

Article | Received 6 August 2024; Accepted 12 September 2024; Published 30 September 2024
<https://doi.org/10.55092/let20240008>

Mitigating bias in generative AI: a comprehensive framework for governance and accountability

Shuchen Tang^{1,*}, Haoming Zhu²

¹ Law School, Tsinghua University, Beijing, China

² Law chamber, Monash University, Melbourne, Australia

Corresponding author; E-mail: tsc22@mails.tsinghua.edu.cn.

Abstract: The pervasive risk of bias in generative artificial intelligence (AI) systems necessitates robust measures to protect public rights and enhance regulatory effectiveness. Addressing bias across the lifecycle of AI products—from data collection and training to modeling and application—requires legal and technical strategies tailored to each layer of potential bias. Current regulatory frameworks, such as China's “Interim Measures for the Administration of Generative AI Service,” lack specific guidelines to mitigate bias in AI decision-making. Global regulatory frameworks are still developing, underscoring the need for a comprehensive governance structure that defines the scope of regulation, implements layered measures to address bias, and allocates liability among platform developers.

Keywords: generative AI; Bias; AI governance

1. Introduction

The rapid development of generative artificial intelligence (AI) systems has transformed various industries, offering unprecedented capabilities in content creation, decision-making, and automation. However, the deployment of these systems has brought the issue of bias to the forefront, a critical problem that risks exacerbating existing social inequalities. Bias in generative AI systems can emerge at different stages of the AI lifecycle—data collection, model development, and application—causing the systems to produce skewed outputs that disadvantage certain groups [1]. This paper addresses the multifaceted problem of AI bias by proposing a comprehensive, layered framework for governance and accountability, aimed at mitigating bias across all stages of the AI system lifecycle.

A growing body of research has identified bias as a central issue in AI systems, particularly those that rely on large datasets harvested from the internet. These datasets often contain inherent biases, reflecting societal stereotypes, underrepresentation of certain demographics, or culturally biased labeling practices [2]. For example, Ferrer *et al.* highlight how biases present in training data can lead to discriminatory outputs [1], while Fang *et al.*



Copyright©2024 by the authors. Published by ELSP. This work is licensed under Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

demonstrate the risk of bias in AI-generated news, showing that language models may reinforce societal stereotypes [3].

Bias can also arise during the model development stage. Bird *et al.* note that model architecture and training objectives are often designed without sufficient consideration for fairness, amplifying biases from the data layer [4]. Moreover, emergent behaviors in complex AI models can introduce new biases that are difficult to detect and mitigate, as large-scale systems may exhibit unanticipated discriminatory patterns [5]. Existing regulatory frameworks have made progress in addressing these issues but remain insufficient. For instance, China's Interim Measures for the Administration of Generative AI Services represent an important step toward regulating generative AI but lack specific guidelines on bias mitigation, especially in decision-making processes [6]. Similarly, the European Union's Artificial Intelligence Act (AIA) adopts a risk-based approach to AI regulation, imposing stringent requirements for high-risk applications but offering limited provisions for generative AI [7]. Comparative studies suggest that while the AIA provides a structured regulatory framework, it requires further refinement to address the unique challenges posed by generative AI systems, especially those involving bias.

While substantial progress has been made in recognizing and attempting to mitigate bias, existing research highlights key gaps in both technical solutions and regulatory frameworks. Al-kfairy *et al.* underscore the need for interdisciplinary approaches that combine ethical principles with technical interventions to address bias effectively. Their work advocates for ethical auditing of AI systems, yet it stops short of providing concrete technical solutions that can be systematically implemented during AI model development [8]. Similarly, Laine *et al.* call for more robust ethics-based AI auditing frameworks, which focus on transparency and stakeholder engagement [9]. However, these frameworks often face practical challenges, such as the complexity of auditing advanced generative AI systems and ensuring accountability throughout the AI lifecycle. In the technical domain, bias detection and mitigation strategies have advanced, but they often remain fragmented. Techniques such as adversarial training, which exposes models to challenging counterexamples to correct biased outputs, have shown promise in reducing model bias [10]. However, Mustafa *et al.* argue that current technical solutions, such as counterfactual fairness and reweighting algorithms, may only address bias at a superficial level [11]. They contend that deeper architectural changes are required to tackle structural biases embedded within AI systems, especially in large-scale generative models. Moreover, methods like these are still largely under-researched when it comes to their practical deployment in real-world AI systems.

From a regulatory perspective, comparative studies of different jurisdictions reveal significant variation in how AI bias is approached. Fang *et al.* emphasize that the AIA, though pioneering in its structured approach, may lack the flexibility needed to adapt to the rapidly evolving nature of AI technologies [3]. Specifically, the AIA's emphasis on high-risk applications does not adequately address the unique challenges posed by generative AI systems, where bias can emerge not only in high-risk settings but also in consumer-facing applications. Meanwhile, Mustafa and Al-kfairy suggest that the United States, despite its

sector-specific guidelines, lags behind in creating comprehensive AI regulations, relying more on industry-led self-regulation which lacks uniformity and enforceability [11].

Another dimension of bias mitigation that has received less attention is the post-deployment phase of AI systems. Buolamwini and Gebru stress the importance of ongoing monitoring and auditing of AI systems after they are deployed, as biases can emerge or evolve over time as systems interact with new data or users [12]. However, post-deployment bias mitigation remains under-regulated in most jurisdictions. For example, China's Interim Measures require AI service providers to adopt continuous measures to prevent bias, but they lack specific mechanisms for enforcing or auditing such requirements in the long term. Gebru *et al.* further highlight that a robust governance framework must incorporate clear guidelines for continuous auditing, independent evaluations, and publicly accessible reporting mechanisms to ensure sustained accountability [13].

