Bias and Fairness in Generative Models: Ethical Challenges and Engineering Solutions

Intelligent Systems Engineering
A.Y. 2024/2025

Montanari Nicola
nicola.montanari14@studio.unibo.it

November 2025

Contents

1 Introduction

Generative Artificial Intelligence (GAI) has emerged as a transformative technology, demonstrating a remarkable capacity to autonomously produce human-like content across diverse domains such as text, images, code, and media.

GAI offers substantial benefits and influences critical decisions, particularly in fields such as education, healthcare, and the creative industries. However, it also introduces complex ethical challenges that necessitate careful consideration.

A prominent concern within GAI is the pervasive issue of bias and fairness.
Generative models, trained on vast datasets, can inadvertently learn and amplify societal stereotypes, prejudices, and historical inequalities, leading to discriminatory or unfair outcomes across various applications.
While bias and fairness are central, GAI's generative capabilities also amplify other significant ethical dilemmas, introducing new risks not typically posed by traditional AI systems. These include misinformation and deepfakes, data privacy violations, intellectual property issues, and challenges related to accountability and explainability.
Understanding these interconnected challenges is crucial for a holistic approach to responsible GAI development.

This systematic literature review focuses specifically on bias and fairness in GAI systems, recognizing them as key concerns within the broader ethical context. It explores the complex ethical issues surrounding generative models, highlighting key challenges as well as new technical solutions and mitigation strategies.

By providing a holistic overview, this review seeks to contribute to the discourse on responsible AI development, fostering a deeper understanding of how to build equitable and trustworthy generative AI systems for the future.

2 Background

This section aims to provide the foundational knowledge necessary to understand the ethical challenges and engineering solutions discussed in this systematic literature review.

It will introduce key concepts related to Generative AI, define what bias and fairness mean in an AI context, highlight the unique characteristics of generative models that amplify these issues, present concrete examples of biased outputs, and underscore the critical importance of fairness in this rapidly evolving field.

2.1 Introduction to Generative Artificial Intelligence (GAI)

Generative Artificial Intelligence (GAI) refers to a subset of deep learning models designed to generate new, realistic content (including text, images, audio, and code) based on patterns learned from vast training datasets.

Unlike traditional discriminative AI, which focuses on classifying existing data, generative models focus on creating novel data distributions that plausibly resemble real-world examples.

Key architectures driving this field include:

- **Generative Adversarial Networks (GANs)**: Frequently used for high-fidelity image synthesis
- **Variational Autoencoders (VAEs)**: utilized for structured data generation
- **Transformer-based models, such as Large Language Models (LLMs)**: Have revolutionized natural language processing tasks.

The capacity of these models to autonomously produce complex, human-like artifacts positions GAI as a transformative general-purpose technology.
However, this is accompanied by significant ethical risks.

2.2 Understanding Bias in Artificial Intelligence

In the context of AI, bias refers to systematic and repeatable errors in a computer system that create unfair outcomes, often privileging one arbitrary group of users over others.
It is distinct from random error, as it consistently disadvantages specific demographics.

Bias typically originates from three primary sources:

- **Data Bias**: Occurs when training datasets reflect historical societal prejudices or fail to adequately represent the diversity of the target population.
- **Algorithmic Bias**: It rises from choices made during model design, such as objective functions that unintentionally prioritize accuracy for the majority group at the expense of minorities.
- **Interaction Bias**: Emerges when models learn and adapt based on biased real-time user interactions.

2.2.1 Generative Models: Unique Challenges for Bias

GAI exacerbates standard AI bias due to its scale and methodology.
Generative models are often trained on uncurated, internet-scale datasets containing toxic language, stereotypes, and misinformation.

Rather than merely reproducing these biases, GAI can amplify them—a phenomenon in which models generate stereotypical outputs at a higher frequency than they appear in the training data. Furthermore, the "black box" nature of complex architectures like LLMs makes it difficult to trace specific biased outputs back to their root causes in the data, complicating mitigation efforts.

2.3 Concepts of Fairness in Artificial Intelligence

Fairness in AI is a complex, multi-faceted concept with no single, universally accepted technical definition.
It often involves trade-offs between competing mathematical objectives.

Common notions of fairness include:

- **Demographic Parity (Statistical Parity)**: Requires that a model's positive outcomes be distributed equally across different demographic groups, regardless of true underlying differences in the data.

- **Equalized Odds:** A model should have equal error rates (false positives and false negatives) across different groups.
- **Individual Fairness**: Focuses on ensuring that similar individuals are treated similarly by the model.

