the decision-making process of AI systems.	Providing users with insight into how decisions are generated.
Accountability	Holds companies or developers accountable for damage caused by AI systems.	Developers being accountable for unfair or discriminatory decisions.
Explainability	Ensures AI decisions are interpretable and understandable to humans.	Providing precise justifications for model forecasts or suggestions.
Intersectionality	Considers how overlapping identity dimensions (e.g., race, gender, socioeconomic status) jointly affect outcomes.	Designing systems that evaluate combined effects of multiple identity factors.

4. Challenges and Limitations

The technical, ethical, organizational, and regulatory challenges to bias mitigation and fairness methods in responsible AI systems are challenging, such as vague definitions, data constraints, incompatibility of fairness objectives, and scalability, contextual dependencies and unintended consequences, among others, which impede them in effective implementation in the real world.

4.1 Challenges and Limitations of Bias in AI

The procedure for lowering bias in ML and AI systems is complex and beset by a number of obstacles at different phases of model development and implementation [36]. These fall under the categories of organizational, ethical, technical, and regulatory impediments. One of the biggest obstacles to applying bias mitigation methods to responsible AI systems is the following challenge.

4.1.1 Ambiguity in Defining and Measuring Bias

Context-dependent bias is situation-specific, domain-specific, and application-specific. The lack of standard definitions and metrics to assess bias across AI systems complicates the process of identifying, quantifying, and comparing bias within AI systems.

4.1.2 Data Quality and Representation Issues

Majority of biases are based on unbalanced, incomplete or historically biased datasets. The acquisition of high-quality, diverse and representative data is expensive and limited by privacy, access, and societal considerations.

4.1.3 Fairness–Performance Trade-off

Increasing fairness tends to reduce model accuracy, efficiency, or utility. Proficiency in balancing between fair results and good predictive performance has been a key technical problem, particularly in high-stakes applications.

4.1.4 Conflicting Fairness Criteria

The various concepts of fairness (demographic parity, equal opportunity and equalized odds) can be inconsistent with each other. The decision on which criterion to put into priority relies on ethical, legal and contextual issues.

4.1.5 Scalability and Computational Complexity

There is a significant cost and complexity of deployment because bias mitigation methods, especially in large or real-time AI systems, demand significant amount of computational resources, continuous monitoring, and regular retraining.

4.1.6 Organizational and Regulatory Barriers

The knowledge gap, the absence of motives, unwillingness to change and adopt systems, and universal global laws discourage the integration of fairness-conscious practices in any industry.

4.2 Challenges and Limitations in AI for Fairness

Even though the known methods of fairness have shown positive prospects of curbing biasness within an AI system, they have a number of inherent limitations that limit their application in, real-life applications. The key obstacles are as shown below:

4.2.1 Trade-offs Between Different Fairness Notions

The primary shortcoming of fairness methods is the trade-off aspect in favor of competing fairness goals. An example is the group fairness procedures that seek to achieve fair results among the demographic groups, but can result in unfair treatment of members within those groups [37]. Individual fairness, on other hand, concentrates on the similarity of individuals and how they can be treated like others, but does not seek to deal with systemic inequality in favor of whole populations.

4.2.2 Difficulty in Defining Fairness

Fairness is a social construct and one that is in a state of change, and various parties have different interpretations of what fair behavior is. The cultural norms, legal standards, and ethical values vary in different societies and evolve as time passes, which is why it is rather hard to create AI systems that please all sides at the same time.

4.2.3 Context Sensitivity and Selection of Appropriate Techniques

It is hard to decide what fairness approach would be most appropriate in this or that application. The methods that are working in one area can be inadvisable or inadequate in another area. The process of selecting fairness constraints and their calibration needs to rely on experts who possess deep knowledge of the domain together with an ethical evaluation process that requires all involved parties to contribute their expertise.