Overall, while regulatory frameworks and technical solutions have evolved to address AI bias, there remain significant gaps that necessitate more comprehensive and actionable approaches. The technical challenges of bias mitigation call for novel methods that go beyond transparency and accountability, incorporating direct interventions into model architecture and training processes. At the same time, regulatory frameworks need to evolve to accommodate the unique complexities of generative AI, with clearer guidelines for bias mitigation throughout the AI lifecycle, including post-deployment monitoring and enforcement mechanisms.

This study is motivated by the gaps in existing research and regulatory frameworks, as well as the growing concern over AI fairness and accountability. The paper aims to answer the following research questions:

What are the primary sources of bias in generative AI systems across the data, model, and application layers?

How can regulatory frameworks be enhanced to more effectively mitigate bias?

What technical solutions can complement governance measures to ensure fairness, transparency, and accountability?

To address these questions, this research employs a mixed-methods approach. First, a legal analysis of national and international AI regulatory frameworks is conducted, focusing on their strengths and weaknesses in handling bias in generative AI systems. This analysis covers China's Interim Measures, the EU's AI Act, and various sector-specific guidelines in the United States. Second, the study conducts a technical review of bias mitigation techniques, including adversarial debiasing, counterfactual fairness, and data augmentation methods. These techniques are evaluated for their potential to address bias at different stages of the AI lifecycle. Through this interdisciplinary approach, the paper seeks to offer a layered governance model that integrates both legal and technical strategies. The proposed framework aims to enhance existing regulatory measures by incorporating specific actions to address bias, while also providing technical solutions that can be implemented during AI system development and deployment.

This study contributes to the literature on AI governance by proposing a comprehensive framework that addresses bias from multiple perspectives—legal, technical, and regulatory.

The paper is structured as follows: Section 2 provides an in-depth review of existing AI bias mitigation strategies, while Section 3 outlines the regulatory landscape, offering a comparative analysis of key jurisdictions. Section 4 presents the proposed framework, detailing how each layer of the AI lifecycle can be governed to minimize bias. The paper concludes with recommendations for policymakers and AI developers, emphasizing the need for coordinated efforts across technical and regulatory domains.

2. The origins of bias in generative AI

Bias in generative AI systems arises from various stages of the AI lifecycle, primarily from three layers: data, model, and application [14]. In Figure 1, it can be seen that each layer contributes uniquely to the development and perpetuation of bias, necessitating targeted interventions to mitigate these biases effectively.

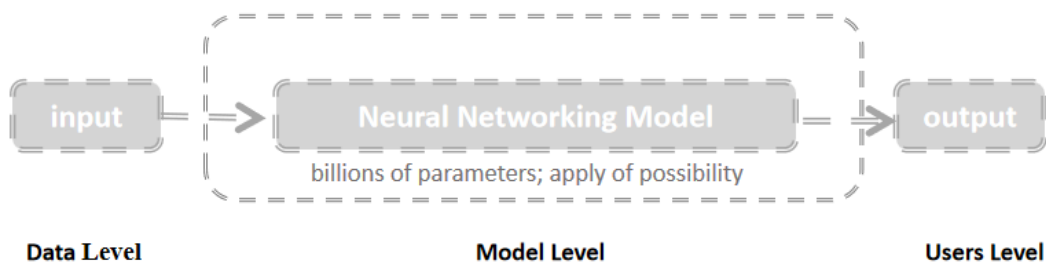


Figure 1. Bias comes from 3 layers in Generative AI systems.

2.1. Data layer

Bias originating at the data layer is a critical concern in the development of generative AI systems, as it directly stems from the datasets used to train these models. Biases at this stage can manifest in various forms, including data sampling bias, representation bias, and labeling bias, each of which uniquely contributes to the propagation of biased outputs.

One of the most significant sources of bias at the data layer is data sampling bias, which occurs when the datasets used for training are not representative of the broader population. For example, datasets that predominantly feature English-language content or data from urban environments may lead to AI models that fail to accurately reflect the needs and experiences of non-English speakers or rural populations [15]. This type of bias often arises due to the over-representation of certain demographic groups on the internet, while others, such as older adults or less-educated individuals, remain underrepresented [16].

Representation bias is another critical issue at the data layer. The internet, serving as a primary source of training data, is filled with content that mirrors societal stereotypes and prejudices. Consequently, AI systems trained on such data may inadvertently internalize and perpetuate these biases [17]. For instance, if training data contains stereotypical portrayals of gender, race, or ethnicity, the AI model may generate outputs that reinforce these biases [18].

This form of bias is particularly harmful because it embeds existing societal prejudices deep within the AI system, making detection and correction more difficult.

Labeling bias further exacerbates bias at the data layer. Human annotators, who are essential in the labeling process, bring their subjective perspectives, which can influence the labels they assign. Such subjectivity can lead to inconsistent or incorrect labels, resulting in AI models producing biased outputs [19]. For example, annotators from different cultural backgrounds may label the same image or text differently, introducing cultural biases into the training data [20].

The scale at which generative AI models operate further intensifies the impact of bias at the data layer. Large-scale datasets, often composed of billions of data points, present significant challenges for identifying and correcting biases. The sheer volume of data means that even small biases can be amplified, resulting in considerable skew in the model's outputs [21]. Additionally, the complexity of these models can lead to interactions between biases in unpredictable ways, compounding their overall effects [22].

