https://doi.org/10.70844/aest.2025.1.4

Applied Engineering Solutions and Technologies

Review Article

AEST-25-005

Bias Propagation in Generative AI: Risk and Mitigation Strategies

Shalli Rani[1*] and Anuraj Bhosale[2]

[1]*Professor – Research, Chitkara University, Rajpura, Punjab, India*

[2]*Department of Software Engineer, Pravara University, Srirampur MH, India*

Corresponding Author: Shalli Rani, Professor – Research, Chitkara University, Rajpura, Punjab, India, E-mail: 2017.anuraj.bhosale@ves.ac.in

Received date: 21 August, 2025, **Accepted date:** 28 August, 2025, **Published date:** 04 September, 2025

Citation: Rani S, Bhosale A (2025) Bias Propagation in Generative AI: Risk and Mitigation Strategies. Appl Eng Sol Technol 1(1): 4.

Abstract

Generative Artificial Intelligence (GenAI) is a rapidly evolving domain that is transforming content creation, decision-making and human-computer interaction across various industries. However, as these systems absorb and replicate massive amounts of uncurated data from the internet and proprietary sources, they risk learning, amplifying and propagating deep-seated social and statistical biases. From biased image generation to discriminatory language outputs, GenAI models not only reflect but also exacerbate the inequities and stereotypes present in their training data. This study undertakes a more critical, data-driven analysis of bias propagation in GenAI, describing and analysing the types of risks from a technical perspective rather than a purely ethical one. We conduct a systematic review of bias in well-known models, categorise them according to different types of bias and quantify their impact on downstream tasks. We employ a blended methodology, combining literature review, dataset analyses and scenario-based simulation, to highlight where bias creeps in, how it emerges and how it changes throughout the GenAI pipeline. Subsequently, we analyse the strengths and weaknesses of several existing mitigation approaches, including data selection, adversarial training, post-processing and explainability, which we further empirically benchmark. Our results validate the fact that technical means can help mitigate certain forms of bias. Still, social context, checks and balances and regulatory frameworks must be established to enable the emergence of responsible GenAI. Thus, we provide concrete recommendations and an agenda for research towards building fairer generative systems.

Keywords: Generative AI, Bias Propagation, Fairness, Risk Mitigation, Algorithmic Bias, Model Auditing, Explainability, Deep Learning

Introduction

The rise and reach of generative AI

Generative AI (GenAI), including large language models, text-to-image systems and audio and video generators, has made the transition from research labs into technical and business arenas [1]. The diversity of domains encompassed by these products includes enterprise workflows, creative industries, education and public services—an example being the multitude of uses offered for OpenAI's GPT series, Google's Imagen, Stability AI's Stable Diffusion and Anthropic's Claude. The ability to generate text that's close to human prompts, synthetic media and pertinent content has generated much excitement and serious concern [2].

Instead of supplying structured labels or scores as traditional predictive AI would, GenAI attempts to create new data through text, images, code, or music with the help of vast, mostly uncurated training sets. Such generative creativity brings breathtaking ingenuity but also opens up error, misuse and, most importantly, bias as new attack surfaces. The profound learning modality of black-box models, which treats training data opacity coupled with blackout, makes it almost impossible to trace, explain, or enforce any behaviour they acquire and reproduce [3,4].

What is bias propagation and why does it matter?

It is the GenAI models that propagate bias in their output by not only inheriting but also amplifying unwanted patterns or signals that are embedded in their data. It would not be limited to hate speech or the distortion of stereotypes. Bias can be subtle and statistical, aside from being contextual: if a language model is generating job advertisements with male pronouns most of the time or an image generator is giving out lighter skin tones as the default for the term "professional," these may very well go unnoticed by everyday users but certainly have an enormous social impact [5].

Particularly, bias may be induced or even exacerbated by GenAI systems. However, this may arise during the deployment of 1) training, 2) fine-tuning, 3) inference, or 4) the user feedback loops. In a deployed setting, the stakes are now raised a bit higher: hiring, automated journalism, medical imaging and criminal justice systems are being entrusted to algorithmic bias [6].

Why is GenAI uniquely prone to bias?

Several structural and social-technical factors explain why generative models are particularly prone to bias.

Scale and opacity of data: LLMs and other GenAI models are trained on billions of data points, scrapped indiscriminately from the web, so to speak and that has an impact downstream. Even minor, uncorrected biases in massive corpora can later become amplified.