4.2.4 Limitations of Statistical Fairness Methods

The current fairness assessment techniques need statistical measurements which require particular statistical assumptions for their operation [38]. The tools tend to simplify social realities which require complex analysis while they fail to capture complete human behavior patterns and structural differences and the root causes of biased outcomes.

4.2.5 Lack of Consideration for Intersectionality

Most fairness metrics investigate bias through one protected characteristic which includes either gender or race but they fail to consider how different identities intersect with one another. When these overlapping factors are disregarded, concealed discrimination of subgroups that have gone benign under the usual measures can be realized.

4.2.6 Risk of Unintended Consequences

Mitigation of bias can also create unintended effects of bias, decrease transparency, or have other detrimental consequences. To use an example, any adjustments meant to equalize the results of a predictive system may cause a change in decision thresholds that cause disproportionately certain groups of people or misrepresent the actual processes taking place in the world.

4.2.7 Transparency, Accountability, and Explainability Challenges

To achieve fairness, it may be necessary to have AI systems that are clear and comprehensible. However, complicated models may be difficult to understand, and it is difficult to hold the individual to account the unjust decision. Whether one or the other take responsibility in the case of harm, whether it is the developers, organization or data sources, is a thorny question that has never been answered.

5. Literature Review

The detection of bias, assessment of fairness, and the ways to mitigate it in AI systems have been studied in several studies that cut across various sectors, including hiring, healthcare, finance, and overall ML applications. These papers show that fairness metrics, debiasing strategies, and governance systems are effective, and that there are practical constraints and a need for consistent, practical solutions, as summarized in Table 4.

M. Loganathan, H. Sharifzadeh, and A. Keivanmarz (2025) present a methodical strategy for improving AI fairness utilizing Adult Census Income dataset, which is regarded as a benchmark for fairness evaluation, and IBM AI Fairness 360 (AIF360) toolbox. SPD and DI are used to evaluate fairness, and methods of bias mitigation include ADB, and CEO. The paper also considers effects of SMOTE variants on fairness. The framework is evaluated on several models, such as RF, XGBoost, LightGBM and TabNet [39].

H. Ashtekar, S. Girme, and R. Khopade (2025) Discuss evaluation metrics and present the different forms of bias that might occur in recruiting process using ML, and classify bias mitigation strategies into three categories: pre-processing, training, and post-processing interventions. Real-world studies indicate that algorithmic-based hiring can help to make the process more efficient, yet it can also contribute to increasing inequalities unless there are strict control, frequent audits, and clear regulations. The paper concludes with the recommendations of applying AI ethically to transform workplaces to become fairer and more positive [40].

A. Trigo, N. Stein, and F. Paulo Belfo (2025) present a systematic literature review of technical, feasible, and practicable ways to enhance fairness in AI based on various perspectives: fairness metrics, research field, datasets, algorithm utilized in the study, and when to intervene (pre-processing, processing, or post-processing). The primary value added by the paper is to reach a common ground when it comes to the methods that should be employed to enhance AI fairness is ability of AI systems to make decisions without prejudice or discrimination [41].

E. Ferrara (2023) provides a concise, broad overview of AI fairness and bias, their origins, effects, and ways to avoid them. They assess the total consequences of biased AI systems and consider bias-causing factors such data, algorithms, and human decision biases, and consider how it supports inequalities and strengthens harmful stereotypes. They discuss different mitigation measures suggested and refer to the ethical implications of their application and interdisciplinary collaboration as key to their efficiency. Define AI bias and its various forms, and explain the adverse effects of AI bias on people and society using a systematic literature review, which encompasses various academic fields. Further give a summary of existing methods of reducing AI bias, such as model selection, post-processing, and data pre-processing [34].