Efforts to mitigate bias at the data layer must focus on improving the diversity and representativeness of training datasets. This involves actively seeking out and incorporating data from underrepresented groups and contexts [23]. Moreover, it is essential to establish rigorous standards for data collection and annotation to reduce the introduction of subjective biases [24]. Advanced techniques such as adversarial training and counterfactual data augmentation can also play a critical role in identifying and correcting biases during the training process [10].

In conclusion, the data layer forms a foundational aspect of generative AI systems where bias can be both introduced and propagated. Addressing these biases requires a multifaceted approach that encompasses improving dataset representativeness, enforcing rigorous annotation standards, and employing advanced bias detection and correction techniques. By tackling bias at this layer, it is possible to develop more equitable and unbiased generative AI systems.

2.2. Model layer

Bias in generative AI systems can also originate from the model layer, where key decisions are made during the design, development, and training of AI models. This layer introduces bias through several mechanisms, including modeling decisions, the amplification of data biases, and emergent behaviors in complex models.

Modeling decisions are critical junctures where bias can be introduced. These decisions include choices made by developers regarding problem specification, training objectives, and model architecture. For example, the criteria used to define a "successful" model output may reflect developers' biases or subjective judgments [25]. If the primary goal of model development is to maximize accuracy without considering fairness, the resulting AI models may perpetuate existing biases present in the training data [26].

The process of algorithm development itself can further contribute to bias. The selection of features, formulation of algorithms, and tuning of model parameters are all influenced by

developers' perspectives and the technological constraints they face [27]. For instance, choosing a particular machine learning algorithm might inherently favor certain data representations, inadvertently reinforcing existing biases [28]. Additionally, a lack of diversity within AI development teams can narrow perspectives on what constitutes bias, leading to decisions that do not account for the experiences or needs of all user demographics [29].

A significant issue at the model layer is the amplification of biases present in the training data. When models are trained on biased datasets, these biases can be magnified through the learning process [30]. For instance, if a model is trained on data that underrepresents certain demographic groups, the model may perform poorly on data from those groups, further entrenching disparities [31]. This creates a feedback loop in which biased model outputs are used to generate new training data, exacerbating the bias over time [32].

Emergent behaviors in complex AI models present additional challenges. As models become more sophisticated, they may exhibit unexpected behaviors that were not anticipated during development [33]. These behaviors can include new forms of bias that were not present in the training data or earlier model versions. For example, large language models may generate biased content or make biased decisions based on subtle correlations in the data that were not explicitly programmed [34]. These emergent biases are particularly difficult to predict and mitigate due to the interactions between various components of the model.

Addressing bias at the model layer requires a multifaceted approach. One key strategy involves the implementation of fairness-aware machine learning techniques, which explicitly incorporate fairness criteria into the model development process [35]. These techniques may include fairness constraints to ensure equitable treatment across demographic groups or optimization methods that balance accuracy with fairness [36]. Additionally, regular audits and evaluations of AI models are essential to identify and address biases that may emerge during deployment [37].

Transparency and accountability are also crucial in mitigating bias at the model layer. Developers should document their modeling decisions, such as algorithm selection and parameter tuning, along with the rationale behind these choices [38]. This documentation helps external auditors and stakeholders understand potential sources of bias and hold developers accountable. Furthermore, involving diverse teams in the development process can offer broader perspectives on potential biases and ensure that models are evaluated against a wider range of criteria [27].

In conclusion, bias at the model layer of generative AI systems is a complex issue stemming from modeling decisions, the amplification of data biases, and emergent behaviors in large-scale models. Mitigating this bias requires a combination of fairness-aware machine learning techniques, transparency, accountability, and diverse development teams. By addressing these factors, it is possible to create AI models that are more equitable and less likely to perpetuate harmful biases.

2.3. Application layer

Bias in generative AI systems can also be introduced at the application layer, where AI are deployed and utilized in real-world contexts. This layer encompasses user interactions, decision-making processes, and the environments in which AI outputs are applied. Several factors contribute to the propagation of bias at this stage, including user reliance on AI outputs, design and testing procedures, and the diversity of development teams.

One of the primary sources of bias at the application layer is user interaction. When users engage with AI systems, they may unknowingly reinforce biases present in the outputs. This is particularly problematic in decision-making scenarios such as hiring, lending, and law enforcement [39]. Users often assume AI-generated recommendations to be objective and unbiased, which can perpetuate existing social hierarchies and systemic discrimination, as biased outputs reinforce users' preconceived notions.

The design and testing of AI systems also play a critical role in the introduction of bias at the application layer. If the test cases and decision-making processes are not developed with diversity and inclusion in mind, the system may produce biased outcomes. For instance, if an AI system used in hiring is trained and tested primarily on data from a homogeneous group, it may perform poorly for candidates from diverse backgrounds [27]. This can result in discriminatory practices and reduced workplace diversity.

The diversity of development teams is another crucial factor. Homogeneous teams may overlook or underestimate the impact of bias on various social groups. For example, a development team consisting primarily of men may not fully consider how their AI system might disadvantage women [12]. In contrast, diverse teams are more likely to recognize and address potential biases, leading to more equitable AI systems [40].

Additionally, the context in which AI systems are applied can exacerbate biases. In settings where AI is used without proper oversight or regulatory guidelines, the risk of biased decision-making increases. For example, when AI is used in judicial settings to assist with sentencing, biased outputs may lead to unjust outcomes for certain demographic groups [41]. Therefore, proper oversight and clear guidelines on how to interpret and apply AI-generated outputs are essential for mitigating these risks.