2.3.1 Unique Challenges and Importance in GAI

Applying these traditional metrics to GAI is uniquely challenging.
Unlike binary classification tasks, GAI outputs are high-dimensional and open-ended.

Measuring "demographic parity" in generated content requires complex (often subjective) evaluation of language nuances.
Despite these challenges, fairness is critical in GAI. As these tools are increasingly integrated into critical sectors like hiring, education, and healthcare, biased outputs can automate discrimination at an unprecedented scale, reinforcing historical inequalities and causing direct representational or allocational harm to marginalized communities.

Ensuring fairness is therefore not merely a technical optimization, but a fundamental requirement for a correct AI deployment.

2.4 Examples of Biased Outputs in Generative AI

Concrete examples illustrate the pervasiveness of these issues across modalities:

2.4.1 Text Generation

LLMs frequently exhibit occupational gender bias, defaulting to male pronouns for high-status roles (e.g., "CEO", "surgeon") and female pronouns for care-oriented roles (e.g., "nurse", "secretary").

They have also been shown to generate toxic or stereotypical text when prompted with African American Vernacular English (AAVE) compared to standard English.

2.4.2 Image Generation

Text-to-image models often fail to represent diversity unless explicitly prompted. For instance, prompts for "a standard family" or "professional person" may overwhelmingly yield images of white individuals.

Furthermore, they can generate harmful caricatures or hyper-sexualized images of certain demographics based on embedded cultural stereotypes.

3 Methodology

This chapter details the systematic approach taken to identify, select, and synthesize relevant literature concerning bias and fairness in generative models.
This methodology aims to ensure transparency and provide a clear framework for addressing the review's objectives.

3.1 Research Questions (RQs)

This systematic literature review is strictly guided by the following research questions, focusing on the core aspects of bias and fairness in generative models:

RQ1: *How do bias and fairness issues manifest and are characterized in generative models?*

RQ2: *What ethical issues arise from bias and fairness concerns in generative models?*

RQ3: *What engineering solutions and mitigation strategies have been proposed to address bias and fairness in generative models?*

3.2 Search Strategy and Information Sources

3.2.1 Information Sources

The electronic academic database ***Scopus***[1] was exclusively utilized for this systematic review. Scopus was chosen for its extensive coverage of peer-reviewed literature across scientific, technical, medical, and social science fields, providing a comprehensive collection of relevant publications.

3.2.2 Search Terms and Strategy

A carefully constructed search string was employed to retrieve articles most relevant to the review's research questions.

This string combined keywords related to generative AI with terms specific to bias, fairness, and potential solutions, using Boolean operators (AND, OR) and wildcards (*) to ensure both breadth and precision.
The keywords were derived directly from the core concepts embedded in RQ1, RQ2, and RQ3.

The comprehensive advanced search string was:

```
TITLE-ABS-KEY(("generative AI"
OR "generative artificial intelligence"
OR "large language model*"
OR "LLM*"
OR "GAN"
OR "diffusion model*"
OR "text generation"
OR "image generation"
OR "AI synthesis"
OR "synthetic data")

AND

("bias"
OR "fairness"
OR "discrimination"
OR "equity"
OR "debiasing"
OR "fairness algorithm*"
OR "bias detection"
OR "bias mitigation"
OR "fairness metric*"))
```

Explanation of Keywords:

- **TITLE-ABS-KEY(...):** This Scopus[1] operator ensures that the search terms appear in the article's title, abstract, or keywords.
- **Generative AI terms:** Relevant AI terms were included to cover the various technologies and outputs central to generative models.
  The wildcard * captures variations like "model" and "models."
- **Bias, Fairness, and Solutions terms:** These terms are specifically included to capture proposed engineering solutions (RQ3).
  The wildcard * covers plurals and related forms.
- **OR:** Used to combine synonymous or related terms within the same conceptual group.
- **AND:** Used to combine the two main conceptual groups (Generative AI *AND* Bias/Fairness/Solutions) to ensure that retrieved articles cover both aspects.

3.2.3 Search Execution

The search was conducted during October and November 2025 and was strictly limited to publications from January 1, 2024, to November 1 2025, to ensure a sharp focus on the very latest advancements, discussions, and solutions.

3.3 Study Selection and Eligibility Criteria

3.3.1 Study Selection

The rigorous selection process involved the following stages:

1. **Initial Identification and Duplication Removal:** All unique records obtained from the pages of the Scopus database[1], in accordance with the defined search strategy, were compiled.
   Duplicate entries were then systematically removed.