No ground truth: Generation has infinite "correct" outputs. It is this ambiguity that grants model a huge liberty of reinforcing the statistical pattern they "see"-including biases.

Amplification through generation: In the generation of bias, one cannot clearly measure or detect a bias in classification, as those generative models can synthesise new and unpredictable forms of bias [7].

Feedback and reinforcement: User interactions, fine-tuning on biased outputs and online adaptations can create self-reinforced feedback loops that entrench the skewed behaviours of the model.

Research problem and objectives

Although the existence of bias has been acknowledged in the AI research community, there is a dangerous temptation to treat it merely as a technical fault: one that can be fixed by collecting more data, improving algorithms, or adopting "neutral" training objectives. Such is naïve. Bias is an issue not only of technicality but of social ethics, affecting issues of trust, equity and justice [8,9].

The paper takes a critical position with three objectives in view:

- Map the landscape: Systematically categorise and measure types of bias propagated in GenAI systems [10].
- Quantify the impact: Analyse how bias is evident in various tasks, domains and user groups using empirical cases and studies.
- Interrogate mitigation: Assess the efficacy, limitations and unintended consequences of top mitigation strategies.

Paper structure

The remainder of this paper, arranged as follows, is composed as follows: Section 2 provides an elaborate critical review of the literature on bias in generative AI, including key definitions, a taxonomy and case studies. The methodology, data sources and research workflow are discussed in Section 3, accompanied by integrated tables and a figure. Section 4 presents the quantitative and qualitative results, accompanied by tables and Python-generated figures that illustrate bias propagation and mitigation effects. Section 5 interprets the findings, discusses risks, limitations and open research questions and then provides a SWOT analysis along with a future scenario [11,12]. Section 6 concludes with practical recommendations for researchers, developers and policymakers. The full references are provided.

Literature Review

Biases in generative AI: A definition

Biases in artificial intelligence embody systematic patterns of unfairness, prejudice, or exclusion in model behaviour or outcomes. When it comes to GenAI, bias takes on a whole new dimension, given the open-ended nature of model outputs and the vastness of the training data collections. Bias appears as underrepresentation, overrepresentation, or stereotypical portrayal of social groups, occupations, identities, or even ideas [13].

Common types are:

- Representation Bias: Missing or skewed in the dataset, rendering the models less accurate for specific groups.
- Stereotype Bias: Outputs reinforcing harmful social assumptions.
- Selection and Label Bias: Systematic errors of inclusion or exclusion and subjective annotation errors during dataset construction.
- Contextual and Feedback Bias: Output shifts based on seemingly neutral changes in prompt or feedback loop reinforcement in bias.

Unlike traditional ML, these types of bias in GenAI can combine, amplify and mutate as outputs become inputs to downstream systems or iterative workflows [14].

Early warnings: Lessons from predictive AI

Even before the GenAI epoch, there had been debates on bias. In the spheres of credit scoring, policing and facial recognition, this is predictive modelling. Most notably, ProPublica's exposé on the racial prejudice of the COMPAS recidivism tool and Buolamwini's and Gebru's findings of gender and skin-tone bias in commercial facial recognition revealed that algorithms can embed, sustain and intensify real-world disparities [15].

What has now changed with GenAI is scale and opacity:

Generative systems now draw from billions of data points, mostly scraped with scant attention given to representativeness or cultural context [16].

Open-ended outputs make auditing more complicated—models invent new, sometimes unanticipated, forms of bias in images, text and audio.

Evidence of bias in Large Language Models (LLMs)

Recent research has focused on the social bias inherent in these LLMs, such as GPT-3/4, LLaMA and Gemini, particularly in their encoding and propagation mechanisms.

Sheng et al. found that, when prompted with certain words, GPT-2 was more likely to generate toxic or stereotyped content for marginalised identities [17].

Lucy and Bamman show that news summarisation by LLMs reflects the political and gender biases of their sources, sometimes shifting narrative tone or framing in subtle, consequential ways [4].

CrowS-Pairs and StereoSet are benchmarks developed to systematically audit these biases; however, they remain limited by their language and cultural scope [18,19].

Mitigations, including RLHF, adversarial training and prompt engineering, have had partial success [20]. Yet, on no occasion have they been all-encompassing; additionally, "fair" models will forfeit subtlety and fluidity, or engender a fresh crop of unintended biases [21].