To effectively address bias at the application layer, several strategies are necessary. A key initial step is educating users about the inherent biases in AI systems and promoting critical evaluation of AI-generated outputs. Training programs that emphasize the limitations of AI and the need for human oversight can facilitate this education [42]. Additionally, design and testing phases should prioritize diversity and inclusion by utilizing broad datasets and considering multiple perspectives. This helps mitigate biases before and during deployment. Regular audits and evaluations are also crucial, as they can detect and rectify biases that may emerge over time.

Moreover, diverse development teams are critical in minimizing bias. Heterogeneous teams ensure a variety of perspectives are considered, which reduces the likelihood of systemic bias. Organizations should adopt policies that promote diversity and inclusion within their teams to ensure fairer AI outcomes. Lastly, robust oversight and regulatory

frameworks are indispensable. Comprehensive standards for AI usage in sensitive areas such as employment, finance, and law enforcement must be established. These frameworks should mandate transparency, accountability, and periodic audits to ensure that AI systems operate fairly and ethically.

In conclusion, the application layer is a critical stage where bias in generative AI systems can be introduced and propagated. Addressing bias at this layer involves educating users, designing and testing AI systems with diversity and inclusion in mind, building diverse development teams, and implementing proper oversight and guidelines. Tackling these issues can lead to more equitable and unbiased AI systems that better serve the diverse needs of all users.

3. Regulatory approaches to bias in generative AI

Addressing bias in generative AI systems necessitates comprehensive regulatory frameworks that incorporate technical, ethical, and legal considerations. Different jurisdictions have begun to develop regulations to tackle the complex issue of bias in AI, with notable efforts seen in China, the European Union, and the United States. These approaches highlight varying strategies and principles for ensuring fairness, accountability, and transparency in AI systems.

3.1. European union

The European Union (EU) has taken a proactive stance on AI regulation through the development of comprehensive frameworks such as the "Ethics Guidelines for Trustworthy AI" and the "Artificial Intelligence Act" (AIA). The Ethics Guidelines for Trustworthy AI, drafted by the High-Level Expert Group on AI, outline seven key requirements for trustworthy AI: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and fairness, societal and environmental well-being, and accountability [43]. These guidelines emphasize the importance of fairness and bias-free AI, advocating for transparency, traceability, and explainability in AI systems.

The AIA further builds on these principles by introducing a risk-based approach to AI regulation. It classifies AI systems into different risk categories—unacceptable, high, and low risks—and imposes stricter requirements on high-risk AI systems [44]. High-risk AI systems must comply with rigorous standards for data governance, transparency, and human oversight to ensure they do not produce biased outcomes. The AIA mandates that high-risk AI providers implement measures to detect and mitigate bias, document their processes, and allow for third-party audits.

3.2. United States

The regulatory landscape in the United States is less cohesive than in China or the EU, with AI regulation primarily driven by individual agencies and sector-specific guidelines. The Consumer Financial Protection Bureau (CFPB) has issued guidance prohibiting the use of unexplained AI models in credit decisions to prevent discriminatory practices [45]. Additionally, the National Telecommunications and Information Administration (NTIA) has

issued a Request for Comment on AI Accountability Policy, focusing on transparency and accountability in AI systems [46].

The U.S. approach to AI regulation often involves guidelines and best practices rather than comprehensive legislative frameworks. For example, the AI Now Institute and the Algorithmic Justice League have published reports and guidelines advocating for transparency, fairness, and accountability in AI systems [47,48]. These initiatives emphasize the need for regular audits, diverse development teams, and public disclosure of AI system information to mitigate bias..

3.3. *China*

China has made significant strides in regulating AI, particularly with the introduction of the "Interim Measures for the Administration of Generative AI Services" by the Cyberspace Administration of China (CAC) and other departments. These measures aim to promote the healthy development of generative AI, safeguard national security, and protect public interests [49]. The Interim Measures address bias by requiring AI service providers to adopt measures throughout the product lifecycle to prevent biases related to race, ethnicity, religion, nationality, gender, age, and profession. However, the Interim Measures lack detailed guidelines on implementing these requirements, particularly in terms of transparency and accountability [14].

Local governments in China, such as those in Beijing, Shanghai, and Shenzhen, have also issued regulations aimed at addressing AI bias. These regulations include prohibitions on biased behaviors and requirements for AI transparency and fairness [50]. However, similar to the national measures, these local regulations often lack specific guidelines and enforcement mechanisms.

3.4. *Comparative analysis*

Comparing the regulatory approaches of the EU, the U.S., and China reveals both commonalities and gaps in how bias in generative AI systems is addressed. While the EU's risk-based framework provides a robust starting point, its lack of specificity for generative AI leaves room for improvement. In contrast, the U.S.'s sector-specific approach lacks the coherence needed to tackle the cross-sectoral nature of generative AI bias. China's proactive regulatory stance, while commendable, still requires clearer guidelines and stronger enforcement mechanisms.

Targeted Recommendations:

Mandatory Bias Audits: All jurisdictions should require periodic bias audits for high-risk generative AI systems. These audits should include both pre-deployment and post-deployment evaluations to ensure that biases are not only identified early but also monitored over time as AI systems interact with new data.

Transparency and Explainability: Generative AI developers should be mandated to provide clear documentation of model decisions, including data sources, model architectures,

and the rationale behind specific algorithmic choices. This would allow for greater accountability and enable independent audits.