P. Chen, L. Wu, and L. Wang (2023) study delivers an extensive examination of fairness problems existing in artificial intelligence systems which includes their definition as well as their historical development and current state. The research study investigates two fairness issues through practical case studies and recent developments in the field, while identifying bias analysis methods and fairness-aware training techniques as crucial advancements that drive progress in this area. The article provides a thorough explanation of each technique's idea, application, traits, and use cases. The study examines methods to lessen prejudice and enhance justice in AI systems, evaluates problems and fixes for practical AI fairness applications, and suggests future lines of inquiry. Additionally, this study provides a thorough comparison analysis of the various techniques, using state-of-the-art research data to clarify how they differ in terms of their characteristics, strengths, and weaknesses [42].

J. Huang, G. Galal, M. Etemadi, and M. Vaidyanathan (2022) performed a scoping review to discuss measures that might be utilized to improve algorithmic fairness in clinical ML and to describe how racial bias of ML has been measured. PRISMA Extension for Scoping Reviews was used to conduct the scoping review. 12 studies were included from the 635 records identified in a literature search using PubMed, Scopus, Embase, and Google Scholar. ML was used in a number of applications, such as diagnosis, outcome prediction, and clinical score prediction, using clinical factor databases, images, diagnostic studies, and clinical text [43].

Table 4. Literature Review on Fairness and Bias Mitigation in AI Systems

Authors (Year)	Study On	Applications	Strategies	Key Contributions	Research Gaps
Loganathan et al. (2025)	AIF360-based framework for fairness improvement	Income prediction (Adult Census dataset)	Reweighing, Calibrated Equalized Odds (CEO), Adversarial Debiasing (ADB), and SMOTE variations	Demonstrates the effectiveness of multiple bias mitigation techniques across diverse ML models	Limited validation on real-world, domain-specific datasets
Ashtekar et al. (2025)	Bias in AI-based hiring systems	Recruitment and workforce selection	Pre-processing, in-processing, post-processing, governance, and audits	Highlights risks of algorithmic hiring amplifying inequalities without oversight	Lack of standardized evaluation frameworks for hiring AI
Trigo et al. (2025)	Systematic review of fairness techniques	Cross-domain AI applications	Classification by fairness metrics and intervention stage	Establishes a unified taxonomy of fairness approaches and methods	Need for practical deployment guidelines and real-world testing
Ferrara (2023)	AI bias's causes, effects, and mitigation	Multi-sector societal AI systems	Model selection, post-processing, and data preprocessing, interdisciplinary approaches	Comprehensive synthesis of bias types, impacts, and ethical considerations	Limited empirical validation of mitigation effectiveness
Chen et al. (2023)	Fairness issues and bias analysis in AI	Various real-world AI applications	Bias analysis, fairness training, comparative evaluation methods	Provides a detailed explanation and comparison of fairness techniques	Insufficient focus on scalability and long-term deployment challenges
Huang et al. (2022)	Racial bias in clinical ML	Healthcare diagnosis and outcome prediction	Bias assessment methods using clinical datasets	Determines how often racial bias and fairness techniques are used in clinical ML	A small number of studies and a limited diversity of datasets

6. Conclusion and Future Work

Prejudice and impartiality are the main issues in creation and implementation of AI systems. This paper has discussed basis of bias and its various origins as well as the enormous effects of bias in decision-making in key areas. It emphasized the idea that biases may be developed based on data, algorithms, and interactions between users and tend to reflect the existing inequalities in society. The types of bias and various metrics of fairness were examined to have a structured picture of how discrimination may occur in AI results. Other key detection methods that were reviewed in the paper are statistical methods and explainability tools, as well as mitigation measures that are taken at various levels of the AI lifecycle. However, even with this important progress, there remains the difficulty of having truly fair AI due to the conflicting definitions of fairness, performance trade-off, data limitations, and context dependencies. Thus, to create responsible AI, it is necessary to manage the process, be transparent, engage stakeholders and have regulatory support so that the systems are just, trustful, and the values of society are aligned.

The future study should be directed towards the creation of single universal rules of fairness, environment-conscious mitigation strategies and practical solutions to actual implementation. More focus should be on intersectionality, causal fairness, systems of human in the loop and more robust regulatory structures and interdisciplinary cooperation to be used to create transparent, accountable and social responsible AI systems.

