AI and Ethics
https://doi.org/10.1007/s43681-025-00721-9

REVIEW

Systematic literature review on bias mitigation in generative AI

Juveria Afreen[1] · **Mahsa Mohaghegh**[2] · **Maryam Doborjeh**[3]

Received: 29 August 2024 / Accepted: 19 March 2025

Abstract

In the era of rapid technological advancement, Artificial Intelligence (AI) is a transformative force, permeating diverse facets of society. However, bias concerns have gained prominence as AI systems become integral to decision-making processes. Bias can exert significant and extensive consequences, influencing individuals, groups, and society. The presence of bias in generative AI or machine learning systems can produce content that exhibits discriminating tendencies, perpetuates stereotypes, and contributes to inequalities. Artificial intelligence (AI) systems have the potential to be employed in various contexts that involve sensitive settings, where they are tasked with making significant judgements that can have profound impacts on individuals' lives. Consequently, it is important to establish measures that prevent these decisions from exhibiting discriminating tendencies against specific groups or populations. This exclusive exploration embarks on a comprehensive journey through the nuanced landscape of bias in AI, unravelling its intricate layers to discern different types, pinpoint underlying causes, and illuminate innovative mitigation strategies. Delving deeper, we investigate the roots of bias in AI, revealing a complex interplay of historical legacies, societal imbalances, and algorithmic intricacies. Unravelling the causes involves exploring unintentional reinforcement of existing biases, reliance on incomplete or biased training data, and the potential amplification of disparities when AI systems are deployed in diverse real-world scenarios. Various domains such as text, image, audio, video and more significant advancements in Generative Artificial Intelligence (GAI) were evidenced. Multiple challenges and proliferation of biases occur in different perspectives considered in the study. Against this backdrop, the exploration transitions to a proactive stance, offering a glimpse into cutting-edge mitigation strategies. Diverse and inclusive datasets emerge as a cornerstone, ensuring representative input for AI models. Ethical considerations throughout the development lifecycle and ongoing monitoring mechanisms prove pivotal in mitigating biases that may arise during training or deployment. Technical and non-technical strategies come to the forefront of pursuing fairness and equity in AI. The paper underscores the importance of interdisciplinary collaboration, emphasising that a collective effort spanning developers, ethicists, policymakers, and end-users is paramount for effective bias mitigation. As AI continues its ascent into various spheres of our lives, understanding, acknowledging, and addressing bias becomes an imperative. This exploration seeks to contribute to the discourse, fostering a deeper comprehension of the challenges posed by bias in AI and inspiring a collective commitment to building equitable, trustworthy AI systems for the future.

Keywords Generative AI · Bias mitigation · Text · Image · Audio · Video · Societal impact · Hallucinations · Misuse systematic literature review

✉ Juveria Afreen
juveria.afreen@autuni.ac.nz

Mahsa Mohaghegh
mahsa.mohaghegh@aut.ac.nz

Maryam Doborjeh
mgholami@aut.ac.nz

[1] Department of Computer Science and Software Engineering, Auckland University of Technology, Auckland, New Zealand

[2] Department of Data Science and Artificial Intelligence, Auckland University of Technology, Auckland, New Zealand

[3] Computer and Information Sciences Department, Knowledge Engineering and Discovery Research Innovation, ECMS, Auckland University of Technology, Auckland, New Zealand

Published online: 25 August 2025

1 Introduction

Artificial general Intelligence was used to create high-quality contextual data as created by humans, as stated by [23, 149]. On the other hand, in the digital world, Generative Artificial Intelligence (GAI) was used to create proficient, knowledgeable and skilled data specialised in all fields [43, 147, 159]. In other words, GAI, defined as an exploration of AI, was improved by creating realistic, creative and unique content based on the latest advancements using Deep Learning (DL) and Machine Learning (ML) technologies depending on various classifications and predictions. Process optimisation and decision-making enhancement entirely depend on adopting AI (Artificial Intelligence) to help collaborate intelligent and human systems [22, 25]. More exploration in the growth of AI leads to the development of GAI, providing innovative services in all domains such as natural language generation, image generation, video generation, healthcare, engineering and design and marketing by implementing automation and novel augmentation [67, 106].

More innovative and capable models were researched in assisting to complete required tasks in a faster and easier way by using GAI [102]. In academic discourse, enhancing GAI helps to positively impact fundamental principle applications and various innovations in different domains [143, 159]. The successful implementation of GAI was made possible by providing solutions to various problems caused by misuse, bias and transparency [65, 126, 140]. Various concepts, applications, and challenges caused by adopting Generative AI were addressed in this research to enhance its ability to produce bias-free content. The method of creating new content based on existing data and creating a variety of new applications was advanced with the help of Deep Learning (DL) integrated with AI techniques [150]. The integration of DL with AI failed to produce realistic data, but this was addressed with the help of GAI. Using a finite dataset, the deep generation model produced a high dimensional probability distribution [123]. The inherent organisation of data and data processing focusing on discriminative models was able to predicate the ship between output label and input feature [72].

1.1 Background

Training the GAI model helps enable various use cases and generate new data. Based on proximal policy optimisation, the GAI model uses a supervised fine-tuning model that differs from the semi-supervised learning used in AI [112]. GAI can easily process vast sets of data using a unique approach. GAI integrated with natural language processing (NLP) helps to produce desired and realistic output on text, images and other types of data [39, 91].

GAI combines artistic creativity, social intelligence, and cognitive intelligence [77]. Users can view automatically served digital ads by opening websites [11]. Programmatic creative and programmatic buying play a significant role in programmatic advertisements created using GAI [33]. In the recruitment process, GAI was used to make decisions and was considered a substitute for human decisions [18]. The manual intervention of work was replaced with GAI with a reduction of cost and numerous benefits [21].

The author [155] has proposed Fairness-Aware Adversarial Perturbation (FAAP) approach to mitigate bias in already-deployed models. FAAP focuses on scenarios where the deployed model's parameters are inaccessible. Instead of modifying the model, it perturbs inputs to render fairness-related attributes undetectable. A discriminator is used to identify these attributes within the model's latent representations, while a generator acts adversarially to 'fool' the discriminator.

Data Oversampling has been studied [48], where underrepresented groups have been addressed by generating synthetic data. The study has examined bias mitigation through targeted oversampling to address the underrepresentation of specific groups within a dataset. By expanding training data in areas where positive base rate differences exist (leading to bias), fairness and overall accuracy can be improved. This is different from other de-biasing methods that may remove information deemed sensitive.

Other researchers, such as [41] and [130], have evaluated fairness without considering demographics. They have focused on techniques that work even without explicit sensitive attributes. Both tackle the challenge of ensuring fairness in scenarios where sensitive attributes are either unknown or unavailable due to privacy concerns [41] and have focused on human-centred federated learning, which has introduced a method that minimises the top eigenvalue of the Hessian matrix to reshape the loss landscape, thus indirectly addressing bias. [130] has presented a Distributionally Robust Optimization (DRO) approach for recommender systems, which aims to minimise the worst-case unfairness over reconstructed probability distributions of missing sensitive attributes, accounting for potential reconstruction errors [15]. They have tried to understand Implicit Bias, probing gender bias in models in scenarios without explicitly gendered language. The UnStereoEval (USE) framework has been introduced to generate benchmarks without prominent gender-related associations. It is seen that a wide range of tested language models still display significant bias on these stereotype-free datasets, highlighting that bias does not solely stem from stereotypical word correlations [134] have also predicted fairness. They have

explored whether the hyperparameter configuration of the model can be used to predict fair outcomes. It has offered a novel angle by exploring the relationship between machine learning hyperparameters and fair outcomes. The goal is to predict the fairness of an ML configuration given a particular dataset. It is seen that tree regressors outperform other methods, particularly under temporal distribution shifts.

1.2 Significance

The imperative to eliminate bias from generative AI models arises from various ethical, social, and technological factors. The potential for biased AI outputs to perpetuate stereotypes, inequality, and discrimination is a significant concern, as it can amplify existing societal biases and inflict harm upon individuals and marginalised communities. We aim to cultivate artificial intelligence (AI) systems that exhibit equity and fairness by deliberately and proactively mitigating bias. This entails generating content that faithfully reflects various perspectives and experiences. Enhancing the quality and reliability of AI-generated outputs is crucial in fostering user trust and confidence in these technologies. In addition, it is worth noting that with the growing integration of AI systems across diverse domains, there is a heightened focus within legal and regulatory frameworks on the imperative of ensuring the impartiality and absence of discriminatory biases in AI technologies. Eliminating bias in AI research and development is crucial for upholding ethical responsibilities, fostering social progress, and facilitating innovation. The underlying drive to mitigate bias in generative AI reflects our unwavering dedication to constructing an AI ecosystem that is characterised by inclusivity, fairness, and accountability, thereby yielding advantages for all individuals involved.

Generative AI models, despite their potency, can inherit and exacerbate biases present in their training data, resulting in substantial real-world ramifications. These biases show in diverse applications, including facial recognition systems that inaccurately identify persons from minority groups [87, 115] and text generation models that reinforce gender or racial stereotypes. AI-generated information in media may exhibit biased perspectives, shaping public opinion [51 ;53]. Moreover, deepfake technology has been exploited malevolently, frequently aiming at demographics such as women or marginalised groups, underscoring the ethical dilemmas inherent in generative AI systems—some significant examples of bias in generative models in real-world scenarios.

Image generation—Academic research [148] found bias in the generative AI art generation application Midjourney. When asked to create images of people in specialised professions, it showed both younger and older people, but the older people were always men, reinforcing gendered bias of the role of women in the workplace.

Predictive policing tools—AI-powered predictive policing [63] tools used by some organisations in the criminal justice system are supposed to identify areas where crime is likely to occur. However, they often rely on historical arrest data, which can reinforce existing patterns of racial profiling and disproportionate targeting of minority communities.

ChatGPT—ChatGPT is tested for identifying six types of bias, including racial bias, gender bias, cognitive bias, text-level context bias, hate speech, and fake news, giving an overview of ChatGPT's capability of dealing with different biases.

Stable Diffusion—Generative AI tools present similar problems. For example, a 2023 analysis of more than 5000 images created with the generative AI tool Stable Diffusion to generate images related to titles and crime found that it simultaneously amplifies gender and racial stereotypes [109]. The text-to-image model is used to create representations of workers for 14 jobs—300 images each for seven jobs typically considered "high-paying" in the US and seven considered "low-paying" — plus three crime- related categories.

This research holds significant value as it aims to address the pressing issue of bias in large language models (LLMs) within the financial sector, specifically in loan assessments. LLMs have become critical tools in decision-making processes across various industries, including healthcare, transportation, and finance. However, these models are susceptible to inherent biases, such as demographic misrepresentation and stereotype reinforcement, that can influence outcomes in unfair and discriminatory ways. In decision-making, such biases can disproportionately affect marginalised groups, leading to unequal access to resources like loans and perpetuating economic disparities.

2 Research methodology

A systematic review provided a better understanding and consolidation of the research. In the field of research, systematic review was widely used to collect information. For the analysis and classification of domains, a systematic mapping study was considered an efficient method for examining primary studies. Further, the need for systematic literature was predicated on the help of systematic mapping. Mitigation of Biases in Generative AI based on text, images, audio, video, and others and challenges faced during enhancement of Generative AI based on bias, transparency, misuse, hallucinations and societal impact is considered an output. Planning, scoping, searching, assessing

and synthesising are different steps in processing research data based on systematic reviews and meta-analyses [105].

3 PRISMA model

To comprehensively investigate various types of bias, their impacts, and mitigation strategies in the realm of artificial intelligence, a systematic literature review (SLR) was conducted. This approach involves a meticulous and structured analysis of existing scholarly works, research articles, and publications related to bias in AI. SLR is considered a robust, explicit, and reproducible method that facilitates theory development by uncovering new research areas while closing up areas where abundant research exists [156]. For this study, Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) compliant systematic literature reviews are conducted to analyse the fairness, consequences, and challenges in generative AI. PRISMA methodology is adopted to reduce bias and add rigour and research transferability to the review [46]. The survey will begin by systematically identifying relevant literature through established databases such as Scopus, Web of Science, Google Scholar, IEEE, and ARXIV, ensuring comprehensive field coverage. Each selected study will undergo a rigorous review process, focusing on the types of bias examined, their observed impacts, and the proposed or implemented mitigation strategies. By synthesising insights from diverse sources, this research methodology aims to provide a nuanced and up-to-date understanding of the complex landscape of biases in AI, offering valuable insights for future developments and fostering a more ethical and unbiased artificial intelligence landscape.

The systematic literature review was conducted in four different stages. The initial identification of the research starts with searching based on the keyword image, video, audio, code, text and others based on mitigation of bias in generative AI, as shown in Fig. 1 below. In addition to the

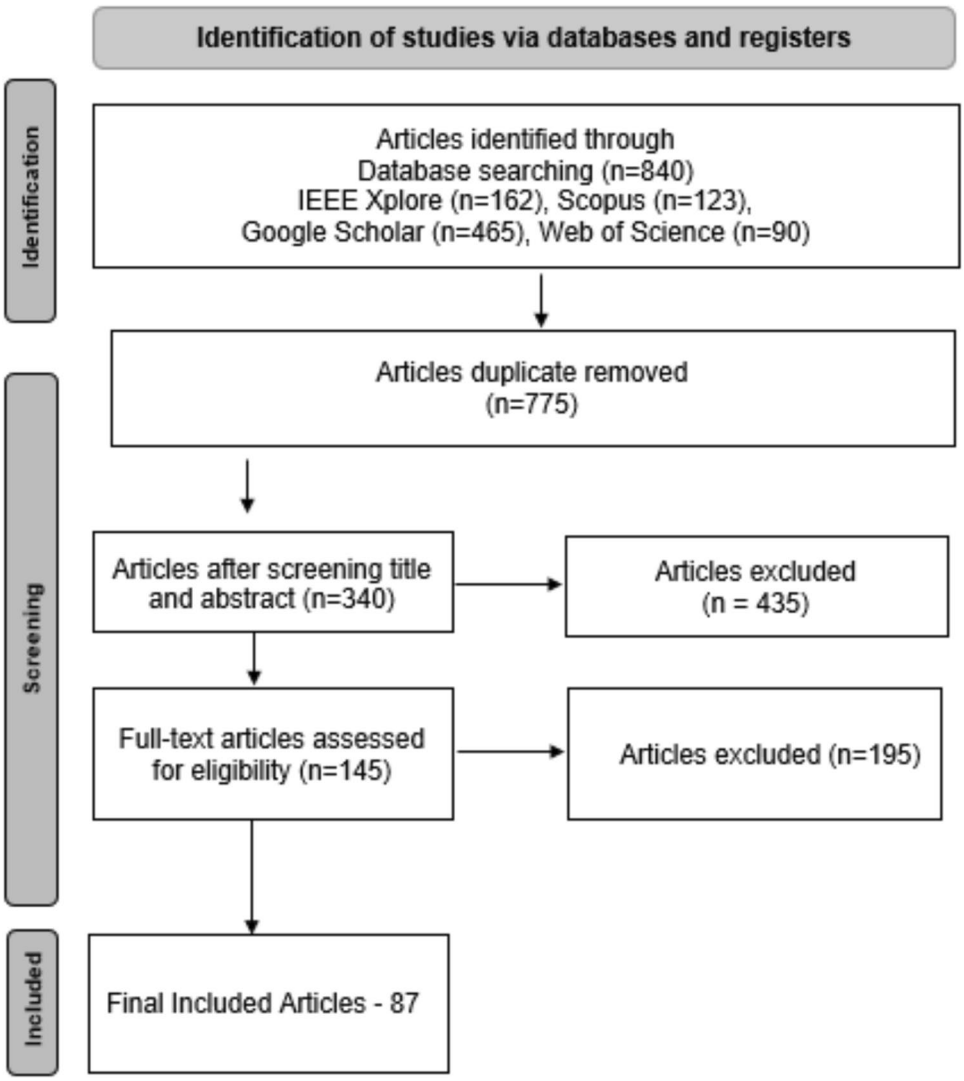

Fig. 1 Prisma model

studies on bias, various challenges caused by generative AI are considered using keyword bias, transparency, hallucinations, misuse, and societal impact, which are included based on optimal search. The usefulness of the PRISMA model was improved with the developed explanatory document.

