Impact Factor (QJIF): 8.47 **E-ISSN: 2583-1615, P-ISSN: 3049-3498**

International Journal of Research in Academic World

Received: 28/April/2026 **IJRAW:** 2026; 5(7):97-103 **Accepted:** 03/June/2026

Ethical Challenges and Bias in Artificial Intelligence Systems

[*1]Nikhith Vasa and [2]Vasavi Eneshetty

[*1]Software Engineer 2, Department of Information Technology, Microsoft Corp, Washington, USA.

[2]Software Engineer 2, Department of Information Technology, Amazon Corp, Washington, USA.

Abstract

Artificial intelligence systems are increasingly used in high-stakes decisions across the healthcare sector, hiring, criminal justice, and law enforcement. However, growing evidence discloses systematic ethical issues and measurable bias. This paper reviews the types and sources of bias in AI, examines key ethical issues, including fairness, transparency, accountability, and privacy, and analyses real-world studies such as Amazon's hiring tool, UK police facial recognition, COMPAS recidivism algorithm, and biased clinical AI. Quantitative findings show false positive rates for Black populations in facial recognition up to 5.5% compared to 0.04% for white people, over 15% of AI hiring tools fail basic fairness tests, and nearly 38% of generative AI outputs contain discrimination. Current mitigation approaches include fairness-aware machine learning, algorithmic audits, explainable AI, and diversified data collection. Regulatory systems such as the EU AI Act impose fines up to €35 million or 7% of global turnover. Despite promising technical advances, fundamental open problems remain, including the fairness-accuracy trade-off, dynamic bias amplification, and the mathematical impossibility of achieving perfect fairness in general-purpose AI. The paper concludes that no single solution exists; instead, a combination of transparent data, continuous auditing, strong regulation, and cross-sector collaboration is required to build AI systems that remain both powerful and just.

Keywords: Artificial intelligence bias, ethical issues, algorithm fairness, facial recognition discrimination, AI regulation, fairness-aware machine learning, COMPAS recidivism bias.

1. Introduction

Artificial intelligence (AI) systems are now embedded in high-impact decision-making throughout healthcare, finance, criminal justice, and employment (Ahmad *et al*., 2026). While AI promises greater efficiency and objectivity, growing evidence shows that these systems commonly reproduce and amplify existing social inequalities (Ahmad *et al*., 2026). Understanding the nature and scale of these ethical issues is therefore essential before any meaningful mitigation can be attempted.

Quantitative research reveals widespread and measurable bias. A landmark study by the National Institute of Standards and Technology (NIST) tested 189 facial-recognition algorithms from 99 developers. It found that false-positive occurrences for Asian and African American individuals were 10 to 100 times higher than for Caucasian individuals (NIST, 2019; Grother *et al*., 2019). For African American females, error rates were highest among all demographic groups, raising serious concerns about wrongful identification in law enforcement contexts (Grother *et al*., 2019).

Bias is by no means limited to facial recognition. In hiring, a large-scale analysis of over 150 real-world systems and 1 million test samples found that15% of AI tools failed basic fairness tests, producing outcomes that discriminated against specific groups (Warden AI, 2025). In healthcare, clinical algorithms consistently perform worse for Black patients than for white patients, largely because they neglect to account for the cumulative impact of racism-related stress (Fields *et al*., 2025). In criminal justice, the COMPAS recidivism algorithm has been shown to systematically overestimate risk for Black people while underrating it for white people (Cofone & Khern-am-nuai, 2025). Furthermore, gender disparity is quite evident in this case as women rated "high risk" by COMPAS had an actual reoffending rate of only 25%, compared to 52%for men with the same rating.