Diverse Development Teams: Regulatory frameworks should incentivize the formation of diverse development teams, as research shows that homogeneous teams are more likely to overlook biases that affect underrepresented groups. Policies that promote diversity in AI development can help mitigate systemic biases at the design stage.

Global Standards: International collaboration is critical for harmonizing AI bias regulations across borders. Organizations such as the OECD and the Global Partnership on AI should work towards creating global standards that address bias in generative AI systems. This would not only foster consistency but also prevent regulatory arbitrage, where companies move operations to jurisdictions with weaker regulations.

In conclusion, while current regulatory frameworks have made strides in addressing bias in AI systems, they fall short when it comes to generative AI's unique challenges. A more nuanced approach is needed—one that incorporates mandatory bias audits, transparency measures, and international collaboration to ensure that generative AI systems operate fairly and ethically across different regions and sectors.

4. Recommendations for mitigating bias in generative AI

This section provides actionable recommendations to mitigate bias in generative AI systems, supported by real-world examples and empirical research to enhance their practical applicability. It begins by proposing technical solutions at the data and model layers, such as using diverse datasets, implementing fairness-aware algorithms, and conducting ongoing bias monitoring. These technical measures are then linked to policy reforms, with concrete recommendations including mandatory bias audits, transparency requirements, and fairness testing, particularly in high-risk sectors like finance and healthcare. By integrating technical solutions with regulatory oversight, this approach offers a comprehensive strategy to address the unique challenges posed by bias in generative AI systems, ensuring fairer and more responsible AI deployment.

4.1. Data layer: ensuring representative and diverse data

Mitigating bias at the data layer is foundational to reducing bias in generative AI systems, as unrepresentative or inherently biased training data can skew outputs. To address these issues, a combination of technical strategies and regulatory frameworks is necessary.

First, clear and enforceable definitions of bias must be established. Regulatory bodies should mandate definitions that cover different types of bias, such as sampling, representation, and labeling bias, with explicit guidelines that protect demographic categories like race, gender, and socioeconomic status. These definitions should be embedded in legal frameworks to ensure compliance across sectors, particularly in areas like finance and healthcare, where bias can have significant impacts. Second, diverse and representative datasets are essential for fair AI systems. Developers should be required to source data that reflects the diversity of the population the system serves, and regulators should enforce

minimum diversity standards. One approach is to implement mandatory audits of datasets to assess their representativeness before AI systems are deployed. This would ensure that underrepresented groups are adequately included, reducing the risk of biased outputs. Third, rigorous data annotation standards should be enforced to minimize subjectivity and bias introduced by human annotators. Clear guidelines and comprehensive training programs should be established, coupled with regular audits of annotation practices to maintain consistency. Additionally, regulators could mandate the use of third-party oversight during the annotation process to ensure impartiality. Incorporating stakeholder input is crucial throughout the data collection and annotation stages. By consulting diverse stakeholders, particularly those from marginalized communities, potential biases can be identified early, leading to more inclusive AI systems. This can be facilitated by requiring organizations to include stakeholder consultations as part of their data collection protocols, ensuring diverse perspectives inform dataset development. Finally, advanced debiasing techniques, such as adversarial debiasing and counterfactual data augmentation, should be employed to further detect and mitigate biases in training data. Regulatory bodies could incentivize the use of these techniques by offering certifications for AI systems that meet fairness criteria. Such certifications could be made a requirement for high-risk AI applications, ensuring that systems are rigorously tested for bias before they are deployed.

In conclusion, addressing bias at the data layer requires a multi-pronged approach, combining technical solutions like diverse datasets and debiasing techniques with regulatory enforcement of data standards and stakeholder involvement. By embedding these practices into both technical and policy frameworks, generative AI systems can be made fairer and more inclusive.

4.2. Model layer: incorporating fairness-aware techniques

Addressing bias at the model layer is essential to ensure that generative AI systems produce fair and unbiased outputs. Bias at this stage can arise from decisions made during model development, the training data, and the optimization methods employed. To effectively mitigate these biases, several key strategies must be implemented, combining technical approaches with organizational practices.

First, integrating fairness-aware machine learning techniques during model development is crucial for reducing bias. This involves embedding fairness constraints and objectives directly into the training process. Techniques such as adversarial debiasing, reweighting, and fairness-aware optimization can adjust the learning process to ensure equitable treatment across different demographic groups. These methods help mitigate the risk of models perpetuating existing biases present in the training data and ensure that outputs reflect fairness from the ground up.

Second, transparency in model documentation is critical for fostering accountability and enabling external audits. Developers should comprehensively document all decisions made during model creation, including algorithm selection, parameter tuning, and any fairness constraints applied. This transparency allows for a clearer understanding of how and why

certain decisions were made, making it easier to trace the origins of bias and implement improvements. By maintaining detailed documentation, stakeholders can better evaluate the fairness of the model and ensure continuous learning and development in future iterations.

Third, regular audits and bias testing are necessary to monitor and address biases that may arise during both training and deployment. These audits should involve testing the model with diverse datasets and employing fairness metrics specifically designed to detect bias. Regular evaluations provide feedback throughout the model lifecycle, identifying areas for improvement and ensuring that the system remains fair as it interacts with new data. This continuous auditing process is vital for ensuring the long-term fairness of AI systems.

Fourth, diversity within AI development teams plays a critical role in reducing bias. A diverse team, composed of individuals from various backgrounds and perspectives, is more likely to detect and address potential biases that might be overlooked by homogeneous groups. Encouraging diversity in hiring and team formation should be a priority for organizations seeking to build fairer AI systems. Diverse teams are better equipped to understand the nuanced ways in which bias can manifest and to develop solutions that are inclusive of all user groups.