Search Terms—We carried out a search for scientific documentation in Google Scholar, IEEE, Science Direct, Scopus and Library databases using keywords associated with Bias in Generative AI (Artificial Intelligence OR Generative AI OR Machine Learning OR various biases in AI OR Generative AI models OR Survey on ChatGPT OR Dalle OR Different types of bias in Generative OR what are sources of bias in Generative AI OR different methodologies to mitigate bias in AI OR what is bias in AI OR Bias in Machine Learning OR what is fairness OR how can we quantify fairness OR categorise of different bias OR Gender Inequality in Machine Learning survey in fairness in Machine Learning OR what are different methods incorporate to mitigate the bias OR what are the effects of bias in AI OR what are the Machine Biases OR what are the causes of bias OR what are the factors for bias in AI OR algorithms used in bias mitigation OR how to remove bias from AI OR fairness and bias in AI OR reducing bias in AI/ML OR bias classification OR Survey on Gender Bias OR Bias in generative AI in images OR real examples of bias in AI. The search for documents was limited to publications that appeared in scientific journals from 2018 to 2023. The same descriptors were used to search the internet for relevant grey literature using the Google search engine. We similarly undertook a manual search using the bibliographic references of the selected publications.

All the research studies scrutinising the affiliation to varied strategies and methods for mitigating AI bias were desirable and qualified for a systematic review. However, because of the literature available on the search objective, specific inclusion and exclusion criteria were applied to narrow the pre-existing literature selection.

3.1 Inclusion criteria

The inclusion criteria for this review included research published in peer-reviewed, open- access journals and written in English between 2018 and 2024. In addition, grey literature was excluded, and works were published in languages other than English. Both qualitative and quantitative research were included, as well as research with descriptive and experimental approaches. The study mainly includes bias in AI/ML.

Considering the search terms and inclusion and exclusion criteria, the first study was published in 2014 with only one article. Most papers were published between the last 3 years (2020–2024), demonstrating a notable growth of interest in the subject by researchers in the area. Indeed, Shrestha and Das [136] have stressed how algorithmic fairness has been a topic of interest in the academy for the past decade,the increase in the number of studies that seek to understand the effects of bias is a growing concern in understanding this phenomenon. They argue that the discourse on fairness in Machine Learning (ML) and Artificial Intelligence (AI) is a relatively recent phenomenon. However, it is essential to recognise that prejudice has long been ingrained in human society.

3.2 Exclusion criteria

The exclusion criteria for the PRISMA model in a systematic literature review include several vital points to ensure the inclusion of relevant and high-quality studies. Non-English publications are excluded unless translations are available, maintaining consistency in data interpretation. Studies that do not align with specific research questions or the relevant domain are excluded as irrelevant topics. Research lacking full-text availability or accessibility is also omitted from consideration. Low-quality studies that do not meet predefined quality benchmarks, such as those lacking robust methodology or peer review, are excluded to maintain high standards. Duplicate publications are removed to prevent redundancy in the analysis. Outdated research, which falls outside the set time frame and does not reflect recent data, is not included. Non-empirical studies such as opinion pieces, editorials, or commentaries without primary data or original research are excluded. Also, studies presenting incomplete or ambiguous data hindering comprehensive analysis are omitted. These criteria collectively ensure that the review focuses on relevant, reliable, high- quality research.

Research papers of nearly 500 numbers are collected from different Scopus, Science Directive and Springer Link sources to mitigate bias in Generative AI based on Text, Audio, image, video, code and others, as depicted in Fig. 2. An investigation was carried out using topic modelling techniques in compressive analysis. Using an optimal search of sources, nearly 340 papers are included based on various sources of information collected during the search process for challenges based on bias, transparency, hallucinations, misuse, and societal impact (Fig. 3).

In recent years, examining the relationship between average citations per article and total citations across countries has provided valuable insights into global research impact and publication influence. Notably, the United Kingdom has demonstrated a stable and consistent impact, averaging 53.70 citations per article, bolstered by a substantial count of 1020 total citations. This indicates a balanced blend of productivity and citation influence, positioning the UK as a prominent contributor in high-impact research (Fig. 3).

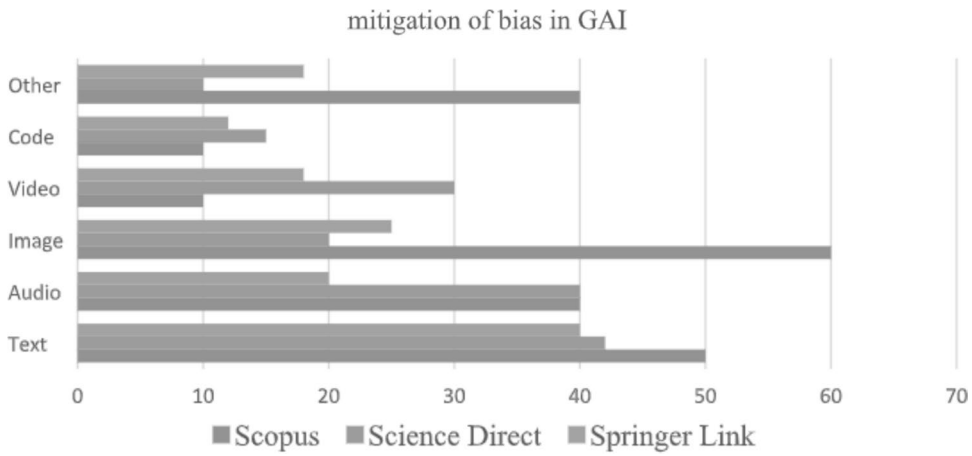

Fig. 2 Distribution of GAI research articles based on mitigation of bias

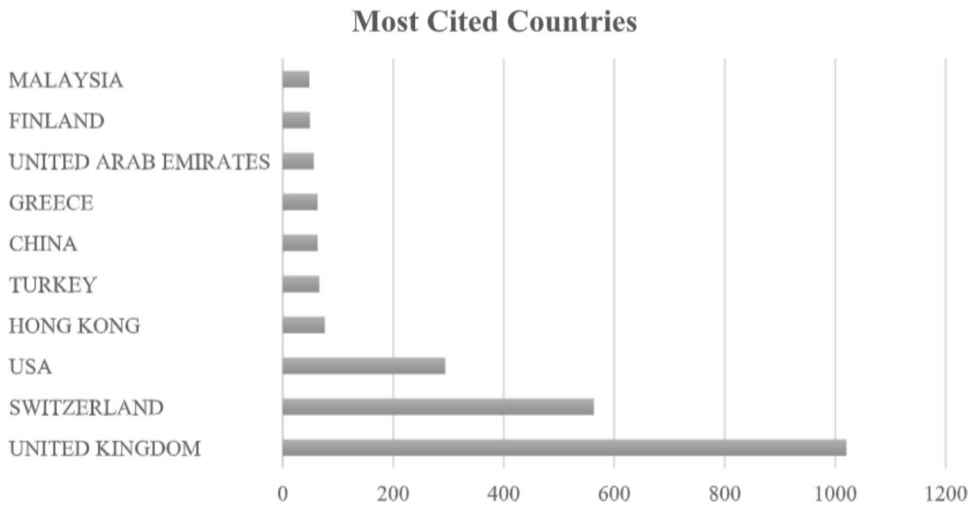

Fig. 3 Top cited countries in research impact

Switzerland, with a distinctly higher citation average of 140.80 per article and 563 citations, reflects a niche yet powerful research presence. This high citation-per-article ratio suggests that Switzerland's publications have significant influence and are recognised for their impact within scholarly communities.

In contrast, despite producing a large volume of publications, the United States exhibits a lower average citation count of 3.20 per article, with 294 citations. This disparity hints at a broad but less concentrated impact, potentially indicative of high research output that includes a range of both high- and low-impact publications.

Other countries, including Malaysia, Finland, the United Arab Emirates, China, Turkey, Greece, and Hong Kong, contribute to the diversity in global research impact. Turkey and Malaysia stand out with higher citation averages, reflecting notable recognition for each published article. Conversely, while contributing significantly to total publication volume, the USA and China display lower average citations per article, indicating a more dispersed influence within the academic community.

These findings underscore the variations in research impact across different countries, emphasising the importance of citation volume and per-article influence in understanding global research landscapes.

Figure 4 presents a thematic analysis of Generative Artificial Intelligence (GAI) research, revealing several core trends and focal areas. Through a structured keyword analysis, it was observed that "Generative Artificial Intelligence" appeared 307 times, while "Artificial Intelligence" occurred 288 times. Key concepts further emerged around terms such as "intelligence model," "learning system," and "generative AI," which capture the essential focus areas within GAI research.

The ethical implications of GAI development were highlighted by frequent references to "ethical technology," underscoring ongoing concerns about responsible AI development. Foundational and practical aspects of GAI were reflected in keywords like "generative adversarial network," "deep learning," and "machine learning," signifying a strong focus on both theoretical and real-time implementations.

The educational potential of GAI was also a prominent theme, with terms like "teaching," "educational computing,"

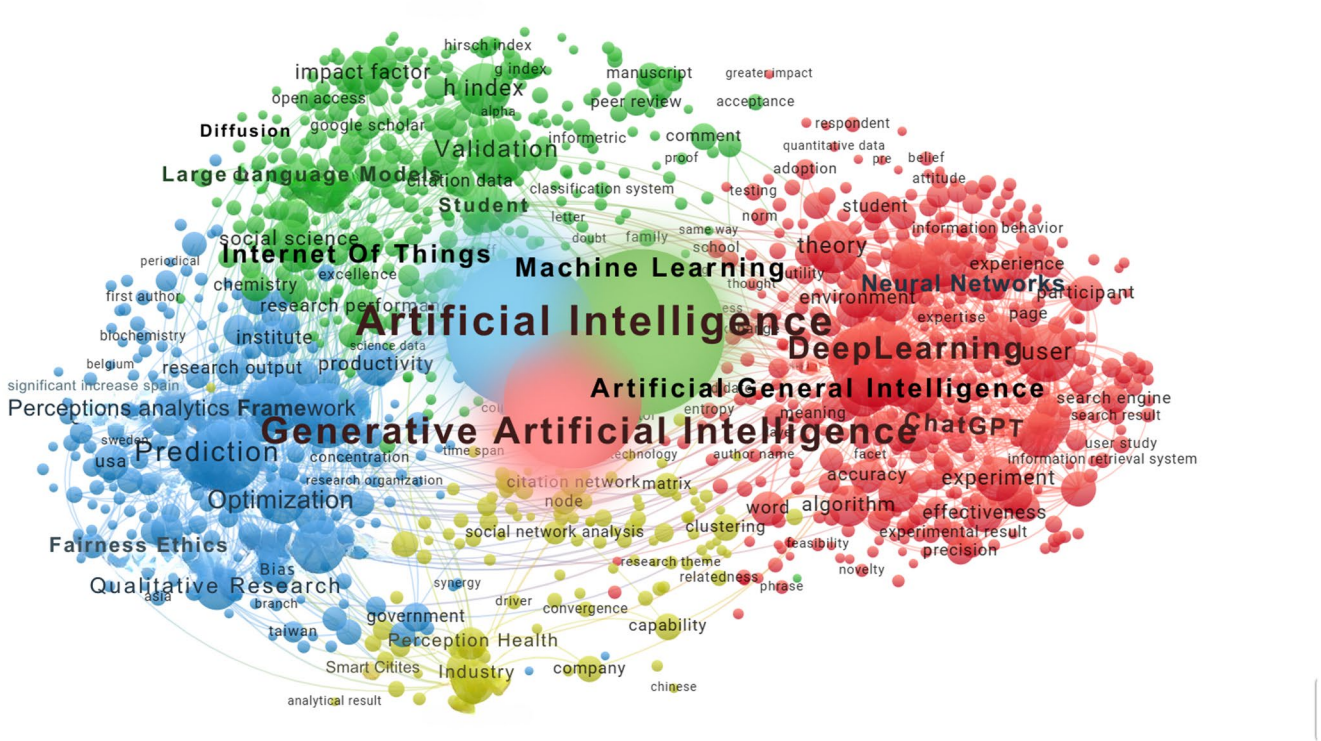

Fig. 4 Thematic analysis of key areas in generative AI (GAI)

and "students" emphasising GAI's impact on learning and pedagogy. Additionally, the keywords "ChatGPT" and "language model" (each cited 108 times) underscored the relevance of natural language processing and conversational AI in this field.

In the context of creativity and user interaction, themes such as "user interfaces," "creative," and "human–computer interaction" illustrated GAI's role in enhancing creative processes and user engagement. The diverse scope of GAI research encompasses multiple perspectives, including application-oriented, educational, ethical, and technical dimensions, collectively highlighting GAI's extensive and varied contributions to various sectors.

This systematic literature review (SLR) confirms the multifaceted nature of GAI research, driven by a blend of theoretical exploration, ethical considerations, practical applications, and educational impacts. These findings underline GAI's significant potential across various domains and the complexities associated with its responsible implementation.

The distribution of GAI research articles highlights diverse vital themes, including bias, hallucinations, misuse, societal impact, and transparency. Each perspective addresses a critical challenge in advancing responsible and ethical GAI. Research on *bias* focuses on identifying and mitigating inherent prejudices within models, while *hallucinations* examine the generation of inaccurate or fabricated outputs. Studies on *misuse* explore the potential for GAI to be applied in harmful or unintended ways, whereas *societal impact* considers the broader effects of GAI on communities and social structures. Lastly, *transparency* emphasises the need for precise, interpretable models to build trust and accountability in AI systems. This distribution reflects the multifaceted nature of GAI and the comprehensive efforts to address its implications across these critical areas.