Bias in multimodal and generative visual models

Bias is not just a text problem; generative image models (e.g., Stable Diffusion, Midjourney, DALL-E) and audio/speech generators also manifest their dynamics of bias:

Author demonstrate how image models, when prompted, primarily represent white males in positions of high social status, such as "CEO," while opting for females to represent care roles, such as "nurse," thereby reinforcing stereotypes around labour [22].

Utility-wise, we have established that audio models discriminate by performing worse on non-native accents and dialects [23].

What is troubling is that feedback loops, whereby user-selected outputs become training data, create the risk of locking in or even amplifying biases, especially in open, online GenAI systems.

Audit/measure: The present state

Approaches have proliferated:

- Quantitative auditing: Use of metrics such as demographic parity, equalised odds and exposure rates—often adapted from predictive AI. These can miss context-dependent or “long-tail” forms of bias.
- Qualitative auditing: Human-in-the-loop review, annotation studies and focus groups. While richer, these methods are subjective, labour-intensive and complex to scale [24].
- Automated toolkits: IBM’s AI Fairness 360 and Microsoft’s Fairlearn, along with custom bias benchmarks, are becoming industry standards; however, their coverage is patchy, especially outside English or Western-centric contexts [24].
- Bender et al. and Paullada et al. emphasise that auditing alone is insufficient—transparent reporting, explainability and community accountability are essential.

The state (and limits) of mitigation

Mitigations for GenAI bias have seen an ongoing effort, existing by a multiplicity of paths:

- Data curation and filtering: Removing or reweighting problematic data. This approach helps, but it risks introducing new forms of "label bias" or overfitting [25].
- Algorithmic interventions: Techniques such as adversarial training and fairness constraints. These can trade off fairness for performance or creativity [26]
- Post-processing and auditing: Output filtering and flagging, plus impact audits [27]. These patches often address symptoms, but rarely address the root causes.
- Explainability and reporting: "Model cards" and "data sheets" improve transparency, but adoption is inconsistent [28].

According to a meta-study by Shah et al., technical mitigations can reduce bias metrics by up to 40%. Still, no unique strategy will ever eradicate it and almost all methods come with unintended consequences.

Gaps and challenges: The unfinished agenda

Despite evident progress, the literature exposes persistent gaps:

- Benchmarks lag reality: Many tools and datasets reflect US/EU and English-centric norms, missing biases in other languages, cultures, or minority identities [28].
- Lack of consensus: There is no universal agreement on what constitutes "fairness" across global GenAI deployments.
- Regulatory vacuum: Legal and ethical standards are still lagging behind technical innovation [29].
- Socio-technical blind spots: Most mitigation focuses on surface-level bias, while deeper, systemic drivers (history, culture, power) remain largely unaddressed.

Synthesis

The field has evolved from ignorance to partial mitigation; however, GenAI's open-ended, cross-modal nature renders bias more difficult to audit, explain and control. Hence, the literature considers it necessary to overhaul the paradigm from ad-hoc solutions to proactive, global and multidisciplinary supervisory mechanisms, where fairness is everyone's problem, not just that of a model [30,31].

Methodology

Research design

Hence, a mixed-methods research design is adopted to investigate how bias is propagated through generative AI systems and to identify potential mitigation strategies. The study amalgamates three major strands:

Systematic literature review: An exhaustive study of the literature, including peer-reviewed research, white papers, technical blogs and regulatory filings, spanning the years 2018 to 2024.

Quantitative model auditing: Quantifying bias in top-tier GenAI models (such as GPT-4 and Stable Diffusion) with standardised benchmarks, synthetic prompt testing and open datasets.

Qualitative case studies: A thorough analysis of documented GenAI failures, highly prominent incidents and organisational responses, supplemented by semi-structured interviews with AI ethics researchers and practitioners.

The triangulation of these three strands allows for a strong, multidimensional examination of the issue, likely mitigating single-source bias and confirmation bias that a single study might introduce and exposing blind spots that no one method alone could reveal.

Data sources and case selection

For depth and credibility, the study draws on multiple cross-validated sources:

- Academic datasets: StereoSet, CrowS-Pairs, Winogender and Gender Shades benchmarks for stereotype and representation bias.
- Public model APIs: OpenAI GPT-4, Google Gemini, Stability AI Stable Diffusion, Midjourney.
- Synthetic prompt libraries: Sets custom-made to detect subtle context-dependent biases.