2. **Title and Abstract Screening:** The titles and abstracts of the remaining unique records were thoroughly screened against the predefined inclusion and exclusion criteria.
   Articles clearly irrelevant were excluded at this initial stage.

3. **Full-Text Review:** After screening titles and abstracts, potentially relevant papers were reviewed in full to confirm their eligibility based on the inclusion and exclusion criteria.

4. **Reference Chaining**: To identify any additional highly relevant studies not captured by the initial database search, the reference lists of the finally included articles were examined. These newly identified papers, if relevant, underwent the same screening process.

3.3.2 Inclusion Criteria

Studies were included if they met all of the following criteria:

- **Topic Focus:** Directly addressed bias and/or fairness challenges, their ethical implications, or proposed engineering solutions specifically within generative AI models.

- **Model Type:** Focused on generative AI models, such as Large Language Models (LLMs), Generative Adversarial Networks (GANs), diffusion models, Variational Autoencoders (VAEs), or other advanced generative architectures.
- **Content Type:** Presented empirical research, theoretical frameworks, methodological proposals, or comprehensive literature reviews.
- **Language & Publication Type:** Published in English as a peer-reviewed journal article, conference paper, or reputable workshop proceeding.
- **Publication Date:** Published between January 1, 2024, and November 1, 2025.

3.3.3 Exclusion Criteria

Studies were systematically excluded if they met any of the following criteria:

- **Irrelevant AI Focus:** Primarily focused on discriminative AI models without substantial discussion of generative applications.
- **Broader Ethical Scope:** Discussed general AI ethics or other ethical considerations (e.g., deepfakes, general privacy) without direct relevance to bias and fairness in generative models.
- **Non-academic:** Were opinion pieces, editorials, news articles, blog posts, or white papers without empirical or robust theoretical backing.
- **Availability/Redundancy:** Not available in full text, were duplicate publications, or early versions superseded by a final, more comprehensive publication.
- **Out of Date Range:** Published outside the specified date range.

3.4 Data Extraction and Synthesis

3.4.1 Data Extraction

From each included study, key information was extracted.
This focused on details directly answering the research questions:

- **Generative Model Details:**
  Type of generative AI model and its application area.

- **Bias and Fairness Characterization (RQ1):**
  How bias or fairness issues were identified, described, and exemplified.

- **Ethical Consequences (RQ2):**
  The specific ethical problems arising from these bias and fairness issues.

- **Engineering Solutions (RQ3):**
  Proposed technical methods or strategies to address bias and fairness.

3.4.2 Data Synthesis

A narrative synthesis, combined with thematic analysis, was used to organize and interpret the collected information. The findings were grouped into clear themes that matched the research questions. This process gave a structured summary of the literature and informed the "Findings" and "Discussion" chapters of this review.

4 Findings

This chapter presents a systematic synthesis of the extracted data from the literature selected, directly addressing the research questions of this review.
The findings are organized to detail the manifestations of bias and fairness issues, their resulting ethical implications, and the proposed engineering solutions in generative models.

4.1 Manifestations and Characterization of Bias and Fairness in Generative Models (RQ1)

Bias and fairness issues are inherent and often amplified within generative models, leading to systematic errors that result in unfair outcomes. These issues manifest through various forms, primarily influenced by training data and model design.

4.1.1 Types of Bias Identified

Generative models tend to associate specific traits or behaviors with certain social groups, reflecting and frequently amplifying stereotypes embedded in their training data. Key manifestations of bias include:

- **Representational Bias:**[2] Generative models inaccurately or incompletely portray demographic groups, or even omit their presence. Similarly, models can misrepresent ethnic groups or generate stereotypical text when prompted with minority dialects.
- **Output Bias:**[2] This bias reflects internalized stereotypes. LLMs have been shown to produce recommendation letters that describe individuals with typically female names as "warm and amiable" but those with male names as "leaders and innovators". This category also includes the generation of toxic, derogatory, or offensive language.
- **Stereotype Amplification:**[5] Generative models often do not just mirror existing biases but can amplify them, creating outputs that are more stereotypical than the original training data.
- **Disparate System Performance:**[4] Bias can manifest as degraded understanding or generation of language quality for specific social groups or linguistic variations, as evidenced by some models misclassifying minority dialects more often than standard English.