4 Literature review

Generative AI technologies are predominantly centred around image creation, using text input as the basis for generating graphics in two, three, and even four dimensions. These systems can produce highly accurate portraits of individuals. However, despite these impressive feats, a significant challenge remains—generative AI still struggles with drawing hands, making it one of its most authoritarian limitations. One of the most popular uses of AI-generated images is for creating avatars, with platforms like TikTok being a significant hub for sharing these AI creations. With generative AI, users can transform into superheroes, kings and queens, or even mermaids. Another emerging trend is the AI-generated QR code, which replaces the traditional black-and-white version with visually appealing designs. Hugging Face has made a free AI-powered QR code

generator available to the public. Investment in generative AI has surged, with funding exceeding two billion dollars, marking a 425 per cent increase since 2020. Looking ahead, generative AI is widely considered one of the most advanced technologies for the next half-century, with some scientists even dubbing it a potential "age-reversing" technology.

They resembled human-generated content or data produced in high quality by using GAI [44, 59]. Since the year 1970, using rule-based patterns and simplistic scripted responses, basic computers have been included in the research on conversational chatbots, such as algorithms on a heuristic for student consultants [28], ELIZA by professor of MIT Joseph Weizenbaum [157]. The early enhancement of AI was not able to explore novel text and not able to proceed with a deep understanding of the content. According to the usage and requirement, the implementation of chatbots varies on the data size, and trained datasets are used in web crawls, social media, crawls and encyclopedias based on the large language model [131]. GAI occupies a significant role in the development process in various fields and domain enhancement based on different paradigms and themes. The research perspective was based on multiple fields: governance, infrastructure and technique, education and human capital innovation.

Previous studies on Generative Artificial Intelligence (GAI) have concentrated on a variety of fields, with a notable focus on education and research [168], healthcare [138], structural and urban design, as well as art and entertainment [2, 73]. Among these, healthcare and education have been the most extensively researched areas. In particular, GAI is pivotal in pediatric radiology, contributing to accurate diagnosis and decision-making. Despite the potential of GAI in clinical adoption, further improvements are necessary for its integration, particularly in predicting accurate diagnostic outcomes.

Ethical considerations have also driven a significant portion of GAI research, particularly in the context of pedagogical and assessment design in education. Several literature reviews have highlighted how GAI applications in education are deeply intertwined with ethical implications, particularly regarding bias and fairness [6, 168]. Furthermore, recent studies suggest that GAI-generated music and visual arts closely match the quality of human-created works, indicating its potential in creative fields [152]. GAI has also been leveraged to optimise structural and urban design project outcomes, aiding in resolving stakeholder conflicts [73].

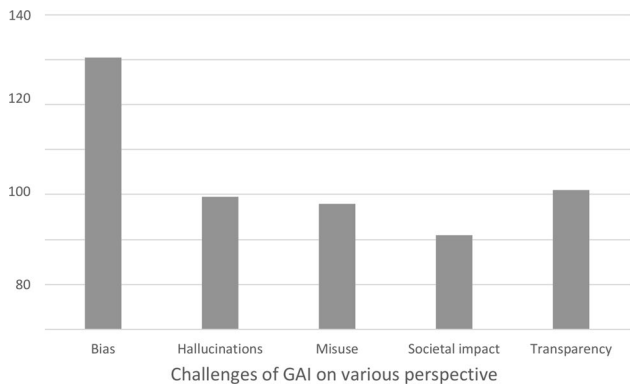

Fig. 5 Distribution of GAI research articles on different perspectives

While previous research has made significant contributions from an interdisciplinary perspective, there remains a gap in comprehensive reviews that address GAI across multiple domains. Much existing literature has focused on specific academic or industry sectors, often using singular tools and methodologies for cross-disciplinary reviews. The literature to date has primarily consisted of cross-disciplinary reviews, with only a limited number of empirical studies included.

Given the multifaceted nature of GAI, there is a pressing need for comprehensive and systematic reviews that provide a holistic understanding of GAI across various applications. A deeper exploration of GAI in human-technology interactions is also critical to better understanding the challenges involved in its implementation. Moreover, systematic reviews on GAI's role in mitigating bias across different modalities, such as text, video, audio, image, and code, have been conducted, highlighting the importance of addressing these issues.

The challenges associated with enhancing GAI, such as mitigating bias, ensuring transparency, reducing hallucinations, preventing misuse, and addressing societal impacts, have been key topics in recent research. These challenges have been examined through various perspectives, further emphasising the need for more in-depth, cross-disciplinary studies to understand and address the complex issues surrounding the development and deployment of GAI (Fig. 5).

4.1 Classification of articles based on key findings

4.1.1 Articles classification by journals

The reviewed articles were published across various journals, underscoring the interdisciplinary interest in addressing bias in generative AI. As shown in Fig. 6, most articles were published in journals dedicated to artificial intelligence and machine learning, reflecting the technical focus of much of the research. The Journal of Machine Learning Research and IEEE Transactions on Neural Networks and Learning Systems emerged as prominent outlets, with seven and six articles, respectively. This strongly emphasises developing new algorithms and methods for detecting and mitigating bias within AI systems.

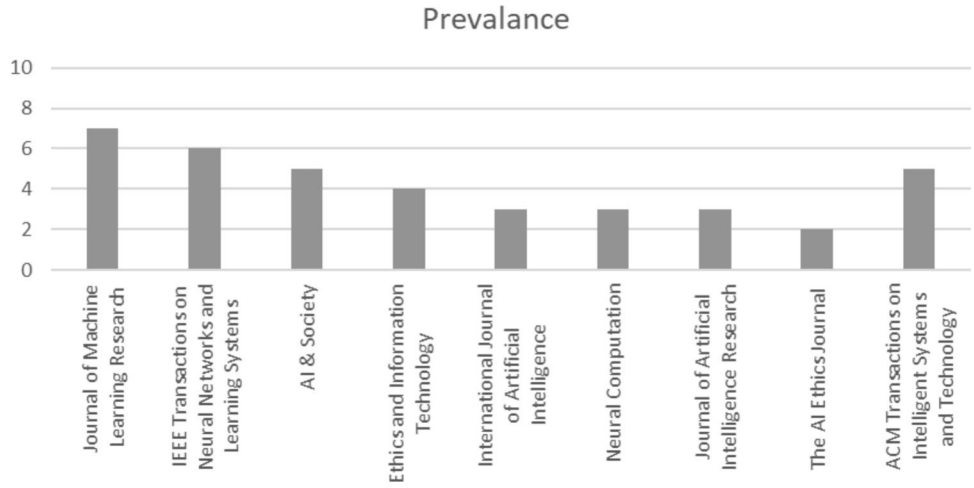

Fig. 6 Articles classification by journal

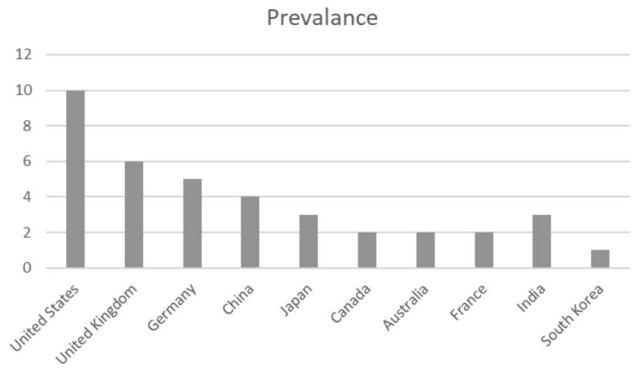

Fig. 7 Articles classification by country

In addition to technical journals, publications in ethics and societal impact journals, such as AI & Society and Ethics and Information Technology, were also featured prominently in the review. Together, these journals published nine articles, highlighting growing concerns about the ethical implications of AI bias and the broader societal impact of AI technologies. These papers advocate fairness in AI systems and emphasise the importance of addressing bias in their design and implementation.

Other notable journals contributing to the discourse include the International Journal of Artificial Intelligence, Neural Computation, and the Journal of Artificial Intelligence Research, each of which published three articles. This diverse distribution of journals reflects the multidimensional nature of the AI bias problem, encompassing technical, ethical, and social considerations. The range of journals involved suggests that bias mitigation in generative AI requires a comprehensive, interdisciplinary approach.

4.1.2 Articles classification by country

The global scope of this study reflects the widespread concern regarding bias in generative AI. Figure 7 illustrates the classification of articles by country, revealing the international nature of this issue. The United States leads with the highest representation, contributing ten articles highlighting the country's significant interest in AI development. This can be attributed to its well-established infrastructure, ample funding, and numerous tech companies and research institutes dedicated to AI.

Following the United States, the United Kingdom contributed six articles demonstrating the country's active engagement in research and industrial practices related to the ethical development of AI. Germany, a key European player in AI ethics, submitted five articles, reinforcing its strong commitment to addressing AI-related challenges. With four articles, China underscores its growing influence in AI and its increasing recognition of addressing bias in AI systems.

Japan reflects the country's ongoing efforts to promote ethical AI development. Canada and Australia each contributed two articles, signalling their participation in this global conversation. France and India also submitted two articles each, underscoring their active roles in working towards solutions for AI bias. South Korea and other countries, with one article each, further demonstrate the global acknowledgment and engagement with the issue of AI bias.

The distribution shown in Fig. 7 highlights that, while some countries are more prominent in the field, there is a broad, transnational effort to address the challenges of bias in generative AI. This global participation reflects the widespread recognition of mitigating bias in AI systems and the collective effort to ensure fairness and accountability in AI development.

4.1.3 Articles classification by methodology

The methodologies employed in the analysed articles varied, reflecting the inherent complexity of studying bias in generative AI. Figure 8 illustrates the classification of articles based on their methodological approach. Quantitative

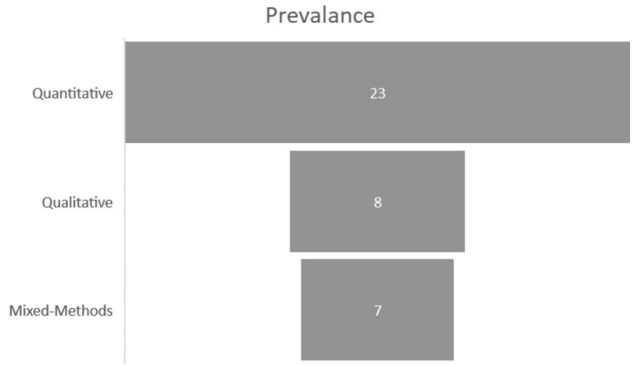

Fig. 8 Articles classification by methodology

methods were the most used, with 23 articles adopting this approach. These studies utilised statistical analysis, machine learning experiments, and simulations to examine and mitigate bias. The quantitative design is particularly effective for defining objective measures and evaluating the extent of bias, making it well-suited for technical research in AI.

In contrast, eight studies employed qualitative methodologies, which included case studies, interviews with AI professionals, and ethnographic research to explore the social and human aspects of AI bias. Qualitative research is valuable for addressing AI bias's ethical and societal impacts, offering insights into the broader contextual and strategic considerations necessary for effective mitigation.

Additionally, seven articles utilised mixed-methods approaches, combining quantitative and qualitative techniques. This hybrid approach allows researchers to leverage the strengths of both methods, providing a more comprehensive understanding of the issues at hand. The diversity of research methods underscores the multifaceted nature of bias in generative AI, highlighting the need for various approaches to understand and address the problem entirely.

4.2 Bias mitigation in generative AI

AI bias refers to prejudiced outcomes of unjust, discriminating, or skewed patterns generated by an algorithm due to erroneous presumptions inside the machine learning (ML) procedure.

Fairness can be widely conceptualised as the ethical principle of ensuring that AI systems produce an equitable output without prejudice or preference for an individual or group, irrespective of their inherent or acquired characteristics [100]. The relationship between bias and fairness in AI can be viewed as interconnected and mutually influential, like two sides of the same coin.

A fundamental distinction between fairness and bias lies in their nature: while bias can arise inadvertently, fairness is fundamentally characterised by conscious and deliberate pursuit. Bias can manifest from multiple sources, including biased data or algorithmic design.

However, achieving fairness necessitates a purposeful endeavour to prevent the algorithm from engaging in discriminatory practices towards any group or individual; prejudice might be perceived as technical, but fairness is regarded as social and ethical.

One notable distinction is that prejudice can manifest as positive or negative, whereas fairness solely pertains to addressing negative bias or discrimination. Positive bias is when an algorithm consistently favours a specific group or individual. In contrast, negative bias is when the algorithm consistently discriminates against a particular group or people. On the other hand, fairness pertains to preventing adverse bias or discriminatory treatment towards any group or individual.

Despite these differences, fairness and bias are often closely related, and addressing bias is an essential step towards achieving fairness in AI. For example, addressing bias in training data or algorithms can help reduce the likelihood of unfair outcomes. However, it is essential to recognise that bias is not the only factor leading to unfairness, and achieving fairness may require additional efforts beyond bias mitigation. [80].

Bias, a pervasive force in human perception and decision-making, encompasses a range of manifestations that impact various facets of our lives. From the cognitive biases that influence individual thinking processes to social biases deeply embedded in cultural structures and algorithmic biases emerging in artificial intelligence systems, Fig. 9 depicts the different types of bias that shape our understanding of the world and drive subtle and overt inequalities. Recognising and comprehending these biases is the first step towards cultivating a more informed and equitable society, fostering awareness and promoting the pursuit of fairness in diverse contexts.

Societal bias, deeply ingrained in cultural norms, institutions, and interpersonal interactions, profoundly influences individuals and communities. This pervasive bias contributes to systemic inequalities, affecting access to opportunities, resources, and justice. Whether manifested in discrimination, stereotyping, or prejudice, societal bias shapes our perceptions, attitudes, and behaviours (Fig. 9).

Generative AI (GAI) has proven to be transformative across various domains, including education, healthcare, and networked businesses, by integrating new tools that enhance operations and facilitate innovation [24, 36, 43]. As depicted in Fig. 10, the influence of GAI spans multiple sectors, highlighting its significance in driving advancements and efficiency.

GAI is crucial in developing novel applications, particularly in generating creative content with distinct features.