Funding: Please add: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
ORCID iD: 0009-0009-6200-9469

References

[1] Gunasekara, L., El-Haber, N., Nagpal, S., Moraliyage, H., Issadeen, Z., Manic, M., & De Silva, D. (2025). A Systematic Review of Responsible Artificial Intelligence Principles and Practice. Applied System Innovation, 8(4), 97. https://doi.org/10.3390/asi8040097

[2] Bhattacharjee, D. (2025). Design and Evaluation of Deep Generative AI Model for Intrusion Detection in Cyber Threat Monitoring. 2025 7th International Symposium on Advanced Electrical and Communication Technologies (ISAECT), 1–6. https://doi.org/10.1109/ISAECT68904.2025.11318752

[3] Reuel, A., Connolly, P., Meimandi, K. J., Tewari, S., Wiatrak, J., Venkatesh, D., & Kochenderfer, M. (2025). Responsible AI in the Global Context: Maturity Model and Survey. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, 2505–2541. https://doi.org/10.1145/3715275.3732165

[4] Rismani, S., Davis, L., Mingole, B., Rostamzadeh, N., Shelby, R., & Moon, Aj. (2025). Responsible AI measures dataset for ethics evaluation of AI systems. Scientific Data, 12(1), 1980. https://doi.org/10.1038/s41597-025-06021-5

[5] Dholariya, H. N. (2026). GVIF: A Governed Vector Intelligence Framework for AI-Driven Cloud Data Modernization in Regulated Financial Systems. International Journal of Computational and Experimental Science and Engineering, 12(1), 371–386.

[6] Yang, J., Blount, Y., & Amrollahi, A. (2024). Artificial intelligence adoption in a professional service industry: A multiple case study. Technological Forecasting and Social Change, 201, 123251. https://doi.org/10.1016/j.techfore.2024.123251

[7] Hasanzadeh, F., Josephson, C. B., Waters, G., Adedinsewo, D., Azizi, Z., & White, J. A. (2025). Bias recognition and mitigation strategies in artificial intelligence healthcare applications. Npj Digital Medicine, 8(1), 154. https://doi.org/10.1038/s41746-025-01503-7

[8] Murumkar, S., & Tayal, C. (2026). Optimizing Big Data Workflows Using AI and Machine Learning. 2026 IEEE 16th Annual Computing and Communication Workshop and Conference (CCWC), 0550–0556. https://doi.org/10.1109/CCWC67433.2026.11393891

[9] Poon, E. G., Lemak, C. H., Rojas, J. C., Guptill, J., & Classen, D. (2025). Adoption of artificial intelligence in healthcare: survey of health system priorities, successes, and challenges. Journal of the American Medical Informatics Association, 32(7), 1093–1100. https://doi.org/10.1093/jamia/ocaf065

[10] Kartal, E. (2022). A Comprehensive Study on Bias in Artificial Intelligence Systems. International Journal of Intelligent Information Technologies, 18(1), 1–23. https://doi.org/10.4018/IJIIT.309582

[11] Biswal, S. K., Somasundaram, N., Jain, S. S., Phalke, S., & Ulaganathan, S. K. (2024). Methods and apparatus for datacenter monitoring (US11966280B2)

[12] Sharde, A. (2025). Bias and Fairness in AI Systems: A Study of Causes, Impacts, and Mitigation Strategies. International Journal for Research in Applied Science and Engineering Technology, 13(VII), 1409–1415. https://doi.org/10.22214/ijraset.2025.73195

[13] Sony, M. M. A. A. M., Amin, M. Bin, Ashraf, A., Islam, K. M. A., Debnath, N. C., & Debnath, G. C. (2025). Bias in AI-driven HRM systems: Investigating discrimination risks embedded in AI recruitment tools and HR analytics. Social Sciences & Humanities Open, 12, 102082. https://doi.org/10.1016/j.ssaho.2025.102082

[14] Macha, Y., & Pulichikkunnu, S. K. (2024). A Data-Driven Framework for Medical Insurance Cost Prediction Using Efficient AI Approaches. IJRAR, 11(4), 887–893.