4.1.2 Sources of Bias

Bias can infiltrate generative models at multiple stages of their development and deployment lifecycle[3]:

- **Training Data:** This is consistently identified as a principal source. Datasets, often massive and uncurated from the internet, inherently reflect historical, societal, and structural biases, leading to underrepresentation, omission of contexts, or the use of biased proxies.
- **Model Design and Learning Procedures:** The dense vector representations learned by models can subtly carry and propagate biases. Algorithmic bias can also arise from choices in model design, such as optimization functions that inadvertently prioritize accuracy for majority groups over minority ones.
- **Human Annotations and Feedback:** When human annotators label data or provide feedback for model alignment, their inherent subjective beliefs and stereotypes can be inadvertently introduced into the model's behavior.
- **Evaluation and Deployment:** Unrepresentative benchmark datasets can steer model development towards optimizing for dominant groups. Furthermore, the context of deployment and user interaction patterns can influence the manifestation and perception of bias.

4.1.3 Characterization of Fairness Concepts

Fairness in AI is a complex, normative concept that lacks a single, universally accepted technical definition, often being subjective and context-dependent. Literature highlights various computational notions[3][4]:

- **Group Fairness:** This concept focuses on achieving equitable outcomes across predefined demographic groups. Common instantiations include *Demographic Parity*, which aims for positive outcomes to be equally likely for all groups, and *Equalized Odds*, which requires equal false positive and false negative rates across groups.
- **Individual Fairness:** This emphasizes treating similar individuals similarly by the model, irrespective of their group affiliation.
- **Counterfactual Fairness:** This newer concept seeks to ensure that a model's decision for an individual would remain constant even if their protected attributes were hypothetically altered.

A key challenge in characterizing fairness in generative models lies in the difficulty of applying these metrics to the high-dimensional, often subjective, and creative outputs of GAI, a task considerably more complex than in traditional classification problems.

4.2 Ethical Issues Arising from Bias and Fairness in Generative Models (RQ2)

Bias and fairness concerns in generative models are not merely technical flaws. They translate into significant ethical challenges that can inflict profound societal and individual harms, compromising the trustworthiness and equitable deployment of GAI systems.

4.2.1 Discriminatory Outcomes and Harm

The most direct ethical consequence is the pervasive potential for discrimination and the perpetuation of existing inequalities.
This manifests in both representational and allocational harms[3]:

- **Representational Harms:** These arise from the generation of content that perpetuates harmful stereotypes, misrepresents, or erases certain social groups, thereby reinforcing existing inequalities. Examples include the amplification of negative or immutable abstractions about groups and the production of derogatory or toxic language.
- **Allocational Harms:** These refer to the inequitable distribution of resources, opportunities, or access to essential services due to biased GAI outputs. This can appear as direct discrimination, where explicit disparate treatment occurs due to underlying, implicit biases, potentially exacerbating inequities in areas such as healthcare or finance.

These discriminatory outcomes can limit individual freedoms, reinforce societal power dynamics, and restrict access to critical services for marginalized communities.

4.2.2 Erosion of Trust and Accountability Challenges

Beyond direct discrimination, bias in generative models presents broader systemic ethical challenges[4]:

- **Erosion of Public Trust:** The widespread deployment of biased GAI systems can severely undermine public trust in AI technology. If these systems

are perceived as tools for discrimination, their adoption, and the realization of their potential benefits could be significantly hindered, potentially leading to societal rejection.

- **Challenges of Accountability:** The "black box" nature of many complex generative models makes it difficult to ascertain why a particular biased output was generated, complicating efforts to assign responsibility. This highlights the critical need for ethical guidelines and regulatory frameworks to hold developers and deploying organizations accountable for discriminatory outcomes.

- **Fairness-Performance Trade-offs:** A recurring ethical dilemma is the inherent tension between achieving fairness and maintaining optimal model performance. Adjusting models for fairness may unintentionally degrade performance for certain groups or tasks, necessitating careful consideration of ethical priorities and the potential for unintended consequences during mitigation.

4.3 Engineering Solutions and Mitigation Strategies for Bias and Fairness (RQ3)

The literature found proposes a comprehensive array of engineering solutions and mitigation strategies aimed at detecting, measuring, reducing, and preventing bias, and enhancing fairness in generative models.
These approaches are broadly categorized by their intervention point within the AI development pipeline.