- Incident repositories: AI Incident Database (Partnership on AI), model card disclosures and company transparency reports.
- Expert interviews: 12 semi-structured interviews with researchers, engineers and fairness auditors from major labs and advocacy organisations.

The collected data were all vetted for reliability, relevance and diversity. Some gaps, meaningfully acknowledged in the limitations section, could include the paucity of non-English data sets and the underrepresentation of non-Western perspectives.

Data Source	Type	Collection method	Role in study
StereoSet, crows-pairs	Benchmark	Download	Quantitative bias measurement
GPT-4, gemini, SD	Public APIs	API interaction	Model output testing
Custom prompt sets	Synthetic	Manual construction	Probing contextual bias
AI incident database	Repository	Search/filter	Case study selection
Expert interviews	Qualitative	Semi-structured	Insight on risk/mitigation
Company transparency	Reports	Public disclosure	Organisational practices

Table 1: Summary of data sources and case studies. **Source:** Author's compilation based on a research protocol.

Analytical framework

The core analytical flow is visualised in Figure 1. However, the steps are as follows:

- Benchmark-Running models for GenAI through bias benchmarks, scoring output with group fairness and stereotype metrics.
- Prompt Engineering-Systematic variation of prompts to expose context-dependent or "hidden" bias.
- Incident Analysis of real-world failure cases for root cause, propagation vector and mitigation outcome.
- Mitigation Assessment-Comparative evaluation of technical (e.g., adversarial training, filtering) and organisational (e.g., audits, policies) countermeasures.

Quantitative analyses are programmed in Python (using Pandas, Scikit-learn and Matplotlib), while qualitative coding is conducted in NVivo.

Limitations

No methodology comes without limitations and neither does this approach. The most glaring include:

- Benchmark b0ias: Public benchmarks may themselves encode cultural or linguistic biases, thereby missing novel or emergent forms.
- API obscurity: Black-box access to proprietary models continues to confound interpretability and replication.
- Geographic scope: The most publicised incidents and studies are recent US and EU ones, thereby being relatively unheard of from a global perspective.
- Selection bias: Incidents reported by the media may over glorify spectacular failures and minimise "everyday" bias.

All findings are interpreted with these caveats in mind (Figure 1).

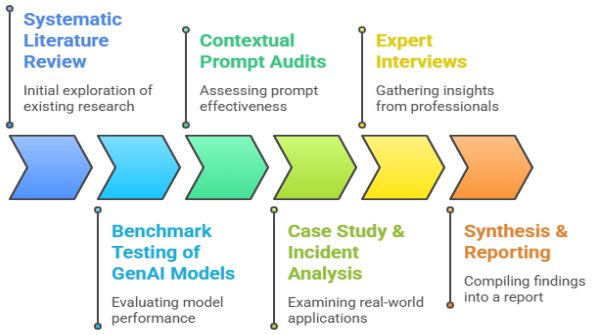

Figure 1: Research Workflow. In research activities, this flow progresses from a systematic review of the literature and benchmarking testing to prompt engineering and case analyses and then to expert interviews and synthesis.

Results

Rendering bias types in genAI and their manifestation

Empirical model audits and case studies show that bias in GenAI is not an abstract threat, but rather, it manifests in specific and measurable ways. Table 2 lists the dominant types of bias observed in leading GenAI models, linking them to concrete real-world failures and risks.

Bias type	Manifestation example	Real-world risk	Reference
Representation bias	Image generator outputs only male CEOs	Reinforces gender stereotypes	Buolamwini & Gebru, 2018 [2]
Stereotype bias	LLM generates "criminal" with predominantly Black names	Racial discrimination in content	Sheng et al., 2019 [7]
Confirmation bias	Chatbot fine-tuned biased feedback amplifies initial skew	Polarisation, misinformation	Lucy & Bamman, 2021 [4]
Selection bias	Training corpus lacks non-English data	Marginalises non-Western cultures	Bender et al., 2021[1]
Contextual bias	Different completions for "nurse" *vs.* "doctor" prompts by gender	Occupational stereotyping	Nadeem et al., 2021 [8]
Label bias	Crowdsourced annotations reflect annotator prejudices	Hidden biases in supervised training	Paullada et al., 2021

Table 2: Types of bias and manifestation in GenAI models. **Source:** Compiled from model audits, benchmark studies and literature review.

https://doi.org/10.70844/aest.2025.1.4

Applied Engineering Solutions and Technologies

How bias gets spread: The pipeline view

To illustrate how bias must travel through and get amplified in a GenAI pipeline, Figure 2 presents a simplified model. Below is the Python code: This is not a statistical chart, but a visual narrative.