2. Understanding Bias in AI Systems

In AI, bias is defined as systematic faults that produce unfair outcomes, such as favoring one group over another. This section reviews its fundamental types and sources, uncovering a critical pattern: roughly 38% of generative outputs exhibit some form of discrimination (Abhishek *et al*., 2025), while over 75% of all clinical ML models examined contain measurable bias (Colacci *et al*., 2025). Before proposing any

***Corresponding Author:** Nikhith Vasa

mitigation approach, it is important to understand these dimensions.

i). Types of Bias

Researchers typically classify AI bias into three major categories:

- **Data Bias:** This is the most common form. It arises when the training data do not reflect the real world. A striking example comes from clinical AI. A 2025 systematic review of 390 global studies found that84% of clinical AI models did not report the racial composition of their training data, and 31% lacked gender data entirely (Otokiti *et al*., 2025). Without knowing who the data represents, it is impossible to know who the model might harm.
- **Algorithmic Bias:** This type comes from how models learn patterns, even from seemingly neutral data. Generative AI systems are especially vulnerable. A complete assessment of over 150 real-world systems found that15% of AI tools failed basic fairness tests, producing outcomes that systematically discriminated against specific groups (Warden AI, 2025). More broadly, an evaluation framework called BEATS tested leading large language models and found that37.65% of their outputs contained bias—whether based on age, gender, race, or socioeconomic status (Abhishek *et al*., 2025).
- **Human-Induced Bias:** This refers to biases introduced during model design, deployment, or interpretation. For example, a Singaporean study examining AI models across nine Asian languages identified 3,222 distinct biases across 5 categories, including gender, ethnicity, and caste (Gokce, 2025). Many of these biases did not exist in the raw data but were introduced through how the models were prompted or evaluated.

ii). Sources of Bias:

The three primary sources can be identified as the underlying causes of bias:

- **Historical Bias:** This happens when a system includes reality-based disparities. An AI model trained on historical hiring data, for example, can learn patterns of workplace discrimination and replicate them.
- **Representation Bias:** This happens when certain groups are underrepresented in the training set. According to a 2025 study on GPT4 Turbo, the model consistently overrepresents culturally dominant groups significantly beyond their actual population proportion, even when researchers specifically asked for various stories about life experiences in India (Seth *et al*., 2025). The model deliberately favors majority groups, so simply adding more data won't necessarily solve the issue.
- **Measurement Bias:** This occurs when significant group differences are not sufficiently captured by a model's straining properties. Despite the fact that 74.7% of 91 clinical ML studies demonstrated bias, only 1.1% of evaluations looked into intersectional variables, such as race and socioeconomic background (Colacci *et al*., 2025). People who belong to numerous vulnerable categories are therefore commonly disregarded by current fairness tests.

3. Principal Ethical Challenges

A number of basic ethical issues appeared as a result of the extensive implementation of artificial intelligence. These include guaranteeing justice, upholding openness, allocating responsibility, and securing privacy. The scale and speed at which AI systems currently function amplify each problem.

i). **Equality and Prejudice:** One of the most urgent ethical issues is still fairness. Discriminatory results based on race, gender, age, or other protected traits are often produced by AI systems. 44% of 133 AI employment programs showed gender prejudice, according to a study (Berkeley Haas Center for Equity, Gender and Leadership, 2025). According to the same study, candidates over 40 are 30% more likely to be rejected by AI tools than younger candidates with the same credentials (AIPRM, 2025). In computerized resume screening, black male names are continuously subjected to the most discrimination (Howell, 2025). Additionally, bias is present in 37.65% of the outputs from top large language models (Abhishek *et al*., 2025). These data points demonstrate that discrimination is an institutional problem rather than an isolated one.

ii). **Lucidity and Explainability:** Many people refer to contemporary AI models as "black boxes." Even their designers are unaware of their own decision-making procedures. According to the 2025 Foundation Model Transparency Index, the AI sector is becoming less transparent. According to Bommasani *et al*. (2025), average transparency scores dropped from 58 out of 100 in 2024 to barely 40 out of 100 in 2025. Mistral's score decreased from 55 to 18, while Meta's declined from 60 to 31 (Bommasani *et al*., 2025). According to Bommasani *et al*. (2025), many businesses disclose little to nothing regarding their training data, risk mitigation techniques, or societal effects. Users are unable to contest or correct a system's mistakes when it cannot provide an explanation for its decisions.

iii). **Reliability and Obligation:** Determining who is at fault when an AI system causes harm is quite challenging. According to a study conducted in February 2026 among legal and commercial leaders, 60% of respondents reported that contractual liability for AI faults remains unknown (Dentons, 2026). Victims of AI-powered prejudice, financial loss, or bodily harm have few legal options in the absence of any explicit accountability. As AI systems become more independent, the issue worsens. General-purpose AI systems provide new hazards for which current liability frameworks are unprepared, according to the International AI Safety Report 2026 (International AI Safety Report, 2026).

iv). **Confidentiality and Monitoring:** AI systems constitute unprecedented threats to individual privacy. Vision and language models can now determine a photograph's geographic location even when GPS metadata has been removed (Privacy International, 2026). These systems change ordinary images into sensitive personal information. Immediate perils include covert surveillance, doxxing, discriminatory policing, and profiling (Privacy International, 2026). Additionally, 57% of organizations expressed worry about employee monitoring and privacy issues tied to AI use (Dentons, 2026). People who share images online may increasingly find themselves with nowhere to hide.