[15] Mohammed, K. K. (2025). A Survey on Digital Health Care Data Analysis Techniques for Developing Machine Learning Models. International Journal of Science, Engineering and Technology, 13(5), 1–6.

[16] Mahmud, S., Mishra, D. P., Ramani, G. G., Patel, M. I., Soumik, M. S., & Manivannan, R. (2025). Design of intelligent healthcare IT infrastructure using graph theory, network analysis, and artificial intelligence. International Journal of Applied Mathematics, 38(12s), 2267–2280.

[17] Chaturvedi, V. (2025). Disease Diagnostic Systems based on AI-Applications in Healthcare: Models, Challenges, and Future Directions. International Journal of Emerging Research in Engineering and Technology, 6(4), 207–217, Dec. https://doi.org/https://doi.org/10.63282/3050-922X.IJERET-V6I4P125

[18] Sajja, J. W., & Nerella, A. (2024). Enterprise Finance Reimagined: Harnessing ERP and Data Innovation for Next-Generation Value Creation. Computer Fraud and Security, 2024(4), 17–26. https://doi.org/10.52710/cfs.743

[19] Pareek, C. S. (2025). Testing Ethical AI in Life Insurance: Ensuring Fairness, Transparency, and Accountability in Automated Decisions. International Journal of Future Engineering Innovations, 2(2), 43–50. https://doi.org/10.54660/IJFEI.2025.2.2.43-50

[20] Afreen, J., Mohaghegh, M., & Doborjeh, M. (2025). Systematic literature review on bias mitigation in generative AI. AI and Ethics, 5(5), 4789–4841. https://doi.org/10.1007/s43681-025-00721-9

[21] Prajapati, N. (2025). The Role of Machine Learning in Big Data Analytics: Tools, Techniques, and Applications. ESP Journal of Engineering & Technology Advancements, 5(2), 16–22. https://doi.org/10.56472/25832646/JETA-V5I2P103

[22] Ramkumar, K., Sagunthala, R. R., Professor, A., Nerella, A., Kilaru, S., & Gladson Battu, G. (2025). A Temporal Graph Neural Network Approach for Deep Fraud Detection in Real-Time Financial Transactions. 16th International IEEE Conference on Computing, Communication and Networking Technologies (ICCCNT), 1–5.

[23] Patel, M., & Chaturvedi, V. (2025). A Survey on Artificial Intelligence Techniques for Disease Prediction in Healthcare. ESP Journal of Engineering & Technology Advancements, 5(4), 201–210. https://doi.org/10.5281/zenodo.19509023

[24] Dattangire, R., Vaidya, R., Biradar, D., & Joon, A. (2024). Exploring the Tangible Impact of Artificial Intelligence and Machine Learning: Bridging the Gap between Hype and Reality. 2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET), 1–6. https://doi.org/10.1109/ACET61898.2024.10730334

[25] Pal, V. (2021). Bias Detection and Mitigation in Foundation AI Models: A Human-Centric Approach. TIJER – International Research Journal, 8(2), 1–7.

[26] Belenguer, L. (2022). AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. AI and Ethics, 2(4), 771–787. https://doi.org/10.1007/s43681-022-00138-8

[27] Toorpu, A. R., Vududala, S. K., Nerella, A., & Madupati, B. P. (2025). Hybrid AI Models for Privacy-Preserving Big Data Analytics in Distributed Environments. 2025 Global Conference in Emerging Technology (GINOTECH), 1–8. https://doi.org/10.1109/GINOTECH63460.2025.11076666

[28] González-Sendino, R., Serrano, E., & Bajo, J. (2024). Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making. Future Generation Computer Systems, 155, 384–401. https://doi.org/10.1016/j.future.2024.02.023