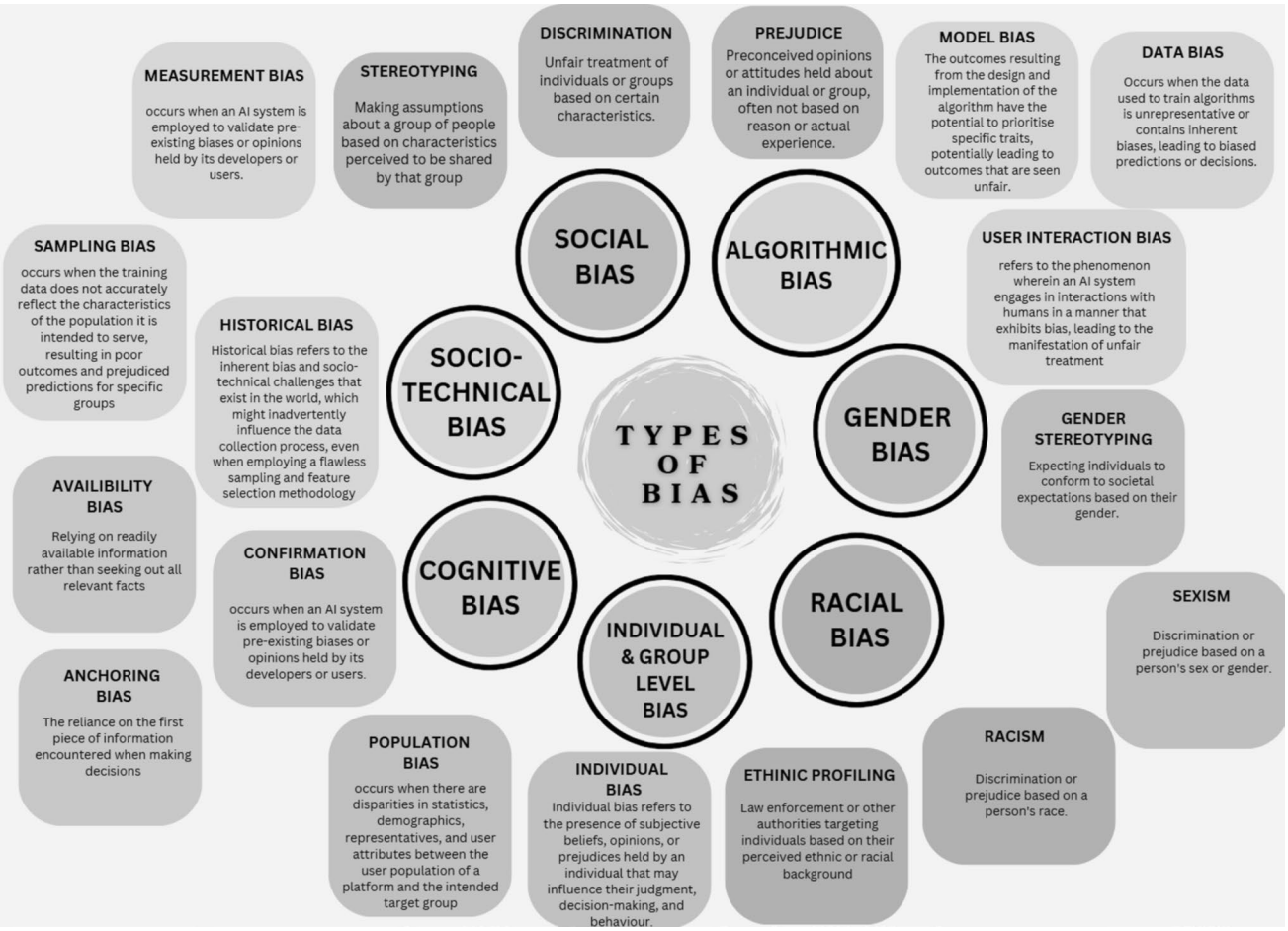

Fig. 9 Types of biases

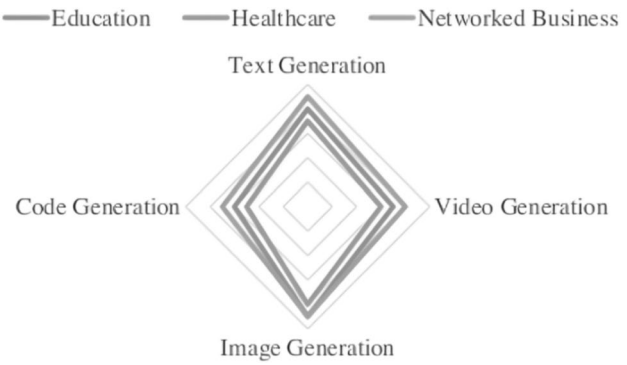

Fig. 10 Influence of generative AI across various sectors

In text and video generation areas, GAI contributes to creative outputs and fosters disruptive innovations, enabling the automation of complex tasks. Furthermore, GAI significantly impacts content creation across various modalities, including text, video, audio, image, and sound generation, offering diverse solutions for both personal and professional applications.

GAI has a notable positive influence in business, particularly in generating marketing content, software development, and multimedia such as videos and images. These capabilities facilitate the creation of high-quality, engaging content that enhances business outcomes [96]. Additionally, GAI supports various forms of content generation through different modalities, such as image-to-image and text-to-text generation, which produce outputs of the same type as the input. Furthermore, GAI enables the combination of various input and output types, such as code-to-text, facilitating the creation of multi-modal models [1, 32].

This diversity in GAI's application demonstrates its vast potential to revolutionise content creation, offering unique and versatile tools that can be leveraged across multiple sectors.

4.2.1 Bias mitigation in generative AI based on text

Natural language processing (NLP) and text generation have seen significant advancements with the development of X-to-text applications [52]. As depicted in Fig. 11, various aspects of text generation have been explored, with a

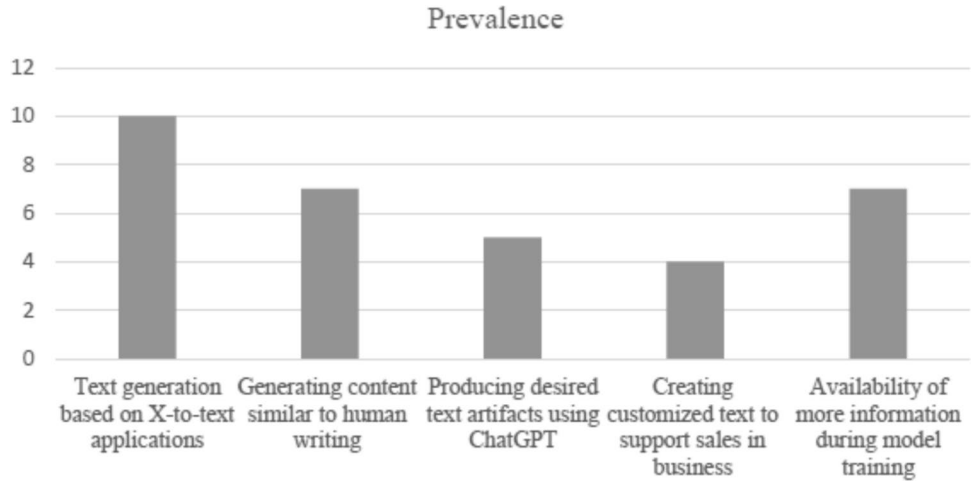

Fig. 11 Key Aspects of text generation in generative AI

primary goal of creating content that mirrors human writing. Generative AI systems, such as ChatGPT, have been instrumental in generating desired textual artifacts, enabling the production of coherent and contextually appropriate conversations [20]. Additionally, integrating text-producing applications further enhances content creation, making these tools indispensable in diverse applications such as customer support and business sales [106].

In the text generation process, GAI models like GPT-4 and BERT have made substantial contributions by advancing the capability to generate high-quality, contextually accurate text [89, 167, 172]. These models are trained on large, diverse datasets encompassing a wide range of text types, enabling them to understand language nuances, semantics, and contextual meaning. Transfer learning and fine-tuning techniques optimise these models for specific tasks, improving text summarisation, translation, and content generation [85].

Moreover, reinforcement learning from human feedback (RLHF) has become essential in fine-tuning these models, reducing the biases present in the training data [29, 31]. For example, in the case of GPT-4, RLHF relies heavily on human oversight to identify and remove biased, false, or misleading information, enabling the model to improve over successive feedback loops [7]. Additionally, adversarial training, where models are exposed to biased inputs and trained to generate unbiased outputs, has effectively mitigated biases in the text generation process [162].

The real-world implementation of these models has demonstrated promising results. In customer support, GAI-driven text generation systems enhance user satisfaction by providing faster, more relevant responses to queries. In content generation, these models facilitate the rapid production of high-quality articles, promotional messages, and other forms of writing. As a result, industries such as journalism, advertising, and literature are poised for significant transformation due to the efficiencies and capabilities offered by generative AI technologies [30, 78].

4.2.2 Bias mitigation in generative AI based on images

Generative AI has profoundly transformed the landscape of visual art, enabling the generation of images across diverse domains such as marketing, design, fashion, and entertainment. The advent of generative models has expanded the scope of creative expression by automating the process of creating visually compelling content. As shown in Fig. 12, image generation has evolved significantly, with applications extending to synthetic image creation, visual art production, and the development of interactive design elements for various industries [71]. Generative AI systems, such as DALL-E, Stable Diffusion, and other innovative platforms, allow users to create novel visual works by processing textual descriptions, thus bridging the gap between textual input and visual output.

One of the critical strengths of generative AI in image creation lies in its ability to produce highly customisable and contextually relevant images. The process begins with machine learning models trained on vast paired images and text datasets. This enables the model to learn the relationships between visual and textual information, allowing it to generate visually accurate images aligned with the input text's semantic meaning. For example, using natural language processing techniques, users can provide specific textual descriptions, and the system will generate an image that reflects the details of the prompt. This opens up many possibilities in fields like advertising, where generative AI can create tailored visual content for campaigns, or in fashion design, where designers can quickly prototype new styles and patterns.

A handy application of generative AI in image creation is its ability to edit and manipulate existing images. The image-to-image technique enables the transformation of

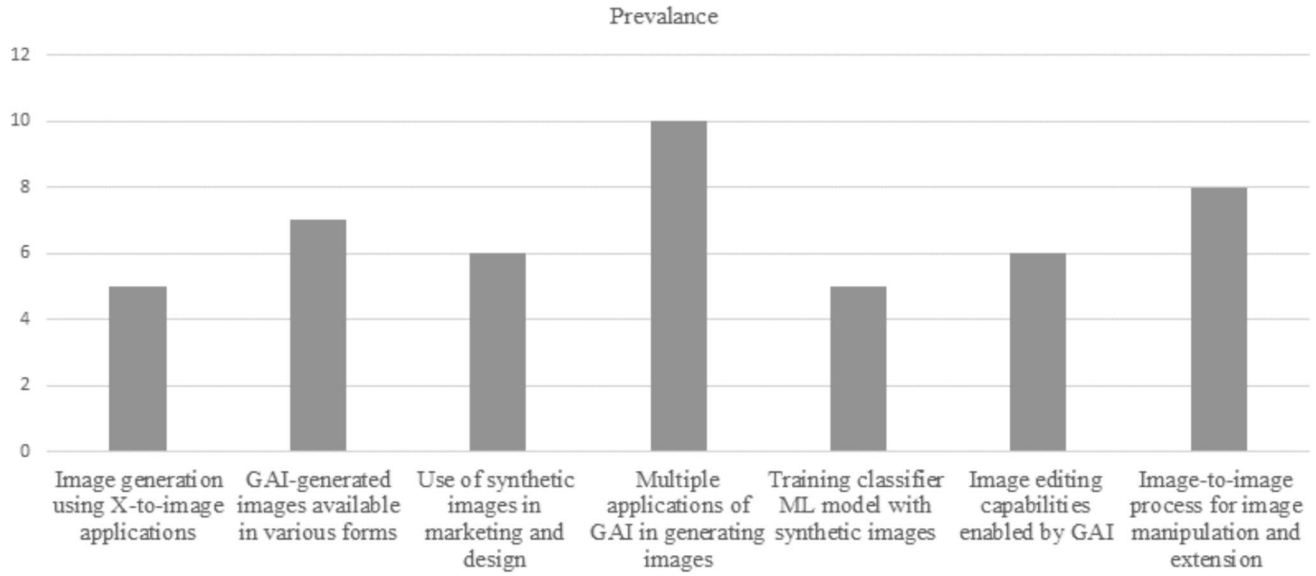

Fig. 12 Key aspects of image generation in generative AI

images based on user specifications. This capability is critical in creative fields such as fashion, graphic design, and entertainment, where customisation and iteration are often required. For example, an initial design can be manipulated to fit different contexts, colour schemes, or artistic styles, allowing designers to explore various possibilities efficiently. Image editing processes powered by generative AI also enable the enhancement of low-quality images or the creation of high-fidelity visuals from rough sketches, which is a valuable tool for both professional artists and hobbyists alike [27, 103].

Among the leading models in image generation, DALL-E and Stable Diffusion have garnered significant attention due to their groundbreaking capabilities in creating high-quality images from textual descriptions. DALL-E, for instance, utilises a transformer-based architecture and a two-step generation process. Initially, the input text is converted into a series of embeddings through a text encoder, which the diffusion model then uses to generate a detailed and contextually appropriate image. This model's capacity to create novel and diverse images based on simple text prompts has revolutionised marketing and content creation industries, enabling quick and highly customisable design workflows. On the other hand, Stable Diffusion utilises a continuous diffusion process to refine images from noise to clarity, progressively enhancing the image's quality and structure. This method produces visually striking images with minimal distortion, making it ideal for generating photorealistic visuals or abstract artistic compositions [128].

The application of Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) further enhances the quality and variety of the images generated by these models. GANs utilise a dual-network structure where a generator creates images, and a discriminator evaluates their authenticity, improving image quality as the two networks evolve. This process helps refine the output, ensuring the images are realistic, detailed, and aligned with the user's intent. VAEs, on the other hand, allow for the generation of diverse images by sampling from a continuous latent space, providing a broader scope for creativity and variation in generated content. When used together or in parallel, these models produce images with high fidelity and creativity, suitable for various industries, from entertainment to education.

Ethical concerns are significant in generative AI, particularly regarding the biases embedded in training data and the potential for harmful content generation. Several AI image generation models, including DALL-E and Stable Diffusion, have implemented Ethical AI frameworks to address these issues. These frameworks include content filtering mechanisms that ensure the generated images are free from offensive, biased, or inappropriate material. The Bias Detection Arrangements incorporated into these models help identify and mitigate any biases in the generated content, ensuring that AI-generated images align with ethical standards and avoid perpetuating harmful stereotypes or misrepresentations [79, 81].

In addition, the integration of Reinforcement Learning from Human Feedback (RLHF) has emerged as an effective method to fine-tune generative models. This process involves human evaluators providing feedback on the AI-generated outputs, which are then used to adjust the model's behaviour. This iterative approach helps improve the accuracy and relevance of the generated content, minimising biases and ensuring that the images produced are more aligned with ethical guidelines. For example, GPT-4, which

is often used in conjunction with DALL-E for text-to-image generation, benefits significantly from human oversight, allowing it to filter out harmful or biased elements and ensure higher-quality outputs [7, 29].