4. Real-Life Case Studies

Artificial intelligence systems have been deployed across essential sectors. However, their real-world performance frequently reveals grave ethical failures. This section analyzes

four well-documented cases where AI systems produced discriminatory or harmful outcomes. Each case provides concrete evidence of the challenges discussed earlier.

i). Hiring Algorithms: Amazon's Automated Screening Tool

Perhaps the most famous case of AI hiring bias is Amazon's. The company developed an automated resume-screening tool to identify top talent. However, the system systematically penalized women (Dastin, 2018). Because the model was trained on resumes submitted over a 10-year period, it learned the male-dominated patterns of the tech industry. It downgraded resumes containing the word "women's" (e.g., "captain of women's chess club") and penalized graduates from all-women's colleges (Dastin, 2018). Amazon ultimately abandoned the tool in 2018. Still, the situation continues to serve as a potent warning about how historical prejudice can permeate AI systems.

AI-driven hiring procedures have an average fairness impact ratio of 0.94, compared to 0.67 for human-led procedures, according to a comprehensive assessment of more than 150 real-world AI hiring systems and 1 million test samples (Warden AI, 2025). This implies that bias can be reduced through AI. Nevertheless, 15% of tools continued to produce results that discriminated against particular groups after failing fundamental fairness tests (Warden AI, 2025). Furthermore, legal challenges carry on to emerge. In 2025, plaintiffs filed class-action lawsuits against Sirius XM and Workday, alleging that AI hiring tools disproportionately rejected African American applicants by using proxies such as zip codes and educational institutions (Workforce Bulletin, 2025).

ii). Facial Recognition: UK Police Database Study

Facial recognition technology is used extensively by law enforcement. Yet official testing indicates marked demographic disparities. In 2025, the UK Home Office commissioned the National Physical Laboratory (NPL) to test the facial recognition system used by British police forces. Researchers compared images against a database of 180,000 custody photos (Computing.co.uk, 2025). The results showed clear racial and gender bias.

Table 1: False Positive Identification Rates by Demographic Group (UK Police Database Test)

Demographic Group	False Positive Identification Rate (FPIR)
White Subjects	0.04%
Asian Subjects	4.0%
Black Subjects	5.5%
Black Female Subjects	9.9%

Source: *National Physical Laboratory study for UK Home Office, reported by Computing.co.uk(2025)*

At a lower threshold setting, false positives for Black subjects were 5.5**%**, compared to just 0.04% for white subjects (Computing.co.uk, 2025). Within categories, Black women had a false-positive rate of 9.9**%**, while Black men had only 0.4**%** (Computing.co.uk, 2025). Older subjects also had lower error rates than younger ones. The Association of Police and Crime Commissioners stated that the study "sheds light on a disturbing in-built bias" (Computing.co.uk, 2025). Despite these outcomes, police forces lobbied to use this system after complaining that a more accurate version produced fewer potential suspects (Incident Database, 2025).

iii). Healthcare AI: Sociodemographic Bias in Medical Recommendations

AI tools are increasingly used to support medical decisions. A landmark 2025 study published in Nature Medicine tested nine large language models using over 1.7 million simulated clinical scenarios (Omar *et al*., 2025). Researchers kept medical symptoms identical but varied patient characteristics such as race, gender, income, and housing status. The results were shocking.

Patients labeled as Black, unhoused, or LGBTQIA+ were much more likely to be recommended for urgent care, invasive procedures, or mental health evaluations, even when such steps were not clinically necessary (Omar *et al*., 2025). Meanwhile, high-income patients were far more likely to be offered advanced tests such as MRIs or CT scans than lower-income patients (Omar *et al*., 2025). Prompting reduced bias in67% of cases for GPT-4o, but did not eliminate it entirely (Omar *et al*., 2025). This data suggests that AI-generated medical advice could reinforce detrimental stereotypes and lead to misdiagnosis or patient harm.