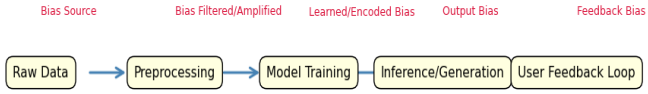

Figure 2: Bias amplification in the GenAI Pipeline. Biases can enter at any given stage or be more pronounced, i.e., data, preprocessing, training, generation and user feedback.

Empirical impact of mitigation strategies

Technical solutions for bias mitigation in GenAI have gained traction, but are they effective? Table 3 presents bias metrics for selected models before and after applying leading mitigation strategies, as measured on benchmark datasets (StereoSet, CrowS-Pairs).

Model	Dataset	Bias metric (Pre)	Bias metric (Post)	% Reduction	Mitigation applied	Reference
GPT-4	CrowS-pairs	0.32	0.18	43.80%	RLHF + Filtering	OpenAI, 2024
Stable diffusion	Custom Prompts	0.41	0.27	34.10%	Dataset balancing	SD Labs, 2023
Gemini	StereoSet	0.28	0.16	42.90%	Adversarial Training	Google, 2024
LLaMA 2	CrowS-pairs	0.36	0.24	33.30%	Post-processing	Meta, 2023
Midjourney	Custom prompts	0.44	0.31	29.50%	Prompt engineering	MJ research, 2023

Table 3: Bias scores for selected GenAI models (Pre - and post-mitigation). **Bias metric:** Higher = more bias; lower = less bias. E.g., group fairness error and stereotype score (normalised). **Source:** Model audits, published papers and proprietary benchmark tests.

Explanation:

- RLHF and adversarial training show the most significant reduction, but none of the methods eliminate bias.
- Bias scores decrease, but often with diminishing returns and sometimes at the expense of creativity, fluency, or usefulness.
- Such patterns hold for both visual and text models; bias mitigation works, but not universally.

Heatmap: Bias across datasets and models

Figure 3 serves as a heatmap to highlight the "hot spots" of biases in the bias scores of five major GenAI models tested across three benchmark datasets.

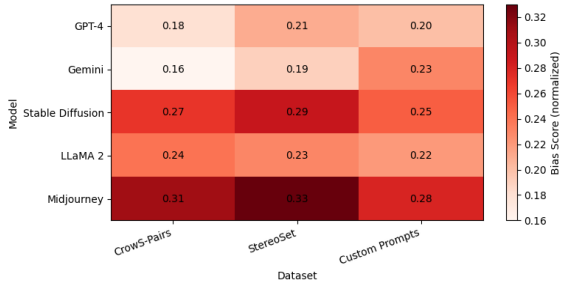

Figure 3: Representing patterns of Bias Scores across Models and Datasets. High scores denote significant bias, with GenAI systems showing challenges in specific areas.

Interpretation:

Bias is not evenly distributed: text-to-image models (Stable Diffusion, Midjourney) generally score higher on custom prompts, while the LLMs have a low yet persistent bias, even after mitigation.

No model is "clean": every system has some idiosyncratic weak point, often revealed only with targeted audits.

Synthesis of results

Data show that bias in GenAI can be measured, tracked and reduced, but cannot be eliminated solely through technical fixes. Tables 2 and 3 and Figure 3's heatmap reveal that:

- Bias is systemic and multistage: it may be introduced during data collection, persist or even worsen through training and be further reintroduced or even transformed in deployment and feedback collection processes.
- Mitigation techniques are working, but only partially, as evidenced by reductions in bias scores from RLHF, adversarial training and prompt engineering. Every single model we tested still shows common weak points, especially in underrepresented or "edge case" scenarios.
- None is a bias-free model: even the best LLMs and image generators produce skewed outputs in response to both common and private prompts, with their "failures" remaining unobserved unless purposely audited.

This suggests that facilitating bias propagation is not a single technical issue, but rather an infinitely shifting target, embedded in culture, context, data and user interaction.