Furthermore, Adversarial Training is another technique that has proven helpful in reducing bias in AI-generated images. The model is exposed to biased inputs during its training phase, and the system is trained to produce unbiased outputs. This technique helps ensure the model is resilient to bias, creating more equitable and ethically sound images over time [162].

Generative AI's image-generation capabilities have also found practical applications in customer service, content creation, and entertainment. For instance, AI-generated images are used in advertising to quickly create promotional materials tailored to specific products or campaigns, significantly reducing business time and costs. In journalism, AI-based text-to- image systems assist journalists in visualising complex stories, making the news more engaging and accessible to audiences. Similarly, generative AI creates realistic and immersive environments and characters in video game design and animation, further blurring the lines between human creativity and machine-assisted innovation [30, 78].

As generative AI evolves, its ability to create high-quality, contextually relevant images will improve. The potential for new applications across industries such as healthcare, fashion, marketing, and entertainment will grow, providing users with increasingly sophisticated tools to create, edit, and enhance visual content. However, as these technologies advance, so too must the ethical frameworks and bias detection mechanisms that guide their development, ensuring that generative AI is used responsibly and equitably across all fields.

4.2.3 Bias mitigation in generative AI based on video

The development of generative AI has significantly advanced the field of video creation, particularly with the introduction of X-to-video applications. These models, which generate dynamic and synthetic video content from textual or image-based inputs, are revolutionising video production. As shown in Fig. 13, generative AI in video generation allows for the rapid creation and editing of video footage, which is particularly valuable in sectors requiring time-sensitive content production.

The X-to-video applications use natural language processing (NLP) and multimodal input to generate desired video sequences. These models create coherent video content by processing text descriptions, image prompts, and other relevant data, making the video creation process significantly faster and more accessible. The ability to generate videos based on simple text prompts or style-based inputs has empowered users without traditional video production skills to create high-quality videos [47]. This has led to the democratisation of video content creation, allowing individuals and organisations to produce professional-grade videos without needing specialised filming or editing training [173].

X-to-video models are particularly beneficial in generating creative content, such as short films, advertisements, or visual art. However, they have also found applications in more practical fields, including customer support (for creating product usage tutorials), education (for developing instructional and training videos), and sales/marketing (for producing product marketing videos). In these contexts, generative AI helps create engaging and informative video content efficiently, enhancing communication, training, and customer experience. The ability to generate these videos

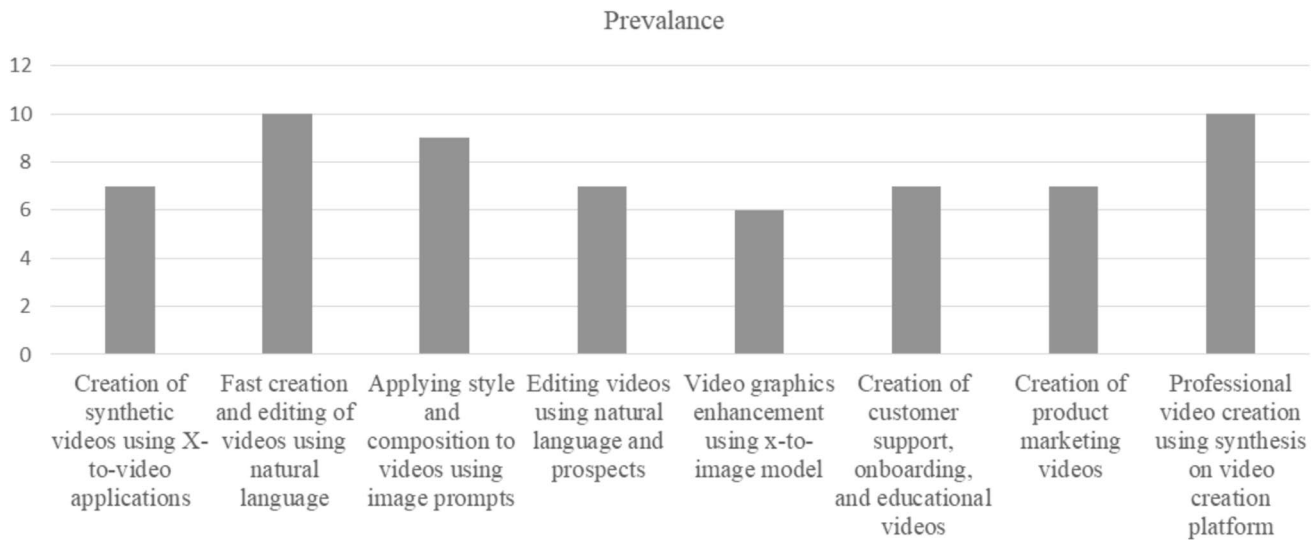

Fig. 13 Key aspects of video generation in generative AI

with minimal manual effort makes video production scalable and cost-effective, offering businesses and content creators the ability to produce videos at a fraction of traditional production costs [9].

Among the cutting-edge generative models, Vid2Vid and VideoGPT have made substantial contributions to synthetic video creation. Vid2Vid uses Generative Adversarial Networks (GANs) to generate high-quality video sequences from image sequences or text prompts. The GAN architecture is designed with a generator that creates video frames and a discriminator that assesses the realism and coherence of the generated content. This adversarial setup helps produce highly realistic video sequences that align with the input specifications [95, 175]. In contrast, VideoGPT employs a transformer-based architecture that generates video frames sequentially. Using self- attention mechanisms, VideoGPT models both temporal and spatial correlations in the video data, ensuring that each frame is contextually relevant and coherent with the entire video sequence [83, 166].

These advanced generative models have created videos that resemble real-life footage, opening up new entertainment, marketing, and education possibilities. Generating meaningful video content from text descriptions or existing video footage enables users to quickly produce high-quality videos that meet specific needs, enhancing creativity and productivity. Moreover, these tools have made video creation more inclusive, as individuals without video production expertise can now engage in video creation and editing [64].

One of the significant advantages of these generative models is their accessibility. With simple interfaces and step-by-step guidance, anyone can generate a video with minimal technical knowledge. This ease of use has fostered innovation and creativity across various industries. It has enabled individuals and businesses to quickly bring their ideas to life without expensive equipment, skilled personnel, or complex editing software. The accessibility of these tools has contributed to the rise of new content creators, democratising video production and giving rise to novel forms of digital expression.

However, as with any technology, ethical considerations around using generative AI in video creation remain paramount. The ability to create realistic videos with little effort raises concerns about the potential for misuse, such as creating deepfakes or misleading content. Addressing these concerns requires robust mechanisms to detect and prevent harmful content generation. Many generative AI platforms already incorporate features such as bias detection, content filtering, and ethical guidelines to ensure the technology is used responsibly [9].

In conclusion, the advancements in X-to-video generative AI models have revolutionised video production and expanded creative possibilities across various sectors. These models empower individuals and businesses to create high-quality video content quickly and affordably by enabling easy and efficient video generation. As these technologies evolve, they are expected to drive further innovation, providing new opportunities for creativity and productivity. However, it is crucial to ensure these technologies are used responsibly and ethically to avoid potential misuse and ensure their positive impact on society.

4.2.4 Bias mitigation in generative AI based on code generation

Applying X-to-code generative AI models has significantly transformed the software development process by enabling the automatic generation of programming code from natural language prompts. As illustrated in Fig. 14, these applications have revolutionised code writing, offering powerful tools that streamline development, enhance productivity, and reduce time-to-market for software products. By integrating X-to-text technologies, generative AI models can generate code in specific programming languages by

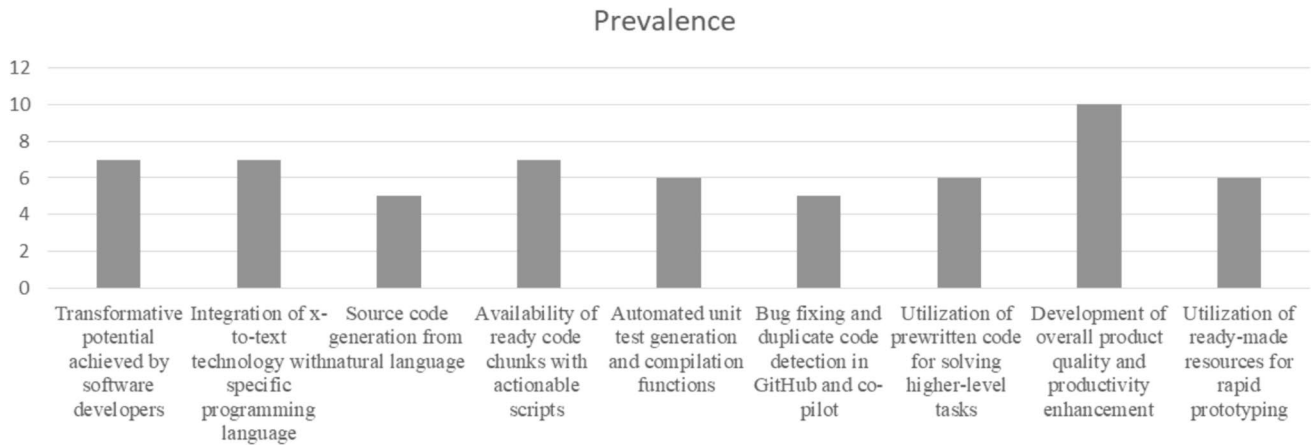

Fig. 14 Key Aspects of Code Generation in Generative AI

interpreting textual descriptions or instructions [90]. This advancement allows software developers to leverage AI-driven code generation tools that translate human-readable prompts into executable code, reducing the effort and time required for coding tasks [49, 135].

Generative AI in code development offers numerous functionalities such as generating chunks of code, automating unit test creation, compiling code, bug fixing, and detecting duplicate code. Platforms like GitHub Copilot and CoPilot provide developers with ready-to- use code snippets, enhancing their ability to solve higher-level tasks quickly. Developers can significantly improve their productivity, leverage problem-solving techniques, and reduce development time using these AI-powered tools. These platforms also facilitate rapid prototyping, allowing for faster iteration of products and quicker market deployment [3].

Technically, several advanced generative models have significantly enhanced code generation capabilities. Notable models such as Codex, the engine behind GitHub Copilot, and AlphaCode have substantially contributed to software development. Codex uses a transformer-based architecture to generate code from natural language descriptions, offering suggestions and auto-completions that optimise the coding process. This model is trained on a large corpus of open-source code, enabling it to generate relevant code snippets, functions, and even entire programs, thus assisting developers in writing high- quality code faster [93].

On the other hand, AlphaCode, developed by DeepMind, integrates supervised learning and reinforcement learning to generate high-quality code. Trained in diverse coding challenges, AlphaCode excels in solving complex programming problems and creating fine-tuned solutions. One of its key features is using few-shot learning, which allows the model to generate new code with minimal additional training, thereby enhancing flexibility and efficiency [86]. These generative models have improved coding speed and foster more reliable and innovative software development practices.

The integration of AI-driven code generation tools has dramatically improved the productivity and efficiency of developers by automating routine coding tasks, enabling them to focus on more strategic aspects of software design. The time saved from writing and testing code is reinvested into improving the quality of the software, leading to better development cycles, higher reliability, and more innovative solutions in technology [86, 93, 117, 132].

In conclusion, generative AI in code generation has significantly enhanced software development processes by automating code creation, bug fixing, and testing while increasing overall development efficiency. With models like Codex and AlphaCode, developers can write code more quickly and with greater accuracy, enabling them to focus on software development's creative and design aspects. Integrating these AI tools into development environments can improve software reliability, reduce development time, and foster innovation, ultimately transforming the software engineering landscape (Fig. 15).

4.2.5 Bias mitigation in generative AI based on audio

Generative AI has made substantial strides in producing high-quality, human-like synthetic voices and music, crucial for applications ranging from digital assistants to content creation. As outlined in Fig. 15, X-to-audio applications have transformed how audio content is generated, facilitating various uses such as customer service, audiobook narration, and interactive voice response systems [5, 69]. Using text-to-speech (TTS) and speech-to-speech models, generative AI can produce highly realistic voice outputs, enhancing user engagement and the overall experience in digital interactions. For instance, Microsoft's VALL-E

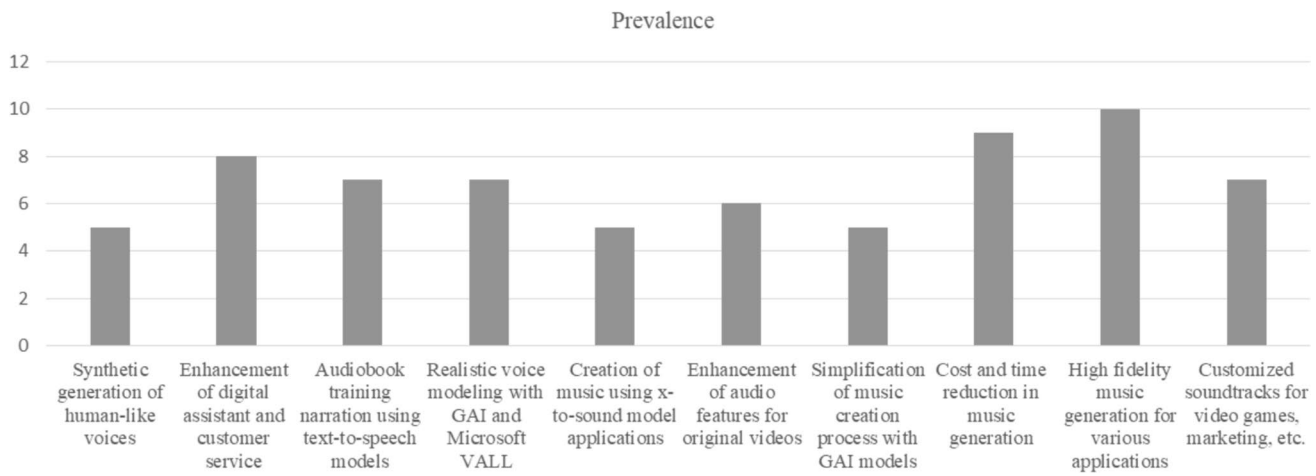

Fig. 15 Key aspects of audio generation in generative AI

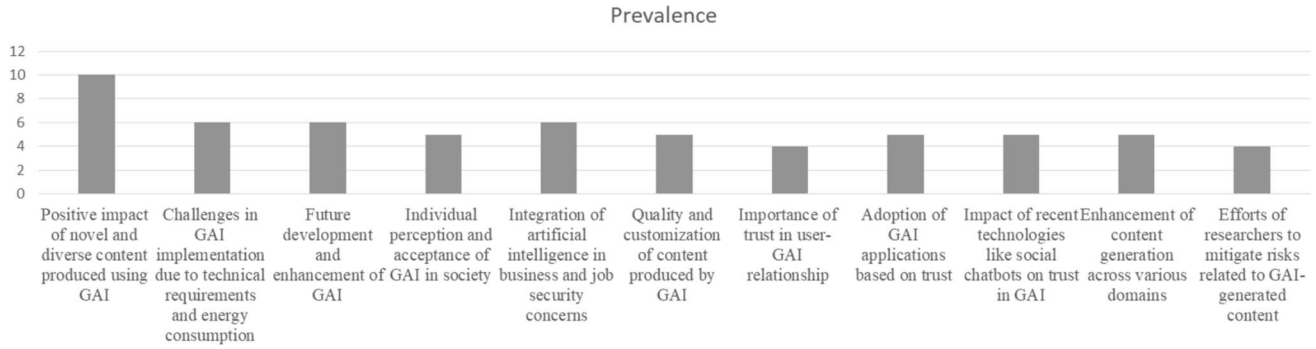

Fig. 16 Different aspects of biases with various challenges as prevalence

model exemplifies this shift, enabling nuanced voice modelling for various content creation needs [4].