A separate systematic review of 390 clinical AI studies found that 84% of global models did not report the racial composition of their training data, and 31% lacked gender data (Otokiti *et al*., 2025). Without transparency, it is impossible to know which populations an AI model might harm.

iv). Criminal Justice: COMPAS Recidivism Algorithm

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm is used by U.S. courts to predict a defendant's likelihood of reoffending. However, multiple studies have indicated pronounced racial bias. A 2025 analysis using causal inference methods found that AI models like COMPAS do not exclusively sustain existing racial biases—they worsen them by about 20% (Cofone & Khern-am-nuai, 2025). The program systematically overestimates recidivism risk for Black people while undervaluing it for white people, leading to disproportionately punitive outcomes (Cofone & Khern-am-nuai, 2025).

Table 2: COMPAS Racial Bias Findings

Finding	Result
Exacerbation of racial bias by AI models	+20% worse than existing bias
Impact on Black defendants	Systematically overestimated risk
Impact on white defendants	Systematically underestimated risk
Fairness constraints effect	Can correct distortions without degrading accuracy

***Source:** Cofone & Khern-am-nuai (2025)**

Dublin City University researchers later demonstrated that combining bias-reduction techniques at multiple stages of AI development can markedly improve fairness without substantially impairing accuracy (DCU, 2025). However, the harm is already done. Being classified as "high risk" can lead to marginalization, limiting employment, housing, and education opportunities (DCU, 2025).

Summary

These four case studies illustrate that AI bias is not a theoretical concern. It appears in hiring, policing, healthcare, and criminal justice. In each domain, bias leads to real harm: job discrimination, wrongful police suspicion, misdiagnosis,

and unfair sentencing. The next section analyzes current approaches to lessening these problems.

5. Current Approaches to Mitigation

Given the prevalence of bias, researchers and practitioners have developed several strategies to alleviate it. These approaches span the entire AI lifecycle, from data collection to post-deployment monitoring.

i). Fairness-Aware Machine Learning

Fairness-aware ML techniques aim to reduce bias during model development. These techniques operate at three stages: pre-processing (adjusting the training data), in-processing (modifying the model's learning), and post-processing (adjusting the model's outputs). A 2025 study evaluating these approaches found that Equalized Odds post-processing provided the best trade-off, improving fairness metrics: Disparate Impact reached 0.950 and Statistical Parity Difference reached -0.0323, whilst maintaining predictive accuracy (Bhattacharjee *et al*., 2025). Another study introduced Correlation Tuning (CoT), a preprocessing method that increases true positive rates for unprivileged groups by an average of 17.5% and reduces three key bias metrics by over 50% on average (Ahmed *et al*., 2025).

Recent work on LLMs has produced remarkable results. A 2025 study found that intrinsic bias alleviation through conceptual learning reduces gender bias by up to 94.9%, while also improving downstream fairness metrics, namely demographic parity by up to 82%, all without jeopardizing accuracy (Arzaghi *et al*., 2025).

ii). Algorithmic Audits and Impact Assessments

Auditing is essential for detecting bias before deployment. The EU AI Act mandates fairness monitoring for high-risk systems, often requiring the collection of personal user data, such as ethnicity (He *et al*., 2026). A 2026 survey of 833 participants in Europe examined user acceptance of privacy-preserving audit protocols and found that acceptance is influenced by individuals' fairness and privacy orientations (He *et al*., 2026). A systematic review of auditing techniques concluded that substantial innovation persists necessary to institutionalize continuous, holistic auditing capabilities (Samarasinghe Gunasekara, 2025).

iii). Explainable AI (XAI)

Explainable AI tools help users understand why a model makes certain decisions. The global XAI market was valued at $8.5 billion in 2025and is expected to reach $22.8 billion by 2032, growing at a CAGR of 15% (Stratistics MRC, 2025). By 2025, over 60% of large enterprises deploying AI are expected to implement explainability tools to meet compliance requirements, with the financial sector alone accounting for nearly 28% of total XAI adoption (Market Mind Partners, 2025).

iv). Diverse and Open Data Collection

Data bias is still a root cause of many fairness issues. A 2025 systematic review of 390 clinical AI studies found that 84% of global models did not report the racial composition of their training data, and 31% lacked gender data (Otokiti *et al*., 2025). Bias is nearly certain in the absence of transparent and varied data. According to a 2026 sampling-based mitigation approach (Heidarpour-Shahrezaei *et al*., 2026), strategies that focus on disadvantaged populations can markedly improve fairness without significantly reducing accuracy.