4.4 Pre-processing Mitigation (Modifying Model Inputs)

Pre-processing strategies address bias by modifying the data or prompts before the model undergoes training or inference[2][3][4][6]:

- **Data-Centric Solutions:** These include *Data Augmentation* techniques, such as *Counterfactual Data Augmentation (CDA)*, which generates new data instances by swapping protected attributes (e.g., gendered pronouns) to balance representations. Also, *Data Filtering and Reweighting* modify existing datasets by removing biased or harmful texts, or by reweighting instances to increase the influence of underrepresented examples. Furthermore, *Data Generation* focuses on producing entirely new datasets curated to specific ethical standards.

- **Prompt-Centric Solutions:** This involves *Instruction Tuning*, where specific textual instructions or "control tokens" are added to prompts to guide

the model toward less biased outputs. Alternatively, *Continuous Prompt Tuning* utilizes trainable prefixes that dynamically update to reduce bias without altering core model parameters.

- **Representation-Level Solutions:** *Projection-based Mitigation* techniques transform contextualized embeddings to remove dimensions of bias by projecting them onto a subspace corresponding to protected attributes.

4.5 In-training Mitigation (Modifying Model Parameters)

These methods integrate fairness constraints directly into the model's training procedure, typically requiring retraining or fine-tuning[3]:

- **Loss Function Modification:** The model's objective function is adjusted to promote fair representations. This can involve *Equalizing Objectives* that encourage independence between social groups and predicted outputs.
- **Architectural Changes:** This includes *Architecture Modification*, where specialized layers are integrated and trained for debiasing while keeping original model parameters fixed.
- **Learning Paradigms:**
  - **Adversarial Learning:** A *discriminator* predicts protected attributes from model representations, while the main model is trained to fool this *discriminator*, learning independent representations.
  - **Contrastive Learning:** Focuses on learning representations that group similar instances and separate dissimilar ones based on sensitive attributes.
  - **Reinforcement Learning from Human Feedback (RLHF):** Utilizes human input to train a reward model that guides the generative model towards unbiased text aligned with human values.
- **Parameter Management:** *Selective Parameter Updating* freezes most model parameters, fine-tuning only a subset contributing to bias reduction. *Filtering Model Parameters* identifies and removes specific parameters linked to bias.

4.6 Intra-processing Mitigation (Modifying Inference Behavior)

These techniques adjust the model's behavior during the inference stage without requiring additional training, often by altering the decoding process[2][3]:

- **Decoding Strategy Modification:** This involves altering how the model generates outputs to enforce fairness. Techniques include *Constrained Next-token Search*, which prohibits the generation of offensive tokens or re-ranks

token probabilities, and *Discriminator-based Decoding*, which uses auxiliary classifiers in real-time to guide the model towards less harmful outputs.

- **Modified Token Distribution:** Adjusting token probabilities during decoding can increase output diversity or reduce biased sampling.
- **Modular Debiasing Networks:** These integrate stand-alone debiasing components with the original model at inference time to remove specific dimensions of bias.

4.7 Post-processing Mitigation (Modifying Output Text Generations)

Post-processing methods operate on the model's generated outputs after they have been produced, making them particularly useful for "black box" models where internal access is limited[2][3][4]:

- **Rewriting Techniques:** These methods identify and modify biased or discriminatory language in the model's outputs. This can involve *Keyword Replacement*, where biased tokens are substituted with neutral alternatives while preserving context and style, or *Machine Translation for Debiasing*, which trains models to translate biased source sentences into debiased versions.

4.8 Cross-cutting Challenges in Mitigation

The implementation of these diverse mitigation strategies consistently encounters several challenges:

- **Fairness-Performance Trade-offs:** A pervasive issue is the inherent tension between improving fairness and maintaining other performance metrics like accuracy or efficiency. Adjustments for fairness often require careful balancing and manual tuning, potentially leading to performance degradation in some contexts.
- **Lack of Consensus on Fairness:** The absence of a universal, context-independent definition for "fairness" poses a significant hurdle in designing and evaluating solutions that satisfy all stakeholders, as different fairness metrics can conflict.
- **Computational Expense and Scalability:** Many advanced in-training mitigation strategies are resource-intensive. Furthermore, methods relying on limited word lists or human annotations can restrict the scope of biases and social groups effectively addressed.

- **Limitations of Current Methods:** Specific techniques have inherent limitations; for example, embedding-based debiasing may have weak correlations with downstream biases, while rewriting techniques can sometimes introduce new biases or oversimplify linguistic nuance, potentially marginalizing minority voices.
- **Unintended Consequences:** Attempts to mitigate one form of bias may inadvertently introduce or amplify other biases, or lead to less diverse or accurate outputs.