Generative AI also plays an integral role in music generation through X-to-sound models, which streamline the music production process by automating the creation of customised soundtracks and compositions. This approach is especially beneficial in gaming, film, and marketing, where high-quality, tailor-made music can significantly elevate the content. Unlike traditional music production, which can be time-consuming and costly, AI-driven music generation offers an efficient, cost-effective solution for creating soundtracks that match the specific requirements of a project [88].

Technological advancements in generative models such as Tacotron, WaveNet, and Jukedeck have been pivotal in pushing the boundaries of audio synthesis. Tacotron, a model developed by Google, uses sequence-to-sequence architecture to generate lifelike speech from text inputs. It leverages recurrent neural networks (RNNs) combined with an attention mechanism to map text sequences to mel-spectrograms, which are then converted into audio waveforms by a WaveNet-based vocoder. This approach enhances the naturalness of generated speech, making it highly intelligible and suitable for TTS systems [133].

WaveNet, also developed by Google, takes a different approach by directly generating high- fidelity audio waveforms through dilated convolutions. This technique enables the model to synthesise natural speech patterns with detail and accuracy that closely resembles human speech. WaveNet's ability to model intricate audio waveforms has made it a prominent tool in TTS and voice generation applications, where audio output quality is crucial [160].

Jukedeck uses neural networks and rule-based systems to create original music across various genres and tempos in music generation. By analysing large datasets of musical compositions, Jukedeck can produce realistic, genre-specific music that aligns with a user's specifications, making it ideal for background scores in multimedia content like videos and games [139]. The versatility of Jukedeck and similar models demonstrates the potential of AI to enhance creative industries by enabling musicians, content creators, and developers to generate high-quality audio content with minimal time investment.

In summary, the use of generative AI in audio and music synthesis has significantly impacted various sectors by streamlining audio production, reducing costs, and expanding creative possibilities. Models like Tacotron, WaveNet, and Jukedeck exemplify the advancements in human-like audio generation, underscoring AI's role in transforming speech and music generation. As these technologies evolve, they offer the potential to revolutionise content creation in customer service, media production, and entertainment, setting new standards for quality and customisation in audio content.

4.2.6 Bias mitigation in generative AI based on other aspects

Generative AI applications have proliferated across diverse fields, enabling advancements in drug discovery, molecular modelling, and 3D rendering. These applications utilise specialised GAI models tailored to each domain's unique demands. In pharmaceuticals and bioengineering, models like **AlphaFold**, developed by DeepMind, provide high-precision predictions of protein structures from genetic sequences. This breakthrough, initially developed to understand protein folding [35], has revolutionised drug discovery by enabling accurate molecular modelling, accelerating the identification of drug targets and the design of molecules that effectively interact with these targets. By refining the understanding of protein structures, AlphaFold facilitates faster, more targeted drug development, ultimately improving the efficacy of therapeutic interventions [119].

Similarly, **DreamFusion** represents an innovative GAI model that has impacted 3D modelling and rendering. Leveraging a neural network architecture combined with advanced graphical algorithms, DreamFusion produces high-quality 3D models from minimal input, such

as sketches or text descriptions. This efficiency streamlines the design process by reducing the time-intensive stages of traditional 3D modelling. Consequently, DreamFusion has become widely adopted in fields requiring detailed virtual models, including virtual reality, game design, and film production. By optimising the rendering process, DreamFusion empowers artists and developers to focus on the creative aspects of design, thereby enhancing user engagement and the realism of visual experiences in the entertainment industry [84, 158, 164].

In addition to its use in creative industries, DreamFusion has also been applied in research and development contexts within the pharmaceutical and biotechnological sectors, where high-fidelity 3D representations of molecular structures aid in understanding complex biological interactions. The tool enables companies and institutions to model potential drugs and test their efficacy virtually before moving to costly physical trials, thus speeding up innovation in drug discovery [114].

Below are ten different tools that can be used for various output generation (Table 1).

4.3 Critical challenges in generative AI

4.3.1 Bias

Generative AI (GAI) technologies have undeniably transformed industries by enhancing productivity, creativity, and innovation. However, with these advancements comes a spectrum of challenges related to bias that affects organisations, individuals, and society. These biases often stem from imbalances in training data, algorithmic decisions, and unintended reinforcement of societal stereotypes, presenting profound implications across different domains. Figure 16 shows different aspects of Biases with various Challenges as Prevalence.

Table 1 Generative AI tools

Text	Visuals	Audio	Code	3D Models
Jasper	Fotor	Lovo.ai	ChatGPT	Move.ai
Copy.ai	DALL-E 2	Synthesis	GitHub Copilot	Lumirithmic
Notion AI	Midjourney	Murf	TabNine	Elevate3D
GrammarlyGO	Dream Studio	Voice Over by Speechify	Replit Ghostwriter	Get3D
Writesonic	StarryAI	Altered Studio	MutableAI	3DFY.ai
Sudowrite	Pictory	Listnr	Seek	Sloyd.ai
Frase	Synthesia	Researcher	AI2SQL	Luma AI
Hypotenuse AI	HeyGen	Play. ht	Enzyme	Masterpiece Studio
ParagraphAI	Colossyan	Speechelo	Durable	Google DreamFusion
Chibi	Deepbrain	Amper Music	Stenography	RODIN

The reliability and quality of outputs generated by Generative AI (GAI) are critical, particularly in real-world and business contexts where the technology's adoption hinges on its feasibility and ethical soundness. However, biases within GAI models present notable risks, including discrimination and unintentional disadvantage across various applications. Bias can infiltrate GAI systems during two main stages—training and inference—where flawed data and imbalanced sampling can influence model outcomes and reinforce biases. Examining the origins, impacts, and mitigation strategies for biases within these models is imperative to advance the effective deployment of GAI (Figs. 17).

Biases commonly enter GAI models during the **training phase**, often due to the characteristics of the datasets used. Large-scale training datasets are essential for GAI development, and models frequently rely on scraped publicly available data. However, the quality and representativeness of this data are only sometimes assured, particularly in unsupervised data collection processes, which may result in data bias [94]. This bias stems from imbalanced or unrepresentative sampling practices, ultimately generating skewed and potentially discriminatory outputs. Consequently, models trained on such data may fail to generalise across diverse user groups.

In the **inference phase**, biases can also arise as the model makes predictions based on biased training data, leading to algorithmic bias. This bias can negatively affect user interactions, especially in applications within sensitive domains, such as e-commerce, where biased recommendations or decisions could disadvantage specific customer groups [163]. Overfitting further reinforces algorithm bias, wherein models learn specific patterns from training data that do not represent real-world scenarios. This limitation underscores the importance of representative and diverse datasets, as GAI models lacking in these aspects can produce outputs that amplify existing social or cultural biases [30].

Bias within GAI-generated content, including text, images, and videos, poses serious concerns for industries dependent on user engagement and experience, particularly in **business and e-commerce**. Social bias embedded in GAI models, mainly when used in customer-facing applications, can result in discriminatory interactions, negatively impacting brand reputation and trust. For instance, text and image generation models have been observed to produce socially biased outputs, which can inadvertently harm the perception of fairness and inclusivity within business environments [30]. Therefore, mitigating these biases is crucial to preserving service quality and maintaining equitable user engagement.

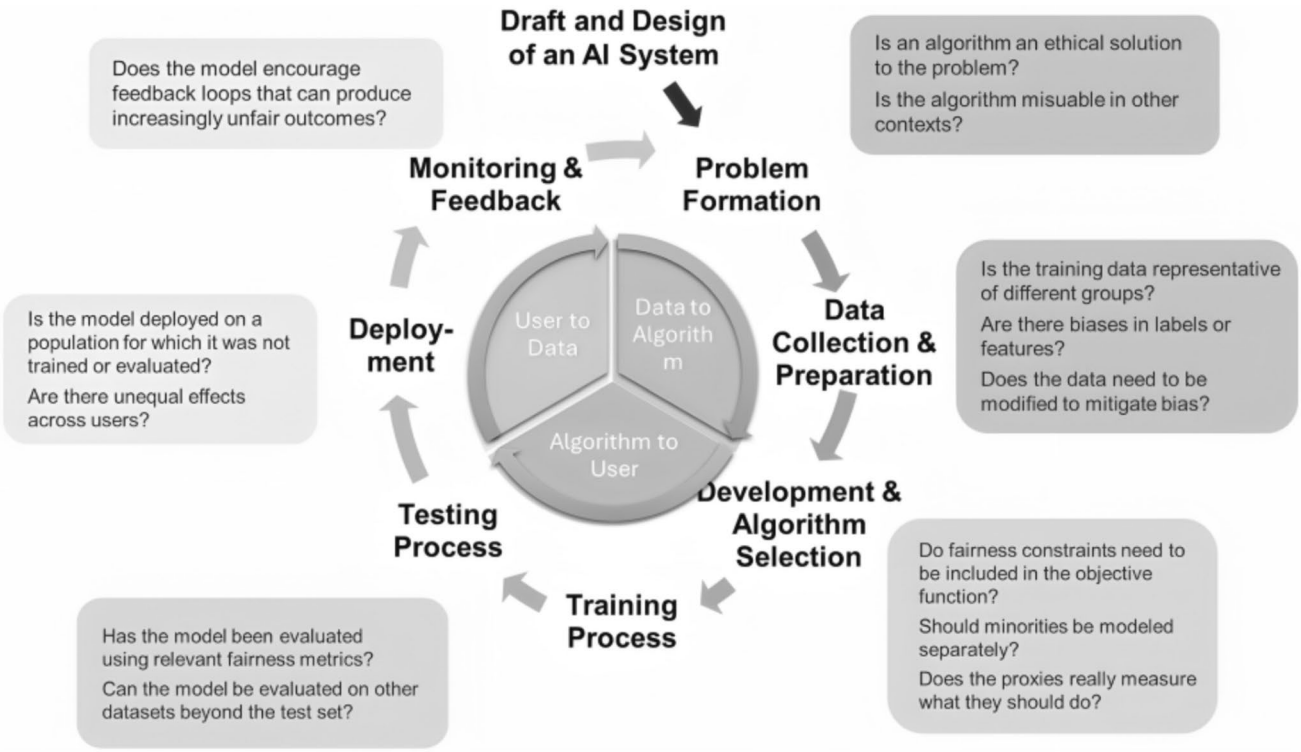

Fig. 17 Lifecycle for addressing bias in generative AI

4.3.2 Transparency

Integrating machine learning (ML) models with the generative nature of Generative AI (GAI) has facilitated complex functionality but often results in unpredictable outputs. While traditional discriminative AI models are designed to classify and make predictions, they need more capacity for creativity than GAI models can provide. This generative capability, however, also presents unique challenges in reliability, transparency, and interpretability [144]. Compared to traditional AI systems, GAI models provide more dynamic and nuanced content but at the cost of increased unpredictability, which raises concerns in sectors that require high accuracy and regulatory compliance.

GAI's interpretability and transparency become essential in digital networks and platforms where large user groups share and access vast information. It is necessary for business applications, as GAI serves as an advisor, interacting with users based on contextual inputs. However, due to the varying accuracy of GAI, there is a risk of misinformation. For instance, inaccurate product recommendations in an e-commerce setting could lead to customer dissatisfaction or even adverse effects, as with recommendations for medical products [101]. This example underscores the importance of validating GAI outputs, especially when high-stakes decisions are involved.

Transparency in Generative Artificial Intelligence (GAI) is a significant challenge, given the complexity of its underlying algorithms and the opaque nature of many of its models.

Transparency refers to understanding, interpreting, and trusting the decision-making process behind GAI outputs. However, achieving full transparency has proven difficult with the widespread use of advanced neural networks and deep learning algorithms. The "black-box" nature of many GAI models, huge language models, means that even developers may have a limited understanding of how specific outputs are generated [121].

GAI models are often trained on vast datasets from the internet, presenting legal and ethical concerns. For instance, closed-source GAI models, such as those used by large tech firms, rely heavily on web-scraped data that may contain copyrighted material, personal information, or biased content. This lack of transparency regarding the source and handling of data complicates users' and regulators' ability to assess the fairness and integrity of the generated content [9, 16].

Moreover, in closed-source models, transparency is further hindered by proprietary constraints that prevent outside parties from accessing the model architecture, training data, or fine-tuning processes. This opacity makes it challenging for users and organisations to understand why a model makes particular decisions, directly affecting accountability

and trust [13]. As GAI continues to be applied in critical areas—such as healthcare, legal services, and finance—understanding the processes driving its outputs becomes imperative to ensure that these applications adhere to ethical and regulatory standards [10].

For business applications, GAI's potential lies in providing reliable, adaptable outputs that align with stakeholder goals [125]. However, the issue of transparency between open-source and closed-source GAI models remains contentious. Closed-source GAI models often rely on extensive web scraping for large datasets, which can introduce challenges related to intellectual property rights, copyright, and licensing concerns [9]. Consequently, organisations using these models must navigate legal risks, while users may need help to trust outputs due to limited insights into the training data and algorithms used.

Increasing data transparency is crucial in open-source models. Techniques like watermarking generated content and images can enhance data traceability and ensure transparency for end users. For organisations, the safest approach to deploying GAI models involves selecting domain-specific models that leverage open-source tools and comply with regulatory requirements. The deployment of a system engineering approach to oversee GAI's role in business environments has been suggested to mitigate associated risks and better align GAI functionalities with organisational goals [13].

The lack of transparency also has implications for trustworthiness, as users are more likely to trust AI systems they can understand and interpret. This is particularly relevant in sensitive domains where biased or incorrect outputs could have significant repercussions. Consequently, addressing transparency is a technical challenge and a socio-ethical imperative, shaping public confidence in GAI's safe and responsible deployment [144].