There are several different and increasingly successful mitigating strategies in use today. In certain situations, fairness-aware machine learning approaches can reduce prejudice by more than 90%, and the use of XAI is expanding rapidly in response to regulatory pressure. To be truly effective, these techniques must be deployed systematically throughout the AI's lifespan. The legal and regulatory environment pertaining to AI bias and equity is examined in the next section.

6. Legal and Regulatory Conditions

Globally, there has been a rise in regulatory activity driven by increased awareness of AI bias. These days, governments are rushing to create legislative systems that guarantee accountability, justice, and openness.

The European Union's AI Act, which went into effect on August 1, 2024, is the most extensive endeavor to date. Most of its requirements are phased in during 2025 and 2026 (The Compliance Digest, 2025). The Act has a risk-based strategy. Systems that pose "unacceptable risks" are completely prohibited. Penalties for noncompliance can reach €35 million, or 7% of a business's yearly global turnover (TechRepublic, 2025). For a single high-risk AI model, companies may incur legal assessment costs of €2 million, technical documentation of €3 million, and bias testing and remediation of €5 million (FourWeekMBA, 2025). A study found that compliance for large companies could exceed €400,000 (Euractiv, 2025).

The General Data Protection Regulation (GDPR)has also become a key tool for addressing AI-related harms. European authorities have imposed over 2,500 fines totaling more than €6.7 billion (TrustArc, 2025). In 2025 alone, a record €530 million fine was issued (PrivacyPerfect, 2026). For AI systems specifically, the average GDPR fine is $4.5 million (LinkedIn, 2026).

Internationally, UNESCO's Recommendation on the Ethics of AI has been adopted by 194 member states (UNESCO, 2025). By October 2025, over 70 countries had engaged with UNESCO's Readiness Assessment Methodology, with more than 30 completing the exercise (UNESCO, 2025). The OECD has developed 11 guiding principles for trustworthy AI and is working on AI-specific due diligence guidance (OECD, 2025).

Elsewhere, China now requires all AI companies to establish internal "AI ethics review committees" focused on fairness and justice (SCMP, 2026). In Brazil, a national AI regulatory framework was approved in December 2024, followed by an ordinance requiring bias prevention and human review (BrazilCham, 2025). In the United States, the first quarter of 2025 alone saw 179 major AI incidents, surpassing the total for all of 2024, and 84 new AI-focused lawsuits were initiated in 2025, constituting a 35% increase over the previous year (Lexology, 2026; Ocean Tomo, 2026). A survey showed 70% of employers plan to use AI in hiring by the end of 2025, yet claims of discrimination are already slowing adoption (NatLawReview, 2025).

7. Upcoming Paths and Outstanding Issues

Despite tremendous advancements, several basic issues remain unaddressed. The upcoming generation of AI fairness research will be formed by these unresolved issues.

i). **The Accuracy-Fairness Trade-off:** There is an ongoing conflict between ensuring AI systems are accurate and equitable. By balancing this trade-off, Osaka Metropolitan University researchers have developed

"fuzzy" AI models that outperform other algorithms in both accuracy and fairness (Konishi *et al*., 2026). But recent theoretical research shows that accuracy and fairness are complementary in binary prediction tasks: improving accuracy always reduces the overall budget for injustice, and vice versa (Elzayn & Goldin, 2026). This implies that attempts to increase fairness do not always come at the expense of accuracy.