5 Discussion

The current knowledge confirms that bias and fairness are central, complex challenges in generative models, particularly Large Language Models.

The findings underscore that these models frequently exhibit and even amplify various forms of bias, often rooted in skewed training data, model design choices, and human annotations.
These biases are not mere technical glitches but lead to significant ethical concerns, ranging from direct discriminatory outcomes and the perpetuation of societal inequalities to a broader erosion of public trust and challenges in accountability.

Generative models, by their very nature of creating novel content, can translate subtle statistical biases from their training data into tangible representational and allocational harms. Stereotypes, for instance, are not just reflected but can be amplified in generated text or images, leading to misrepresentation or exclusion of marginalized groups.

This creates a direct pathway to ethical problems and can deepen existing societal inequalities.

The dynamic, high-dimensional outputs of generative models make the application and measurement of various fairness definitions particularly complex, highlighting a persistent challenge in moving from abstract ethical principles to concrete technical implementation.

Regarding engineering solutions and mitigation strategies, the literature demonstrates a robust and multi-faceted response, categorized across multiple stages.
However, the review highlights several enduring challenges within these solutions.

The most prominent is the pervasive fairness-performance trade-off, where efforts to enhance fairness often risk degrading other performance metrics.
Furthermore, the lack of a universal consensus on fairness definitions continues to complicate the design and evaluation of universally applicable mitigation strategies. Many advanced solutions also face computational expense and scalability issues, limiting their broad applicability.

5.1 Limitations of This Review

This systematic literature review has certain limitations.

The search was exclusively conducted using the *Scopus*[1] database and limited to publications from January 1, 2024, to November 1, 2025.
While this ensured a focus on the most recent advancements, it may not capture relevant literature indexed predominantly in other databases or foundational works published outside this narrow timeframe.

Furthermore, the review strictly adhered to its focus on bias and fairness, excluding a detailed exploration of other ethical considerations in generative AI.

5.2 Future Research Directions

Based on the identified gaps and recurring challenges in the literature, several areas warrant future research:

- **Empirical Comparison of Solutions:** There is a need for more rigorous empirical studies that compare the long-term effectiveness, trade-offs, and unintended consequences of different mitigation strategies across various generative models and application domains.
- **Context-Aware Fairness Metrics:** Further development of standardized, context-aware fairness metrics and evaluation benchmarks specifically tailored for the high-dimensional and subjective outputs of generative models is crucial.
- **Intersectional Bias Mitigation:** Research should expand beyond single-axis biases to address intersectional biases, understanding how different dimensions of identity interact to create unique challenges and requiring more nuanced mitigation strategies.
- **Transparency and Explainability in Solutions:** Developing more interpretable and explainable bias detection and mitigation techniques would enhance accountability and help practitioners understand why a particular intervention is effective or how it might introduce new biases.
- **Holistic and Hybrid Approaches:** Future work should explore hybrid mitigation techniques that intervene at multiple stages of the AI lifecycle, moving towards more holistic strategies that consider problem formulation, data collection, and deployment alongside model-centric interventions.

6 Conclusions

This systematic literature review comprehensively examined bias and fairness in generative models, exploring their ethical challenges and proposed engineering solutions.

The findings highlight that generative AI, despite its transformative potential, inherently manifests biases stemming from training data, model design, and human input. These biases translate into significant ethical issues.

The review identified a diverse landscape of engineering solutions, categorized according to the stage of the AI development pipeline in which each solution is implemented.

Ultimately, addressing bias and achieving fairness in generative models remains a dynamic and multifaceted challenge. It necessitates a holistic approach that moves beyond superficial fixes, demanding continuous vigilance, interdisciplinary collaboration, and the ongoing development of robust, context-aware strategies.

The future of generative AI hinges on a concerted commitment to building systems that are not only powerful but also fundamentally equitable, trustworthy, and beneficial for all of society.

7 References

1 Scopus Database
https://www.scopus.com/

2 A Comprehensive Survey on Bias and Fairness in Large Language Models
https://www.scopus.com/pages/publications/105009851907

3 Bias and Fairness in Large Language Models: A Survey
https://www.scopus.com/pages/publications/85209882813

4 Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies
https://www.scopus.com/pages/publications/85189164092

5 Evaluating Fairness and Bias in Large Language Models for Tabular Data
https://www.scopus.com/pages/publications/105007711462

6 The Ethics of Generative AI: Analyzing ChatGPT's Impact on Bias, Fairness, Privacy, and Accountability
https://www.scopus.com/pages/publications/105001367528