4.3.3 Hallucinations

The phenomenon of "hallucinations" in Generative AI (GAI) refers to instances where the model generates incorrect or fictional information that appears factual. This challenge arises from the probabilistic nature of the data on which GAI models are trained, leading to variations in output accuracy [17]. Hallucinations are increasingly prevalent as GAI systems become more complex and are used in diverse applications. This issue is compounded when source data contain inaccuracies, fictional elements, or inconsistencies alongside factual information, resulting in outputs that reflect this unreliable mixture. Combining varied input sources fed into models through training data significantly contributes to hallucinations [17]. Closed-source models further exacerbate this issue by limiting transparency, making it difficult to assess the origin and quality of training data that might contribute to erroneous outputs.

In content generation tasks, such as automated social media posts or advertisements, hallucinations compromise the reliability of GAI systems. For instance, GAI image generators sometimes produce distorted or incorrect representations of people, which has been documented as a significant limitation in practical applications [78]. Additionally, basic computational errors—such as incorrect calculations in simple arithmetic operations—highlight potential gaps in the model's foundational abilities, underscoring the need for enhanced accuracy mechanisms.

Evaluating hallucinations is challenging, as the process requires distinguishing between factual and fictional data within highly realistic outputs. Both automated and human evaluations face limitations; while deep learning techniques can conduct statistical evaluations, human oversight is essential to assess nuanced and context-specific errors [153]. Given the inherent subjectivity and contextual complexity in assessing content accuracy, human evaluation remains a critical complement to automated assessment tools.

Current research advocates for open-source models to mitigate hallucination risks. Open- source models enable greater scrutiny of training datasets and algorithms, supporting error reduction through transparency. Furthermore, integrating robust error-checking mechanisms in GAI algorithms has been recommended to enhance reliability, trustworthiness, and the overall quality of output. Transparency in data sources and computational processes is essential to build user trust and improve the utility of GAI models in various applications [9].

4.3.4 Misuse

Generative AI (GAI) tools have presented a dual impact on society: while they foster new opportunities and applications, they also pose significant risks related to misinformation, identity manipulation, and intentional societal harm. Models capable of generating hyper- realistic outputs, such as x-to-sound, x-to-video, and x-to-image, have made it easier to create deepfakes, which can spread misleading or harmful information [125]. In particular, content featuring celebrities, public figures, or political leaders is frequently targeted and manipulated to deceive viewers, often resulting in fabricated media that appears indistinguishably authentic. This poses a challenge for distinguishing fact from fiction in daily life, complicating the task of identifying accurate information produced by GAI.

The misuse of GAI in creating fake online stores, identity theft, and fraudulent service offers highlights the role of GAI in facilitating digital deception [110; 108]. In politics and social media, the rapid spread of misinformation—amplified by GAI—can shape public perception, impact democratic processes, and cause societal harm. Although GAI holds potential for business and creative innovation, it also serves as a tool for criminals and hackers seeking to disseminate falsehoods or mislead users for personal or political gain. Furthermore, GAI models can inadvertently perpetuate discrimination, toxicity, exclusion, and information hazards due to biased or low-quality training data, raising ethical concerns about their deployment.

Researchers and policymakers emphasise the importance of safe, responsible, and ethical use of GAI content to mitigate these risks. Scholars such as Mantelero [97], Gu [58], and Golda et al. [56] focus on establishing guidelines to prevent harmful applications, promote transparent model outputs, and encourage the development of secure and high-quality datasets. Security measures, such as bypass filters to detect and counteract prompt injections by malicious users, represent essential steps in protecting the integrity of GAI applications. However, ensuring the safe use of GAI will require sustained collaborative efforts between researchers, industry professionals, and regulatory bodies to monitor, evaluate, and address the challenges GAI-generated content poses. Further research and cross-sector collaboration will be necessary to establish comprehensive safeguards for GAI's ethical integration into society.

4.4 Analysis of ensuring fairness and mitigating bias in generative AI

Ensuring fairness in AI and mitigating bias is paramount for creating ethical, inclusive, and equitable technological landscapes. If unchecked, AI systems can unintentionally perpetuate and amplify existing societal biases, leading to discriminatory outcomes [107]. Fairness is essential to upholding ethical standards and building trust among diverse user communities [100]. Biased AI can result in unjust treatment, reinforcing inequalities and eroding the fundamental principles of fairness and justice. To harness the full potential of AI for positive societal impact, it is imperative to prioritise fairness in the development, deployment, and ongoing monitoring of AI systems. Proactive measures, such as diverse and representative data, transparent algorithms, and ethical guidelines, are crucial for minimising biases and ensuring that AI technologies contribute to a more just and inclusive future. Achieving equity in AI requires a deliberate commitment to address biases throughout the development life cycle [142] (Table 2).

4.4.1 Development of systematic approaches to monitor and reduce bias in AI

Our analysis emphasises that systematically reducing bias and promoting equity in AI systems requires a multifaceted and continuous approach. Central to this process is algorithmic transparency, which enables a clearer understanding of AI's decision-making and allows for identifying and rectifying biases. As discussed in various sources, the importance of transparency in AI cannot be overstated in the context of fairness and accountability [108, 142] (Figs. 18 and 19).

4.4.1.1 Algorithmic transparency The concept of algorithmic transparency plays a crucial role in mitigating bias within AI systems. It refers to the ability to understand and evaluate the internal workings of an AI model, which is essential for identifying biased patterns within both the algorithm and the training data. Techniques like explainable AI (XAI) and model interpretability are pivotal in achieving this transparency. XAI emphasises designing AI systems that inherently include explanatory capabilities, which allows users to understand how models arrive at specific outputs [142]. In contrast, model interpretability involves creating methodologies that help stakeholders comprehend how decisions are made based on the data fed into the system [108, 169].

Our analysis underscores that increasing the transparency of AI systems helps uncover biases that may otherwise be obscured, enabling corrective actions to be taken. This transparency helps with fairness and empowers stakeholders to evaluate and refine models, fostering continuous improvement. When AI algorithms are transparent, it becomes easier to identify problematic biases in the data or design, which can perpetuate harmful societal inequalities [19]. Moreover, such transparency supports building trust among users, which is critical for the widespread adoption of AI technologies [169].

Additionally, a transparent AI system can facilitate the ongoing monitoring and auditing of models, ensuring that they evolve in line with ethical standards. This dynamic approach is essential in addressing emerging biases that may occur as AI systems are exposed to new data or changing societal norms. Integrating transparency into AI systems mitigates bias and enhances AI technologies' ethical integrity [108, 142]. This approach, grounded in algorithmic transparency, establishes a vital pathway toward reducing bias in AI systems and enhancing their overall trustworthiness and fairness, helping to shape a more equitable future for AI technologies.

Table 2 Comparative analysis of generative AI models: bias, fairness, and performance