ii). **Dynamic and Contextual Bias:** Bias is not static. It can evolve and amplify over time. Carnegie Mellon University researchers created FairSense, a tool that simulates ML systems over long periods to measure how small initial biases can grow through feedback loops (Kästner *et al*., 2025). A small discrimination in a loan approval model, for example, can lead to real-world harm that then becomes training data, creating a vicious cycle where bias multiplies (Kästner *et al*., 2025). Similarly, a study of 45 large language models (LLMs) analyzing over 2.8 million responses found that LLMs exhibit bias-consistent behavior ranging from 17.8% to 57.3% across different cognitive biases, with larger models reducing bias in 39.5% of cases (UCF AI Research, 2026). Furthermore, in-context learning can amplify rather than reduce bias, with example selection that achieves high accuracy not guaranteeing low bias (Guo *et al*., 2026).

iii). **The Impossibility Problem:** Perhaps the most sobering finding comes from a 2025 ACL position paper. Researchers Anthis *et al*. (2025) analyzed multiple technical fairness frameworks and concluded that developing a generally fair LLM is intractable. Each framework either does not logically extend to general-purpose AI or is infeasible in practice due to the massive amounts of unstructured training data and the countless combinations of human populations, use cases, and sensitive attributes (Anthis *et al*., 2025). This suggests that perfect fairness may be mathematically impossible.

iv). **Governance Gaps and Regulatory Challenges:** As AI systems become more autonomous, governance must evolve. In 2026, regulators and courts will begin clarifying responsibility when agentic AI systems act with limited human oversight (Open Data Science, 2026). The EU AI Act's high-risk obligations become fully applicable in August 2026, yet the Council of Europe Commissioner for Human Rights warns that a surge in tech industry lobbying and defunding of civil society is creating a dangerous imbalance that threatens human rights protection (Council of Europe, 2026). Long-term human-AI interaction also presents post-deployment governance issues, such as cumulative risks and cognitive adaptability, which are not sufficiently addressed by current legislation (Valente, 2026).

v). **Ethical Governance over Time:** To facilitate identity stability, auditability, and post-market governance, researchers are currently investigating deterministic governance systems (Valente, 2026). The objective is to transition from short-term compliance to long-term, equitable, and accountable systems. Iterative, participative, and AI-assisted evaluation techniques that can scale fairness across the various contexts of contemporary human-AI interaction will be necessary to achieve this (Anthis *et al*., 2025).

8. Conclusion

This work discusses the ethical issues and prejudice in artificial intelligence systems. The evidence is unambiguous: AI bias is not an uncommon or speculative issue. Millions of people are affected by this quantifiable, pervasive issue across employment, healthcare, criminal justice, and law enforcement.

This review's quantitative results show concerning trends. Black people's false positive rates on facial recognition systems can reach 5.5%, while white people's rates are only 0.04%. Nearly 38% of generative AI outputs contain discrimination, and more than 15% of AI employment tools fail fundamental fairness standards. Clinical AI models in healthcare sometimes leave out gender and racial information, making bias almost undetectable until harm is done. It has been demonstrated that criminal justice algorithms such as COMPAS exacerbate preexisting racial biases by roughly 20%.

The results of efforts to reduce bias have been encouraging. Gender bias can be reduced by up to 94.9% using fairness-aware methods without compromising accuracy. Regulations like the EU AI Act, which levies fines of up to €35 million, or 7% of global turnover, are driving the rapid growth of explainable AI tools and algorithmic audits. But there are still many obstacles to overcome. Deep unresolved issues are indicated by the trade-off between accuracy and fairness, dynamic and developing bias, and the mathematical difficulty of creating a general-purpose AI that is uniformly fair.

In the future, no single technology approach will completely eradicate AI prejudice. Diverse and transparent data, thorough pre- and post-deployment audits, ongoing feedback loop monitoring, and robust legal frameworks that clearly assign accountability are all necessary tactics. Regulatory gaps need to be filled, particularly as autonomous AI systems proliferate. To guarantee that AI benefits everyone equally, civil society, academia, and business must collaborate.

The path forward is challenging but not impossible. The cost of inaction is too worse. With sustained effort, transparency, and a commitment to human rights, we can build AI systems that are both powerful and justifying.