Lastly, advanced techniques like counterfactual fairness and explainability can further strengthen bias mitigation efforts. Counterfactual fairness involves testing whether a model's decision changes when sensitive attributes, such as race or gender, are altered, ensuring that these attributes are not influencing the outcome. Additionally, explainability techniques provide insights into the model's decision-making process, helping stakeholders identify potential biases and understand the rationale behind specific outputs. Explainable models are more transparent, enabling corrective actions when biased decisions are identified.

In summary, mitigating bias at the model layer requires a multi-faceted approach that includes fairness-aware machine learning techniques, transparent documentation, regular audits, diverse development teams, and advanced fairness methods like counterfactual fairness and explainability. By implementing these strategies, AI developers can create more equitable models that minimize harmful biases, leading to fairer and more responsible outcomes for all users.

4.3. Application layer: ensuring fairness in deployment

Bias at the application layer of generative AI systems stems from how these systems are deployed and used in real-world scenarios. To address this bias, it is essential to focus on user education, designing inclusive applications, continuous monitoring, and fostering transparency and accountability. These strategies help ensure that AI systems are applied fairly and ethically across different contexts.

Educating users is a critical first step in mitigating bias at the application layer. Users must understand that AI-generated outputs can carry biases and should not be treated as inherently objective or unbiased. Through targeted training and awareness programs, users can learn to critically evaluate AI outputs and integrate human judgment into decision-

making processes. Educated users are more likely to spot potential biases and make informed decisions, reducing the risk of overreliance on flawed AI systems.

Another important approach is designing inclusive applications that account for a diverse range of user needs. This involves incorporating diverse datasets and consulting with stakeholders from various demographic groups during the design and testing phases. Ensuring that AI systems are effective and fair for all user groups helps prevent the disproportionate impact of bias on any specific demographic. Inclusivity in design also helps identify potential biases early, allowing developers to address them before deployment.

Continuous monitoring is essential for identifying and addressing biases that may emerge once an AI system is deployed. AI systems should be regularly assessed to ensure they operate fairly as they encounter new data and user interactions. Monitoring frameworks should track performance across different user groups and identify any emerging patterns of bias. Regular audits and evaluations enable developers to correct issues promptly, ensuring that the AI system remains equitable over time.

Transparency and accountability are key elements in bias mitigation at the application layer. AI developers and operators must provide clear documentation of how systems are designed, the data used, and the potential biases identified during development. Transparent systems allow for external scrutiny, helping users and auditors understand how decisions are made and where biases might occur. Fostering a culture of accountability ensures that organizations are committed to addressing bias throughout the lifecycle of the AI system.

Finally, collaborating with independent auditors provides an additional layer of oversight. Independent evaluations offer unbiased assessments of AI systems, helping to identify and correct biases that may not be apparent to developers. Regular third-party audits reinforce trust in AI systems and ensure adherence to ethical standards, promoting their fair and responsible use.

In conclusion, mitigating bias at the application layer requires a comprehensive approach that includes user education, inclusive design, continuous monitoring, transparency, and collaboration with independent auditors. By implementing these strategies, organizations can ensure that AI systems are deployed fairly, ethically, and with a focus on minimizing bias across all user groups.

4.4. Regulatory and policy measures: creating a robust governance framework

Regulatory and policy measures are critical for mitigating bias in generative AI systems. A strong governance framework requires a mix of risk-based approaches, transparency standards, international collaboration, and stringent oversight mechanisms to ensure fairness, accountability, and equity throughout the AI lifecycle.

Adopting a risk-based regulatory approach is fundamental for effective governance. AI systems should be classified based on their potential risk, with high-risk applications subject to stricter regulations. This approach ensures that AI systems with significant societal impact, such as those used in healthcare, criminal justice, or hiring, undergo more rigorous scrutiny

to prevent biased outcomes. Implementing such a framework ensures that high-risk systems are thoroughly assessed, reducing the likelihood of biased or discriminatory results.

Transparency and accountability are essential to foster trust and ensure fair practices in AI development. Regulators should mandate clear and comprehensive documentation of AI processes, including data sources, model architecture, and decision-making criteria. Public disclosure of these details enables independent audits and allows stakeholders to understand the sources of bias. Enhanced transparency makes it easier for both regulators and the public to hold developers accountable for addressing potential biases in their systems.

International collaboration is another key component of a robust governance framework. By working together, countries can develop standardized guidelines and best practices for mitigating AI bias, harmonizing regulations across borders. International organizations and initiatives can help ensure that AI ethics and governance are consistent globally, allowing for shared learning and mutual reinforcement of ethical standards. This cross-border cooperation helps create universal norms that prevent bias and promote fairness in AI systems on a global scale.

Robust regulatory oversight mechanisms are necessary to enforce compliance with bias mitigation measures. Dedicated regulatory bodies should be empowered to conduct audits, impose penalties, and mandate corrective actions when biases are identified. These oversight entities should regularly evaluate AI systems, ensuring that they continue to operate fairly and ethically as they evolve. Having the authority to impose sanctions ensures that companies and developers are held to high ethical standards, while the technical expertise within these bodies ensures that AI systems are adequately scrutinized.

Engaging with diverse stakeholders is crucial to ensure that regulatory measures are comprehensive and effective. Policymakers should actively seek input from academia, industry, civil society, and communities affected by AI systems. This inclusive approach helps to identify potential biases and ensures that the regulatory framework addresses the concerns of all stakeholders, especially those from marginalized or underrepresented groups. Engaging a broad spectrum of voices also ensures that policies are well-rounded and capable of addressing real-world impacts.