[29] Warrier, A. (2020). Machine Learning-Enhanced Data Quality Validation for Electronic Health Record Integration. International Journal on Science and Technology, 11(1). https://doi.org/10.71097/IJSAT.v11.i1.8387

[30] González Sendino, R., Serrano, E., Bajo, J., & Novais, P. (2024). A Review of Bias and Fairness in Artificial Intelligence. International Journal of Interactive Multimedia and Artificial Intelligence, 9(1), 5–17. https://doi.org/10.9781/ijimai.2023.11.001

[31] Mali, S. U., & Harshali Ashtekar, Siddhipriya Girme, R. K. (2025). A Survey Mitigation of Bias and Fairness in Machine Learning using LLMs towards Ethical AI. International Research Journal of Modernization in Engineering Technology & Science, 07(11), 2240–2242. https://doi.org/10.56726/IRJMETS85230

[32] Phalke, S., Athave, Y. D., & Ilag, B. N. (2025). A Multi-Layered Approach to IT Infrastructure Governance and Compliance-Security, Hardening, and Audit Readiness V3. International Journal of Computer Applications, 187(12), 29–33. https://doi.org/10.5120/ijca2025925133

[33] Shah, S. B. (2025). Evaluating the Effectiveness of Machine Learning in Forecasting Financial Market Trends: A Fintech Perspective. 2025 3rd International Conference on Integrated Circuits and Communication Systems (ICICACS), 1–6. https://doi.org/10.1109/ICICACS65178.2025.10968297

[34] Ferrara, E. (2024). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. Sci, 6(1). https://doi.org/10.3390/sci6010003

[35] Damani, D. (2022). A survey on Artificial Intelligence Bias. The National High School Journal of Science, 8(1), 1–6.

[36] Kakkar, S. (2025). Advancing Credit Risk Modeling Through Generative Artificial Intelligence: Methods, Applications, And Challenges. INTERNATIONAL JOURNAL OF FINANCIAL DATA SCIENCE, 3(2), 30–48. https://doi.org/10.34218/IJFDS_03_02_003

[37] Wilkinson, C., Yawney, J., & Gadsden, S. A. (2026). Explaining explainability: A comprehensive survey on explainable artificial intelligence and relevant industry applications. Intelligent Systems with Applications, 30, 200647. https://doi.org/10.1016/j.iswa.2026.200647

[38] Memarian, B., & Doleck, T. (2023). Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review. Computers and Education: Artificial Intelligence, 5, 100152. https://doi.org/10.1016/j.caeai.2023.100152

[39] Loganathan, M., Sharifzadeh, H., & Keivanmarz, A. (2025). Towards Improving Fairness in AI Systems: A Framework for Bias Mitigation. 2025 IEEE Region 10 Symposium (TENSYMP), 1–7. https://doi.org/10.1109/TENSYMP63728.2025.11145004

[40] Ashtekar, H., Girme, S., & Khopade, R. (2025). A Survey Mitigation of Bias and Fairness in Machine Learning Using LLMs Towards Ethical AI. International Research Journal of Modernization in Engineering Technology and Science, 07(11), 2240–2242.

[41] Trigo, A., Stein, N., & Belfo, F. P. (2024). Strategies to improve fairness in artificial intelligence: A systematic literature review. Education for Information, 40(3), 323–346. https://doi.org/10.3233/EFI-240045

[42] Chen, P., Wu, L., & Wang, L. (2023). AI Fairness in Data Management and Analytics: A Review on Challenges, Methodologies and Applications. Applied Sciences, 13(18), 10258. https://doi.org/10.3390/app131810258

[43] Huang, J., Galal, G., Etemadi, M., & Vaidyanathan, M. (2022). Evaluation and Mitigation of Racial Bias in Clinical Machine Learning Models: Scoping Review. JMIR Medical Informatics, 10(5), e36388. https://doi.org/10.2196/36388