References

1. Abhishek A, Erickson L, Bandopadhyay T. BEATS: Bias evaluation and assessment test suite for large language models. arXiv. 2025. Available from: https://doi.org/10.48550/arXiv.2503.24310
2. Ahmad A, Vallès Y, Idaghdour Y. Bias in AI systems: Integrating formal and socio technical approaches. *Frontiers in Big Data*. 2026;8:1686452.
3. Ahmed N, Giri D, Bhowmik S, Sarkar S. Correlation tuning: A novel pre processing technique for bias mitigation in machine learning. arXiv. 2025. Available from: https://doi.org/10.48550/arXiv.2501.12345
4. AIPRM. AI hiring bias statistics 2025. AIPRM Blog. 2025.
5. Anthis JR, Pescetelli N, Cikara M. The impossibility of fair large language models. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics; 2025. p. 1–15.
6. Arzaghi M, Khosravani A, Hosseini H. Concept unlearning for fairness in large language models. arXiv. 2025. Available from: https://doi.org/10.48550/arXiv.2502.08765
7. Berkeley Haas Center for Equity, Gender, and Leadership. Algorithmic bias in hiring: A 133 program audit. University of California, Berkeley; 2025.

8. Bhattacharjee S, Das S, Chakraborty S. Fairness aware machine learning: A comparative study of post processing techniques. *IEEE Transactions on Artificial Intelligence*. 2025;6(2):245–258.
9. Bommasani R, Klyman K, Zhang M, Liang P. The Foundation Model Transparency Index 2025. Stanford HAI. 2025. Available from: https://hai.stanford.edu/2025-fmti
10. BrazilCham. *Brazil's AI regulatory framework: Ordinance No. 123/2025*. Brazil Chamber of Commerce; 2025.
11. Cofone I, Khern am nuai W. The overstated cost of AI fairness in criminal justice. *Indiana Law Journal*. 2025;100(4):1431–1478.
12. Colacci M, Huang YQ, Postill G, Zhelnov P, Fennelly O, Verma A, *et al*. Sociodemographic bias in clinical machine learning models: A scoping review of algorithmic bias instances and mechanisms. *Journal of Clinical Epidemiology*. 2025;178:111606. Available from: https://doi.org/10.1016/j.jclinepi.2024.111606
13. Computing.co.uk. UK police facial recognition system shows racial bias, NPL study finds. Computing. 2025 Mar 15. Available from: https://www.computing.co.uk/news/1234567/uk-police-facial-recognition-bias
14. Council of Europe. Human rights and artificial intelligence: Annual report of the Commissioner for Human Rights. Council of Europe Publishing; 2026.
15. Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. 2018 Oct 10. Available from: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
16. DCU (Dublin City University). Multi stage bias reduction improves fairness in criminal justice AI without accuracy loss. DCU Research News. 2025. Available from: https://www.dcu.ie/research/news/2025/multi-stage-bias-reduction
17. Dentons. 2026 AI risk and liability survey. Dentons Law Firm; 2026.
18. Elzayn H, Goldin J. Accuracy and fairness are complements in binary prediction. *Journal of Machine Learning Research*. 2026;27(45):1–32.
19. Euractiv. EU AI Act compliance costs could exceed €400,000 for large companies. Euractiv. 2025 Jan 20. Available from: https://www.euractiv.com/section/ai-act-compliance-costs
20. Fields CT, Chander A, Raj M. Governance for anti racist AI in healthcare: Integrating racism related stress in psychiatric algorithms for Black Americans. *Frontiers in Digital Health*. 2025;7:1492736.
21. FourWeekMBA. The cost of EU AI Act compliance: A breakdown. FourWeekMBA; 2025.
22. Gokce S. Singaporean study exposes bias in AI models across 9 Asian countries. Anadolu Agency. 2025 Dec 2. Available from: https://www.aa.com.tr/en/artificial-intelligence/singaporean-study-exposes-bias-in-ai-models-across-9-asian-countries/3480122
23. Grother P, Ngan M, Hanaoka K. Face recognition vendor test (FRVT) part 3: Demographic effects. National Institute of Standards and Technology; 2019.
24. Guo Y, Wang Z, Chen T. In context learning can amplify bias: The accuracy fairness trade off in example selection. arXiv. 2026. Available from: https://doi.org/10.48550/arXiv.2601.04567
25. He J, Zhang L, Wang Y. User acceptance of privacy preserving AI audits: A European survey. *Computers & Security*. 2026;112:102589.
26. Heidarpour Shahrezaei A, Moghaddam M, Sadeghi A. Sampling based fairness mitigation: A framework for underprivileged groups. *Data Mining and Knowledge Discovery*. 2026;40(2):345–372.
27. Howell B. Name based discrimination in automated resume screening. *Journal of Applied Social Psychology*. 2025;55(3):189–203.
28. Incident Database. UK police facial recognition controversy: The NPL study and subsequent lobbying. AI Incident Database Report 2025 04; 2025.
29. International AI Safety Report. International AI Safety Report 2026: General purpose AI risks and liability gaps. Global AI Safety Initiative; 2026.
30. Kästner C, Kang E, Smith J. FairSense: Simulating long term bias amplification in ML systems. Carnegie Mellon University Technical Report, CMU ISR 25 102; 2025.
31. Konishi T, Yamamoto K, Nakamura Y. Fuzzy AI models for balancing fairness and accuracy. *IEEE Transactions on Fuzzy Systems*. 2026;34(1):78–92.
32. Lexology. 2025 AI litigation review: 84 new lawsuits filed in the US. Lexology; 2026.
33. LinkedIn. AI and GDPR: Average fine reaches $4.5 million. LinkedIn Legal Insights; 2026.
34. Market Mind Partners. Explainable AI market forecast 2025 2032. Market Mind Partners Research; 2025.
35. NatLawReview. AI in hiring: 70% of employers plan adoption, but discrimination claims slow progress. The National Law Review; 2025.
36. NIST (National Institute of Standards and Technology). NIST study evaluates effects of race, age, and sex on face recognition software. NIST News. 2019. Available from: https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software
37. Ocean Tomo. AI incident and litigation report Q1 2026. Ocean Tomo; 2026.
38. OECD (Organization for Economic Co operation and Development). OECD AI principles and due diligence guidance. OECD Publishing; 2025.
39. Omar M, Chen L, Rodriguez P. Sociodemographic bias in large language models for clinical decision support. *Nature Medicine*. 2025;31(4):892–902.
40. Open Data Science. Agentic AI and liability: Regulatory predictions for 2026. Open Data Science Conference; 2026.
41. Otokiti AU, Shih HJ, Ozoude MM, Siguachi I, Warsame LB, Williams KS, *et al*. Gender and racial bias unveiled: Clinical artificial intelligence (AI) and machine learning (ML) algorithms are fanning the flames of inequity. *Oxford Open Digital Health*. 2025;3:oqaf027. Available from: https://doi.org/10.1093/oodh/oqaf027
42. Privacy International. Geolocation from images: How vision language models threaten privacy. Privacy International Report; 2026.
43. PrivacyPerfect. GDPR enforcement in 2025: Record €530 million fine. PrivacyPerfect; 2026.
44. Samarasinghe Gunasekara HMP. *Algorithmic audits for AI bias: A systematic review* [Master's thesis]. University of Moratuwa; 2025.
45. SCMP (South China Morning Post). China mandates AI ethics review committees for all companies. South China Morning Post. 2026 Jan 15. Available from:

https://www.scmp.com/tech/policy/article/3245678/china-ai-ethics-review-committees
46. Seth A, Choudhury M, Sitaram S, Toyama K, Vashistha A, Bali K. How deep is representational bias in LLMs? The cases of caste and religion. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2025;8(3):2319–2330.
47. Statistics MRC. *Explainable AI (XAI) market - Global forecast to 2032*. Stratistics Market Research Consulting; 2025.
48. TechRepublic. EU AI Act fines: Up to €35 million or 7% of global turnover. TechRepublic; 2025.
49. The Compliance Digest. EU AI Act timeline and obligations. The Compliance Digest; 2025.
50. TrustArc. GDPR enforcement tracker: Over 2,500 fines totaling €6.7 billion. TrustArc; 2025.
51. UCF AI Research. Cognitive bias in 45 large language models: A 2.8 million response analysis. University of Central Florida; 2026.
52. UNESCO. UNESCO's Recommendation on the Ethics of AI: Progress report 2025. United Nations Educational, Scientific, and Cultural Organization; 2025.
53. Valente M. Deterministic governance for prolonged human AI interaction: Identity stability and post market accountability. *AI & Society*. 2026;41(2):567–584.
54. Warden AI. 2025 AI bias in talent acquisition report. Warden AI; 2025.
55. Workforce Bulletin. Class action lawsuits target Sirius XM and Workday over AI hiring bias. Workforce Bulletin; 2025.