In summary, regulatory and policy measures must be built on a foundation of risk-based approaches, transparency, international collaboration, and strong oversight. These elements, combined with broad stakeholder engagement, are essential for creating a governance framework that mitigates bias in generative AI systems and promotes fairness, accountability, and equity in AI deployment.

5. Conclusion

The rapid proliferation of generative AI systems presents remarkable opportunities and significant challenges, particularly concerning bias and fairness. Bias in AI systems can perpetuate and exacerbate existing societal inequalities, making it imperative to address this issue comprehensively. Mitigating bias in generative AI requires a multi-layered approach targeting the data, model, and application layers, supported by robust regulatory and policy measures.

At the data layer, ensuring diverse and representative datasets, implementing rigorous data annotation standards, and engaging in continuous stakeholder consultation are essential steps. These measures help to minimize bias introduction from the outset, creating a solid foundation for fair and equitable AI systems. In the model layer, integrating fairness-aware machine learning techniques, maintaining transparency in model documentation, conducting regular audits, and fostering diversity within development teams are crucial. These strategies help identify and mitigate biases that may emerge during the development and training of AI models, ensuring that the systems are robust and fair. The application layer requires educating users, designing inclusive applications, continuous monitoring, and fostering transparency and accountability. These steps ensure that AI systems are used ethically and fairly, reducing the risk of biased decision-making and promoting equitable outcomes across different user groups.

Robust regulatory and policy measures are vital to support these technical and procedural strategies. Risk-based regulation, transparency and accountability frameworks, international collaboration, effective regulatory oversight, and inclusive stakeholder engagement provide a comprehensive foundation for mitigating bias in generative AI. These measures ensure that AI systems are developed and deployed in a manner that upholds fairness and equity, building public trust and confidence in these technologies. By adopting this multi-faceted approach, stakeholders—including developers, users, regulators, and policymakers—can work together to create generative AI systems that are not only innovative and powerful but also fair and just. Addressing bias in AI is not just a technical challenge but a societal imperative, requiring ongoing commitment and collaboration to ensure that the benefits of AI are shared equitably across all segments of society.

The journey towards mitigating bias in generative AI is complex and ongoing. However, by focusing on comprehensive strategies across all layers of AI development and deployment, supported by robust regulatory frameworks, we can make significant strides towards creating AI systems that are fair, transparent, and accountable. These efforts will help harness the full potential of generative AI while safeguarding against the risks of bias, ensuring that these technologies contribute positively to society as a whole.

Conflicts of interests

The authors declare no conflicts of interests.

Authors' contribution

Conceptualization, Shuchen Tang; methodology, Shuchen Tang; writing—original draft preparation, Shuchen Tang; data collection, Haoming Zhu; writing—review and editing, Haoming Zhu. All authors have read and agreed to the published version of the manuscript.

References

- [1] Ferrer X, Van Nuenen T, Such JM, Coté M, Criado N. Bias and discrimination in AI: a cross-disciplinary perspective. *IEEE Technol. Soc. Mag.* 2021, 40(2):72–80.

- [2] Liu Y. An analysis of algorithmic bias and its regulatory paths (In Chinese). *Law Sci. Mag.* 2019, 40(6):55–66.
- [3] Fang X, Che S, Mao M, Zhang H, Zhao M, *et al.* Bias of AI-generated content: An examination of news produced by large language models. *Sci. Rep.* 2024, 14(1):5224.
- [4] Bird C, Ungless E, Kasirzadeh A. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAI/ACM Conference on AI, Ethics, and Society*, Montreal, Canada, 8–10 August 2023, pp. 396–410.
- [5] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, *et al.* On the opportunities and risks of foundation models. *arXiv* 2021, arXiv:2108.07258.
- [6] Migliorini S. China's Interim Measures on generative AI: Origin, content and significance. *Comput. Law Secur. Rev.* 2024, 53:105985.
- [7] European Commission. Proposal for a regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act). European Commission; 2021. Available: <https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence#:~:text=The%20Proposal%20for%20a%20Regulation%20on%20artificial%20intelligence%20was%20announced> (accessed on 16 July 2024).
- [8] Al-kfairy M, Mustafa D, Kshetri N, Insiew M, Alfandi O. Ethical Challenges and Solutions of Generative AI: An Interdisciplinary Perspective. *Informatics* 2024, 11(3):58.
- [9] Laine J, Minkkinen M, Mäntymäki M. Ethics-based AI auditing: A systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders. *Inf. Manag.* 2024:103969.
- [10] Zhao J, Wang T, Yatskar M, Ordonez V, Chang KW. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv* 2018, arXiv:1804.06876.
- [11] Mustafa D, Al-Kfairy M. Ethical considerations in electronic data in healthcare. *Front. Public Health* 2024, 12:1454323.
- [12] Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, New York, United States, 23–24 February 2018, pp. 77–91.
- [13] Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, *et al.* Datasheets for datasets. *Commun. ACM.* 2021, 64(12):86–92.
- [14] Shuchen T, Huiwen J. Bias in Generative AI Systems: A 3-Layer Response and Liability Determination. *Contemp. Soc. Sci.* 2024, 9(2):121–38.
- [15] Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 3–10 March 2021, pp. 610–623.
- [16] Broussard M. *Artificial Unintelligence: How Computers Misunderstand the World*. London: MIT Press, 2018.
- [17] Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Adv. Neural Inf. Process. Syst.* 2016, 29:4349–4357.
- [18] Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science* 2017, 356(6334):183–186.