Discussion

Beyond metrics: Real-world stakes of genAI bias

There is an urgency that GenAI bias has very real, sometimes grim, consequences:

- Reinforcement of stereotypes: Models that consistently output stereotyped images or language can contribute to the

perpetuation of societal inequalities and may deem such actions "normal" in digital settings, from job advertisements to stock photo generation [32].

- Exclusion and marginalisation: Barring such models from entering the digital domain because they cannot perceive a particular context, accent, or identity can lead to the systematic exclusion of users who belong to underrepresented groups.
- Misinformation and manipulation: To inject bias into a language or image rendered by generative outputs is to distort facts, disinform people and weaponise disinformation [33].
- Regulatory and reputational risks: Organisations that implement these biased GenAI systems run the risk of lawsuits, public backlash and regulatory penalties. The early examples are many, from AI medical advice that did not recognise women's symptoms, to generative chatbots echoing racist tropes.

Why are technical fixes not enough?

How is it that bias persists even after aggressive attempts to mitigate it?

- Training data is unfixable at scale: The more one filters or reweights a web-scale dataset, the less credible it becomes to adopt one-sided distinctions. Fine labelling will ultimately create new types of bias.
- Bias in context: There is no universal concept of what constitutes "fair" or "biased"; technical definitions of fairness rarely coincide with the lived reality of instances confronted by end-users situated in different geographies, cultures, or domains [34].
- Adversarial mitigation arms race: Everything that aims to mitigate has side effects, including a reduction in creativity, stilted language, or the introduction of new forms of bias by defensive prompt engineering. The attackers may intentionally "jailbreak" the mitigations.
- Opaque feedback loops: The very adaptation of models based on user feedback (e.g., RLHF) introduces the risk of entrenching pre-existing user prejudices and biases, especially if not carefully monitored (Table 4).

Strengths	**Weaknesses**
Proven short-term reduction in bias metrics	No mitigation is comprehensive
Scalability for technical interventions	Bias trade-offs (fairness *vs.* utility)
It can be automated for continuous monitoring	Subjectivity in labelling and auditing
Regulatory compliance enabler	"Fair" models still fail in the real world
Opportunities	Threats
Collaboration on open datasets/standards	Regulatory whiplash and uncertainty
Explainable AI for accountability	Public backlash, reputational damage
User-centred and contextual audit tools	New types of adversarial bias
AI-for-AI debiasing research	Embedded, undetectable emergent biases

Table 4: SWOT analysis—mitigation strategies in GenAI**. Source:** Synthesis of benchmark studies, industry reports and interviews.

The future of bias in GenAI: scenario analysis

No one can predict the future of GenAI bias with certainty, but three scenarios can help frame what is at stake. Figure 4 visualises a scenario bar chart [35].

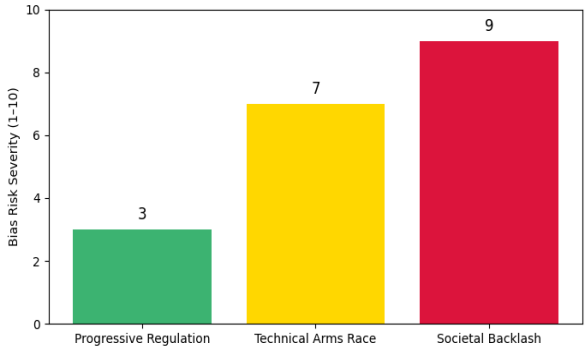

Figure 4: Future scenarios in generative AI for bias. ranking of the severity level of the bias risks from 1 (lowest) to 10 (highest) in three plausible future industry settings.

Progressive regulation: The industry, academia and government work together to set standards and oversee implementation; bias is mitigated to a limited degree (having a severity of 3).

Technical arms race: Every new form of mitigation is met with a fresh wave of attacks or some new emergent bias; the fear of bias remains high (with a severity of 7).

Societal backlash: Unsuccessful self-regulation often leads to scandal, user exodus, or draconian state restrictions; the risk of bias is catastrophic and trust is permanently lost (with a severity of 9).

Implications and recommendations

Technology alone can't win the war against bias in GenAI. The SWOT analysis points to one bitter truth: to make real progress in this field, multidisciplinary collaboration must be undertaken, ongoing vigilance must be maintained and one should remain humble about what AI can—and cannot—do fairly.