- [19] Geva M, Goldberg Y, Berant J. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv* 2021, arXiv:1908.07898.
- [20] Tubadji A, Huang H, Webber D J. Cultural proximity bias in AI-acceptability: The importance of being human. *Technol. Forecast. Soc. Change* 2021, 173:121100.
- [21] Raji ID, Gebru T, Mitchell M, Buolamwini J, Lee J, *et al.* Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, United States, 7 February 2020, pp. 145–151.
- [22] Choudhry M D, Sundarajan M, Sundaram K. *9 Bias and Fairness in Generative AI*. Generative AI and LLMs: Natural Language Processing and Generative Adversarial Networks, 2024, p. 177.
- [23] DeCamp M, Lindvall C. Mitigating bias in AI at the point of care. *Science*. 2023, 381(6654):150–152.
- [24] Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, *et al.* Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, Atlanta, United States, 29–31 January 2019, pp. 220–229.
- [25] Binns R. Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the Conference on fairness, accountability and transparency*, New York, United States, 21 January 2018, pp. 149–159.
- [26] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* 2021, 54(6):1–35.
- [27] Holstein K, Wortman Vaughan J, Daumé III H, Dudik M, Wallach H. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, Glasgow, United Kingdom, 4–9 May 2019, pp. 1–16.
- [28] Hajian S, Bonchi F, Castillo C. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, San Francisco, United States, 13–17 August 2016, pp. 2125–2126.
- [29] Herries G. Navigating the AI Frontier: Bias, Ethics, and the Vital Collaboration Between Engineers and Policymakers. *Chem. Eng.* 2023, (988):32–33.
- [30] Barocas S, Hardt M, Narayanan A. Fairness and Machine Learning. 2019. Available: <https://fairmlbook.org/> (accessed on 16 July 2024).
- [31] Zhao J, Wang T, Yatskar M, Ordonez V, Chang KW. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv*, 2017, arXiv:1707.09457.
- [32] Suresh H, Gutttag JV. A framework for understanding unintended consequences of machine learning. *Commun. ACM*. 2019, 63(5):62–71.
- [33] Radford A, Wu J, Child R, Luan D, Amodei D, *et al.* Language models are unsupervised multitask learners. 2019. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (accessed on 16 July 2024).

- [34] Gichoya JW, Thomas K, Celi LA, Safdar N, Banerjee I, *et al.* AI pitfalls and what not to do: mitigating bias in AI. *Br. J. Radiol.* 2023, 96(1150):20230023.
- [35] Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, *et al.* A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, Atlanta, United States, 29–31 January 2019, pp. 329–338.
- [36] Zafar MB, Valera I, Rogniguez MG, Gummadi KP. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, United States, 20–22 April 2017, pp. 962–970.
- [37] Wen Y, Holweg M. A phenomenological perspective on AI ethical failures: The case of facial recognition technology. *AI Soc.* 2023:1–18.
- [38] Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing.* 2020, 415:295–316.
- [39] O'Neil C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York: Crown Publishing Group, 2016.
- [40] West SM, Whittaker M, Crawford K. Discriminating systems: Gender, race, and power in AI. 2019, pp. 1–33. Available: <https://ainowinstitute.org/wp-content/uploads/2023/04/discriminatingystems.pdf> (accessed on 16 July 2024).
- [41] Angwin J, Larson J, Mattu S, Kirchner L. Machine bias. In *Ethics of data and analytics.* Florida: Auerbach Publications, 2022, pp. 254–264.
- [42] Mittelstadt B, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: Mapping the debate. *Big Data Soc.* 2016, 3(2):2053951716679679.
- [43] Tallberg J, Lundgren M, Geith J. AI regulation in the European Union: examining non-state actor preferences. *Bus. Polit.* 2024, 26(2):218–39.
- [44] Silva N S E. The Artificial Intelligence Act: critical overview. 2024. Available: <https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence#:~:text=The%20Proposal%20for%20a%20Regulation%20on%20artificial%20intelligence%20was%20announced> (accessed on 20 July 2024).
- [45] Consumer Financial Protection Bureau (CFPB). Consumer protection guidance on artificial intelligence in credit decisions. 2023. Available: <https://www.consumerfinance.gov/about-us/newsroom/cfpb-issues-guidance-on-credit-denials-by-lenders-using-artificial-intelligence/#:~:text=%E2%80%93Today%2C%20the%20Consumer%20Financial%20Protection%20Bureau%20%28CFPB%29,accurate%20reasons%20when%20taking%20adverse%20actions%20against%20consumers> (accessed on 20 July 2024).
- [46] National Telecommunications and Information Administration (NTIA). Request for comment on AI accountability policy. 2024. Available: <https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report> (accessed on 20 July 2024).
- [47] AI Now Institute. 2024. Available: <https://ainowinstitute.org/> (accessed on 20 July 2024).
- [48] Algorithmic Justice League. 2024. Available: <https://www.ajl.org> (accessed on 20 July 2024).
- [49] Cyberspace Administration of China (CAC). Interim measures for the administration of generative AI services. CAC; 2023. Available: https://pkulaw.com/en_law/6dc227b9153496c2bdfb.html#:~:text=The%20Interi (accessed on 20 July 2024).

[50] Todaro D. *The Use of Artificial Intelligence in the Public Sector in Shanghai*. Singapore: Palgrave Macmillan Singapore. 2024.