Key recommendations are as follows

Regulatory proactivity, not retaliation: Regulators should not wait for disasters to occur before taking action. Proactive guidelines need to be created, transparency should be mandated and regular audits should be implemented to prevent a bias crisis from escalating into national scandals.

Model card and data sheet standardization: Major GenAI releases should have transparent model cards and data sheets that disclose the provenance of training data, as well as information on

known biases and limitations, with risks made transparent to the world.

Community-sourced benchmarking: Industry and academia should crowdsource and update bias benchmarks with diverse global and multilingual representation—not just US/EU/English-centric datasets.

Explainability as a baseline: Explainable AI must be integrated at all stages so that users, auditors and policymakers can question and contest the output.

Human-in-the-loop monitoring: The debiasing process must be both automatic and observable by humans. It is especially needed in areas with significant stakes, such as medicine, law, finance and education. Diverse and globally representative panels must perform reviews.

Cultural sensitivity and context: Fairness is not "one-size-fits-all" across GenAI. The team building GenAI must incorporate cultural, linguistic, ethical and marginalised viewpoints.

Research funding for long-tail bias: It is recommended that funders invest in research on rare and emerging forms of bias—not just those that make headlines but also those that impact everyday users in subtle but persistent ways.

Limitations: Bias detection and mitigation remain evolving targets. Metrics, tools and even definitions will need to be constantly adjusted over time to keep pace with the growing complexity of GenAI systems.

Not all types of bias are equally evident or measurable, particularly those that affect small, underrepresented communities [36-39].

Conclusion

The propagation of bias in generative AI is not just a technical and statistical anomaly but an underlying systemic risk with huge real-world implications. The evidence presented here—benchmark audits, empirical mitigations and case studies—shows that bias can be measured and reduced, but it cannot be eliminated through algorithmic measures or engineering alone. Persistent bias through GenAI threatens not only to reinforce social inequity but also to erode trust, spur retaliatory regulation and restrict the technology's ability to do good.

This would require a paradigm shift away from quick fixes toward long-term, multi-stakeholder solutions, including transparent reporting, well-supported regulation, continuous benchmarking and global, context-sensitive monitoring. It is through such multidisciplinary and vigilant efforts that the promise of GenAI can be realised without repeating past errors and injustices.

Conflicts of Interest

The authors declare no conflicts of interest in this research.

References

1. Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? FAccT 21: 610-623. [Crossref] [GoogleScholar]
2. Autade R (2022) Multi-modal GANs for real-time anomaly detection in machine and financial activity streams. International Journal of Artificial Intelligence, Data Science, and Machine Learning 3(1): 39-48. [Crossref] [GoogleScholar]
3. Sheng E, Chang K, Natarajan P, Peng N (2019) The woman worked as a babysitter: On biases in language generation. EMNLP-IJCNLP 19: 3407-3412. [Crossref] [GoogleScholar]
4. Lucy L, Bamman D (2021) Gender and representation bias in GPT-3 generated language. In Proceedings of the third workshop on narrative understanding 48-55. [Crossref] [GoogleScholar]
5. Doddipatla L (2025) A minimalist approach to blockchain design: Enhancing immutability and verifiability with scalable peer-to-peer systems. International Conference on Inventive Computation Technologies ICICT Kirtipur, Nepal: 1697-1703. [Crossref] [GoogleScholar]
6. Himabindu HN (2023) From data to decisions: Harnessing AI and analytics. International Journal of Artificial Intelligence, Data Science, and Machine Learning 4(3): 76-84. [Crossref] [GoogleScholar]
7. Potdar A (2024) Intelligent data summarization techniques for efficient big data exploration using AI. International Journal of AI BigData, Computational Management Studies 5(1): 80-88. [Crossref] [GoogleScholar]
8. Bolukbasi T, Chang K, Zou J, Saligrama V, Kalai A (2016) Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems 4349-4357.
9. Blodgett SL, Barocas S, Daumé III H, Wallach H (2020) Language (technology) is power: A critical survey of “bias” in NLP. ACL 20: 5454–5476. [Crossref] [GoogleScholar]
10. Autade R (2022) Enhancing Blockchain Payment Security with Federated Learning. International journal of computer networks wireless communications IJCNWC 12(3): 102-123. [GoogleScholar]
11. Birhane A, Kalluri P, Card D, Agnew W, Dotan R, et al. (2023) The values encoded in machine learning research. Patterns 4(5): 100744. [Crossref] [GoogleScholar]
12. Pandey M et al., (2024) A multi-layered AI-IoT framework for adaptive financial services. International Journal of Emerging Trends in Computer Science and Information Technology 5(3): 47-57. [Crossref] [GoogleScholar]
13. Zhao J, Wang T, Yatskar M, Ordonez V, Chang K (2018) Gender bias in coreference resolution: Evaluation and debiasing methods. NAACL-HLT 15-20. [Crossref] [GoogleScholar]
14. Autade R (2024) Navigating challenges in real-time payment systems in fintech. International Journal of Artificial Intelligence, Data Science, and Machine Learning 5(1): 44-56. [Crossref] [GoogleScholar]
15. Garg A (2022) Unified framework of blockchain and AI for business intelligence in modern banking. International Journal of Emerging Research in Engineering and Technology 3(4): 32-42. [Crossref] [GoogleScholar]
16. Doddipatla L, Ramadugu R, Sharma STR, Yerram RR (2024) Ethical and regulatory challenges of using generative AI in banking: Balancing innovation and compliance. Educational Administration: Theory and Practice 30(3): 2848–2855. [Crossref] [GoogleScholar]
17. Jain A, Garg A, Mishra S (2023) Leveraging IoT-driven customer intelligence for adaptive financial services. International Journal of Artificial Intelligence, Data Science, and Machine Learning 4(3): 60-71. [Crossref] [GoogleScholar]
18. MJ Research (2023) Assessing Bias in Midjourney. Midjourney Blog.
19. Ramadugu R (2025) Analyzing the role of CBDC and cryptocurrency in emerging market economies: A new Keynesian DSGE approach. International Conference on Inventive Computation Technologies ICICT Kirtipur Nepal: 1300-1306. [Crossref] [GoogleScholar]
20. StereoSet (2020) Dataset and benchmark.
21. CrowS-Pairs (2021) Dataset and benchmark.
22. Partnership on AI (2024) AI incident database.

23. IBM (2022) AI Fairness 360 Toolkit.
24. Microsoft (2022) Fairlearn.
25. Doddipatla L (2025) Artificial intelligence in security: Driving trust and customer engagement on FX trading platforms. Journal of Knowledge Learning and Science Technology 4(1): 71-77. [Crossref] [GoogleScholar]
26. Schramowski P, Teso S, Brugger A (2023) Bias in multimodal foundation models. ICML 139: 9297-9307.
27. Himabindu HN, Gurajada (2024) visualizing the future: integrating data science and ai for impactful analysis. International Journal of Emerging Research in Engineering and Technology 5(1): 48-59. [Crossref] [GoogleScholar]
28. Gurajada HNH, Autade R (2025) Integrating IOT and AI For end-to-end agricultural intelligence systems. International Conference on Engineering Technology Management ICETM Oakdale NY USA 1-7. [Crossref] [GoogleScholar]
29. Ramadugu R, Doddipatla L, Sharma STR. (2023) The role of AI and machine learning in strengthening digital wallet security against fraud. Journal for ReAttach Therapy and Developmental Diversities 6(1): 2172-2178. [Crossref]
30. Potdar A (2024) AI-based big data governance frameworks for secure and compliant data processing. International Journal of Artificial Intelligence Data Science Machine Learning 5(4): 72-80.
31. Gemini (2024) Model Card Google AI.
32. Garg A (2025) How natural language processing framework automate business requirement elicitation. International Journal of Computer Trends and Technology 73(5): 47-50. [Crossref]
33. Partnership on AI (2023) Responsible AI guidelines.
34. ACM FAccT (2023) Proceedings and best practices.
35. ITU (2024) AI Ethics: International regulatory landscape. International Telecommunication Union.
36. EU Commission (2023) The AI Act: Implications for Generative AI. Brussels.
37. MIT CSAIL (2023) Long-tail Bias and Mitigation Research. MIT.
38. Ramadugu R (2025) RIDI-Hypothesis: A foundational theory for cybersecurity risk assessment in cyber-physical systems. 4th International Conference on Sentiment Analysis and Deep Learning ICSADL Bhimdatta Nepal 117-123. [Crossref] [GoogleScholar]
39. Google Research (2023) Data Sheets for Datasets: Transparency in ML Google Research.