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
Denoising Diffusion Probabilistic Models (DDPM)	Challenges exist due to dataset bias, emphasising the need for diverse and balanced datasets to eliminate biases in models Models trained on large-scale unfiltered data may reinforce biases, leading to ethical risks in generated images	Innovative Techniques: Text-to-image models utilise techniques like text inversion for concept control and conditional diffusion models for personalised generation Models are assessed based on image quality using metrics like Fr'echet Inception Distance (FID) and CLIP score, ensuring high fidelity and text-image alignment	Evaluation Criteria: Text-to-image models are evaluated based on image quality using metrics like Fr'echet Inception Distance (FID) and CLIP score	360-Degree Manipulation: Some models enable 360-degree manipulation by editing from a single view, expanding the possibilities of image editing Alleviation of Unintended Changes: These models help alleviate unintended changes in image content compared to GAN inversion methods, enhancing image editing precision	Dataset Bias: Models trained on biased text-image pairs may perpetuate biases related to race and gender, impacting the fairness of generated images Ethical Risks: Generated images may contain inappropriate or offensive content, posing ethical concerns and risks	Zhang et al. [172, 171]
Generative Adversarial Networks (GANs)	Diverse Dataset Types: Studies utilise datasets such as MNIST, CIFAR-10, UMN crowd dataset, Credit Card Fraud Detection dataset, and CUHK avenue for various applications, including anomaly detection and image generation	Complexity: Machine learning models can exhibit intricate structures with numerous parameters, leading to challenges in interpretability and transparency Subjectivity: Visual inspection by domain experts is sometimes necessary to subjectively evaluate the quality of generated samples, providing additional insights beyond quantitative metrics	Quantitative Evaluation Metrics: Commonly used metrics include the Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR) for assessing data quality	Enhanced Understanding: Transparency in machine learning models provides a deeper understanding of the algorithm's underlying characteristics and decision-making process Facilitates Evaluation: Transparency enables detailed examination of essential characteristics in decision-making and evaluation based on specific metrics, aiding in model assessment	Privacy Concerns: Transparency efforts may raise privacy concerns, particularly regarding data collection and usage, especially for historically marginalised groups Limited Effectiveness: Transparency may not always guarantee adequate understanding, especially if the data used for training models are already biased	Pagano et al. [113]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
Generative Adversarial Networks (GANs)	GANs address data scarcity in anomaly detection through data generation DCGANs, standard GANs, and cGANs are commonly used architectures	GANs are composed of two neural networks, a generator and a discriminator, engaged in a zero-sum game where the generator aims to produce realistic data samples while the discriminator tries to distinguish between accurate and generated data The training of GANs involves reaching a Nash equilibrium where the generator's output is indistinguishable from accurate data, leading to the discriminator making random guesses about the authenticity of the input data	Area Under the Receiver Operating Characteristic Curve (AUROC) is a widely used metric to evaluate the performance of GANs in anomaly detection applications, providing a measure of the model's ability to distinguish between normal and anomalous data points	cGANs allow for conditional data generation, enabling control over the type of data generated, such as specifying to create data of a particular class or type The conditional model of GAN can be obtained by conditioning both the generator and discriminator on additional information fed through extra input layers, providing flexibility in data generation	In some cases, the performance of cGANs may be sensitive to the quality and relevance of the conditioning information provided, potentially leading to issues in data generation when the conditioning information is noisy or inaccurate	Sabuhi et al. [124]
Generative Adversarial Network (GAN), Variational Autoencoder (VAE)	Training Data: GAI models are trained on diverse, unbiased input data to learn patterns and generate content Bias Mitigation: Datasets play a vital role in addressing biases in GAI models, as biases can arise from the training data or the model's algorithm,	Creative Capabilities: GAI exhibits creative potential by generating novel and realistic content across various domains, such as texts, images, or code, based on basic user prompts Paradigm Shift: GAI represents a shift from data-driven, discriminative AI tasks to more sophisticated, creative tasks, offering unique use cases and opportunities in diverse domains	Data Quality Metrics: Metrics assessing the quality of data used to train GAI models, crucial for minimising biases and ensuring accurate outputs Transparency and Accountability Metrics: Metrics evaluating the transparency and accountability of GAI-based systems to ensure responsible use and decision-making	Creative Potential: Generative AI enables the creation of novel and realistic content across various domains like texts, images, and programming code based on user prompts	Data Quality Challenges: Like traditional AI, GAI models are susceptible to biases and discriminations due to data quality issues, impacting their performance and adoption in real-world business contexts Algorithmic Bias: Biases can be introduced during the inference phase, independent of the model's training dataset, affecting users due to biased algorithmic outcomes	Banh [9]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
BERTopic model for topic modelling and analysis within the field of Generative AI	The dataset underwent topic modelling using the BERTopic algorithm, which successfully extracted 23 topics from the corpus, enabling a deeper understanding of the prevalent themes in Generative AI research	Topic Extraction: BERTopic efficiently extracted 23 topics from the dataset, representing distinct clusters of themes and concepts within Generative AI research Topic Words: The model identified a set of lexemes, or 'topic words,' that encapsulated the principal themes and concepts within each topic, aiding in interpreting and understanding the content	Seven distinct clusters of topics in Generative AI research were identified, including image processing, content analysis, content generation, emerging use cases, engineering, cognitive inference and planning, data privacy and security, and Generative Pre-Trained Transformer (GPT) academic applications	Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction: UMAP retains more local and global features of high-dimensional data compared to Principal Component Analysis (PCA) and t-distributed Stochastic Neighbour Embedding	UMAP may require parameter tuning to optimise its performance, which can be time-consuming and computationally intensive, especially for large datasets The interpretability of UMAP embeddings can be complex, making it challenging to understand the exact relationships between data points in the reduced dimensional space	Gupta et al. [61]
Supply-side platforms (SSPs), real-time bidding systems through demand-side platforms (DSP), and data management platforms (DMPs)	randomly assigned to different experimental conditions to compare human versus machine attitudes towards AI-generated ads	Associated with higher ad attitudes compared to non-humanlike AI, particularly in emotional appeal scenarios Perceived as more machine-like and lacking human characteristics, potentially leading to lower ad attitudes among consumers, especially in hedonic appeal situations	Ad Attitude Measure: Utilized in the study to assess participants' attitudes towards AI-created ads, measured on a seven-point scale ranging from strongly disagree to agree strongly	Participants exposed to humanlike AI showed higher ad attitude scores than those exposed to non-humanlike AI, particularly in hedonic appeal scenarios This preference stems from the perception that AI cannot experience or feel emotions, making human agents more suitable for emotional appeals	Consumers tend to react less favourably to non-humanlike AI when promoting hedonic products, as they perceive AI as lacking the ability to convey emotions effectively	Marat Bakpayev et al. [8]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
Natural language processing (NLP)	Datasets like COMPAS, Communities and Crime (Communities), and Florida Department of Corrections (FDOC) are specifically mentioned for addressing crime-related bias and unfairness	Transparency and interpretability are key characteristics addressed in bias and fairness research, emphasising the importance of understanding how machine learning models make decisions	Various fairness metrics are used to assess bias and fairness in machine learning models, including Equalized Odds, Opportunity Equality, and Demographic Parity [14]	Mitigating bias and unfairness in machine learning models is crucial to ensure fair and ethical decision-making By using tools, metrics, and datasets to detect and mitigate bias, researchers can improve the overall fairness of AI systems Addressing bias in ML models can lead to more transparent algorithms, enhancing the interpretability of decisions made by these models	The lack of standardised fairness metrics across different use cases can lead to challenges in effectively assessing and mitigating bias Some existing solutions to mitigate bias and unfairness are often tailored to specific problems or use cases, limiting their generalizability	[68]
Machine Learning Models for Bias Mitigation	Data label bias refers to biases or inaccuracies in the labels assigned to training data, impacting model performance	Bias mitigation methods can involve pre-processing, in-processing, or post-processing techniques Techniques like adversarial training focus on improving model robustness by exposing it to adversarial examples during training	Fairness metrics like demographic parity, equalised odds, and equal opportunity are commonly used to measure bias in ML predictions [14]	Bias reduction methods in machine learning aim to lessen bias by altering data, the model itself, or adding fairness constraints Techniques like adversarial training can enhance model robustness and generalisation, improving performance in real-world scenarios	Implementing bias mitigation techniques, such as adversarial training, can increase computational and storage requirements for training models The complexity introduced by adversarial training may also raise questions about the necessity of additional techniques for achieving robustness	Siddique et al. [137]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
GPT-4 and Predecessors: Models like GPT-4, GPT-3, and their prior versions are commonly used for training large language models	WebText Dataset: The primary dataset used for training models like GPT-3 and its predecessors, containing text from a variety of web pages gathered by crawling the internet, The Pile Dataset, Social Media Platforms, Conversational Data	Training Data Biases: Biases absorbed from the source material or selection process can influence model behaviour, impacting fairness and inclusivity Algorithmic Biases: Algorithms may prioritise certain features or data points, introducing or amplifying biases in model outputs	Fairness Metrics: Assessing the model's performance in terms of fairness towards different demographic groups, ensuring equitable outcomes Accuracy Metrics: Evaluating the model's accuracy in generating unbiased and reliable outputs	Biases can sometimes help in improving efficiency by making predictions based on learned patterns In some cases, biases can align with user preferences, leading to more satisfactory outcomes	Biases can perpetuate stereotypes, favour certain groups, and lead to incorrect assumptions, impacting the fairness and inclusivity of the model's outputs They can result in unintended consequences, such as reinforcing cultural prejudices, promoting specific political perspectives, or excluding minority groups	[50]
Graph Embedding and Clustering Networks	UCI Adult Dataset- The UCI Adult dataset, also known as the 'Census Income' dataset, contains information extracted WinoBias Dataset- The WinoBias dataset follows the Winograd format, including 40 occupations in sentences linked to human pronouns. It is used to assess gender bias towards stereotypical occupations in reference resolution studies	Hybrid fairness, which requires treating both positively and negatively labelled cross pairs similarly, is a key characteristic in addressing fairness in AI systems Fairness definitions in AI systems aim to avoid biases by considering various sources of bias that can affect applications, leading researchers to develop taxonomies for fairness definitions to guide ethical AI development	Studies have shown that biases in AI systems can lead to unfair outcomes in various real-world applications, such as judicial systems, face recognition, and algorithmic decision-making processes	Addressing bias and fairness in AI systems is crucial to ensure that decisions made by these systems do not discriminate against certain groups or populations By mitigating bias, AI systems can make more equitable and just decisions, leading to fairer outcomes for individuals and society	Despite efforts to address bias, challenges still exist in achieving fairness in AI systems, with issues stemming from both data and algorithms Some widely used commercial AI systems, like COMPAS, have been found to exhibit biases and performance issues, indicating that bias mitigation is an ongoing challenge in AI development	Mehrabi et al. [100]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
Neural Network Architecture	The dataset used for training CodeBERT consists of 2.1 million bimodal data points and 6.4 million unimodal codes across six programming languages, including Python, Java, JavaScript, PHP, Ruby, and Go	Data is collected from non-fork GitHub repositories and filtered based on specific criteria such as project usage, documentation length, function size, and function names	The metrics help in quantitatively measuring the effectiveness of CodeBERT in capturing the relationships between natural language and programming languages, showcasing its state-of-the-art performance in NL-PL applications	CodeBERT is a bimodal pre-trained model for a programming language (PL) and natural language (NL), enabling it to learn general-purpose representations for various downstream NL-PL applications CodeBERT achieves state-of-the-art performance in natural language code search and code documentation generation tasks, showcasing its effectiveness in these domains	Traversing the tree structure of AST did not improve CodeBERT's performance on generation tasks, suggesting a potential area for enhancement by incorporating AST information in the model CodeBERT may achieve slightly lower results compared to models like code2seq, which utilise compositional paths in abstract syntax trees (AST), while CodeBERT only takes original code as input	Feng [49]
Neural Networks in GraphCodeBERT	BigCloneBench dataset for clone detection tasks and a dataset crawled from open-source projects for code translation tasks	Data Flow Extraction: The model utilises data flow, a graph representing dependency relations between variables, where nodes represent variables Semantic Understanding: Data flow helps in understanding the semantics of variables,	Mean Reciprocal Rank (MRR): MRR is used as an evaluation metric in tasks like code search, where the model ranks candidate codes based on their relevance to a given natural language query	Utilizes Data Flow for Learning Code Representation, State-of-the-Art Performance	Limited Comparison, Complexity in Implementation	Guo et al. [60]
Generative AI on the future of visual content marketing	Emphasizes the importance of visual content in marketing strategies and the impact of integrating artificial intelligence with visual content	Data-Driven Insights: AI analyses vast amounts of data from visual content campaigns, providing valuable insights into customer behaviours for informed decision-making in future marketing strategies Competitive Advantage: Companies utilising AI in visual content marketing gain a competitive edge by delivering targeted and relevant content, leading to higher conversion rates and market share	Memory Retention: Studies show that individuals are more likely to remember information presented in the form of relevant images, with a retention rate of 65% three days later, compared to only 10% retention for information presented without visuals	Enhanced Customer Engagement: Integrating artificial intelligence with visual content in marketing Increased Efficiency: AI can automate repetitive tasks in visual content creation, allowing marketers to focus on more strategic aspects of their campaigns	Lack of Human Touch: While AI can automate tasks, it may struggle to evoke empathy and emotional connections in visual content Skill Requirements: Implementing AI in visual content marketing requires specialised skills and knowledge	Mayahi [98]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
GPT-3.5 mode	The study utilised the Generative Pre-trained Transformer 3.5 (GPT-3.5) model for market research purposes, querying it to generate survey responses related to consumer preferences and willingness-to-pay estimates	Distributional Nature: GPT-3.5 exhibits a distributional nature in its responses, generating hundreds of survey responses to each prompt, which can provide a broad perspective on consumer preferences Sensitivity to Response Order: GPT shows a significant preference for options listed first, highlighting the importance of randomising response order to mitigate bias in survey results	Consistency with Economic Theory: GPT's responses broadly align with predictions from economic theory, demonstrating behaviours consistent with well-documented patterns of consumer behaviour	Consistency with Economic Theory: GPT-3.5 responses align with economic principles like downward-sloping demand curves and state dependence, reflecting well-documented consumer behaviour Realistic Estimates: GPT-3.5 provides willingness-to-pay estimates that match those obtained from human consumers, showing realistic magnitudes in its generated data	Diminishing Marginal Utility: GPT may struggle to simulate diminishing marginal utility accurately Prompt Sensitivity: The effectiveness of GPT's responses is highly dependent on how prompts are worded, requiring careful consideration to avoid misinterpretations or inaccuracies in the generated data	Brand [20]
Large Language Models (LLMs)	Dataset: The initial dataset comprised 14,163 videos obtained using search operators with the key phrase "How to" combined with various AI-related terms. After refinement, 68 videos were selected for qualitative analysis, focusing on Gen-AI usage in content creation on YouTube	Diverse Tool Usage: You Tubers utilise a variety of Gen-AI tools for different tasks, ranging from image processors to text-to-speech tools and podcast editing support Domain Specificity: Content creators apply Gen-AI tools in specific domains like marketing, arts, and education, tailoring the tools to suit the requirements of each domain	Tools Utilized: Various Gen-AI tools are employed for tasks like video processing, text processing, and audio processing, with LLMs being the predominant tool in the analysed videos	Enhanced Creativity: You Tubers leverage Gen-AI tools like DALLE, SeaArt, and Lexica to generate imaginative images and videos, fostering creativity in content creation	Overreliance on AI: There is a risk of over-dependence on Gen-AI tools, potentially diminishing originality and human input in content creation	Lyu et al. [92]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
Generative Pre-trained Transformer (GPT) models	The assessment likely involved using various datasets to train and test the GAI software to evaluate its performance and behaviour in generating content	Content Generation: GAI software can autonomously create content like text, images, and videos based on the underlying algorithms and training data Automation: GAI tools automate tasks such as answering questions, creating content, and analysing data, reducing manual effort and improving efficiency	Bias Detection: Metrics to detect and mitigate biases in the generated content are essential to ensure fairness and prevent discriminatory outcomes	Enhanced Productivity: GAI-driven robots and machines are forecasted to significantly boost human productivity, leading to increased efficiency and output Cost-Effectiveness: GAI-enabled Robotic Process Automation (RPA) offers low operational costs, 24/7 productivity, and scalability, making it a cost-effective solution for various tasks	Ethical Concerns: GAI raises ethical issues related to privacy, intellectual property rights, biases in generated content, and the potential misuse of the technology for malicious purposes like deepfake creation and misinformation spread Risk of Bias: There is a risk of bias in generated content due to the heavy reliance of GAI algorithms on specific types of data or input, potentially leading to skewed outcomes	Beerbaum [13]
Generative Adversarial Networks (GANs)	ImageCFGen on the Morpho-MNIST dataset and the CelebA dataset	The method aims to evaluate and mitigate biases in machine learning classifiers, particularly in image classifiers, by generating counterfactuals that satisfy constraints implied by the causal model	ImageCFGen evaluates bias in classifiers by generating counterfactuals that are comparable in quality to prior work on SCM-based counterfactuals and outperforming in generating high-quality valid counterfactuals	ImageCFGen generates counterfactual examples for images based on a structural causal model (SCM) using Adversarial Learned Inference It can evaluate bias in machine learning classifiers, especially in image classifiers, by generating counterfactuals that adhere to causal relationships between image attributes	ImageCFGen relies on accurate knowledge of the causal graph for counterfactual generation, which can be a limitation	Dash et al. [40]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
BlenderBot	Bias in generated self-chats, conversations between two copies of the same generative model, is primarily used to evaluate model biases and quality in the study	The study focuses on measuring and mitigating biases, specifically gender and race/ethnicity biases, in generative dialogue models by analysing the impact of names associated with specific demographics on conversation content and token usage	multidimensional gender bias classifier to measure gender bias in conversation turns between Speaker A and Speaker B in self-chats generated by de-biased models	Using names to measure bias in generative dialogue models allows for the assessment of biases related to gender and race/ethnicity in conversations De-biasing based on names can have benefits for other linguistic proxies for gender, such as adjectives and nouns, aiding in the evaluation and mitigation of biases in conversational language models	The research primarily focuses on binary gender, which is acknowledged as an incomplete representation of human self-reference, potentially limiting the scope of bias evaluation and debasing efforts	Smith and Williams [141]
Neural Network	Label bias can occur when options for labels in a dataset do not capture the full range of possible labels representative of the target population, leading to biased model development	Multiple Definitions of Bias: Bias in imaging AI encompasses various definitions, including unequal preference based on pre-existing attitudes or beliefs, cognitive bias leading to systematic judgment deviations, and statistical bias resulting in differences between actual and expected values in model predictions	Different metrics like accuracy, F1 score, sensitivity, specificity, and area under the receiver operator curve are crucial for evaluating model performance in medical imaging AI	Mitigating bias in imaging AI can help reduce health disparities by ensuring fair and accurate clinical decisions for all patient populations	Inadequate bias mitigation strategies can result in patient harm due to inaccurate AI predictions, potentially exacerbating existing health inequities	Tejani et al. [146]
Generative Adversarial Network (GAN)	Datasets like FFHQ, datasets directly sampled with GAN, and datasets sampled with the proposed method to compare data distributions and biases	The method is generalisable and can handle more than one attribute at a time, allowing for the synthesis of samples from fine-grained subgroups The approach focuses on shifting the latent distribution based on interpretable semantic dimensions identified in the latent space, aiming to sample a set of latent codes that result in a more fair generated dataset concerning specific attributes	The study uses metrics like the Kullback–Leibler divergence and imbalance measurements to quantify bias in the training data	The method improves the fairness of image generation without the need for retraining, reducing the cost and effort typically associated with retraining models It balances output facial attributes without retraining, enhancing the fairness of generated data while maintaining sample diversity	The method may not eliminate biases introduced during training, as biases can still be present in the generated data, albeit to a lesser extent	Tan et al. [145]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
Recurrent Neural Networks (RNN)	The detectability of machine-generated text heavily depends on the dataset used to train the generator and detector, highlighting the importance of diverse and representative datasets in training detection models	Machine-generated text detection methods need to demonstrate trustworthiness through fairness, robustness, and accountability to ensure they do not cause harm and are reliable in identifying non-human-authored text	Metrics for evaluating machine-generated text detection methods include precision, F1 scores, and recall, which assess the effectiveness of these techniques in identifying machine-generated text	Machine-generated text detection methods play a crucial role in countering the abuse of NLG models by identifying text not authored by humans, thus mitigating potential harm Techniques like deep reinforcement learning (RL) and inverse reinforcement learning (IRL) have been utilised to improve text generation quality and address issues like reward sparsity and mode collapse in GAN-based text generation	Defensive detection systems often lack knowledge of the specific parameters, architecture, and training dataset of NLG models used by attackers, posing challenges in effectively countering machine-generated text	Crothers [38]
Neural Network Model	Datasets used for training neural networks can contain biases, noise, and inconsistencies	Datasets used for training neural networks can contain biases, noise, and inconsistencies Data preparation processes can inadvertently introduce biases, affecting the model's predictions in real-world scenarios	Metrics like the number of switched predictions, mean change, and median change are used to evaluate model performance and bias	Neural networks achieve state-of-the-art performance in various tasks due to the quality and quantity of training data Neural networks can learn from data and improve their performance over time through training	Neural networks can be susceptible to biases in the training data, leading to biased predictions in real-world applications They require a large amount of data for training, which can be a challenge in some domains	

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
Deep learning models- BERT, GPT-2, and ELMol	Datasets used for evaluating biases in deep learning language models may consist of text data with annotations related to gender, race, or other sensitive attributes to assess model performance and biases	Biases in neural network language models can manifest in various forms, impacting the model's predictions and associations based on gender, race, or other attributes	Metrics for bias detection and mitigation in deep learning language models may include quantifying biases, characterising their forms, and evaluating the extent of harm they can cause	Detecting biases in deep learning language models allows for the creation of more fair and ethical AI systems Mitigating biases in these models can lead to improved performance and reduced harm in real-world applications	Biases in AI models can lead to unfair treatment of individuals based on attributes like gender, race, or religion Biased models can result in negative impacts on society, especially in critical areas like healthcare, recruitment, and resource allocation	Garrido-Munoz [55]
AI model- (Midjourney, Stable Diffusion, DALL•E 2)	The study used consistent text prompts for three AI models (Midjourney, Stable Diffusion, DALL•E 2) to generate images for occupations in the ONET database, which contains information on 1,016 occupations	Generative AI models exhibit biases against women and African Americans, with more subtle biases in emotional portrayals and appearances, potentially shaping perceptions unconsciously	The study found systematic gender and racial biases in popular generative AI tools, with biases against women and African Americans being more pronounced than current societal disparities	Generative AI can revolutionise creative content generation in marketing and sales, enhancing customer interactions and potentially adding trillions of dollars in value to the global economy It has the potential to personalise learning experiences in education, improving the learning process, stimulating creativity, and reducing content creation costs	Generative AI models can perpetuate harmful biases, including gender and racial biases, which can intensify societal disparities and shape perceptions in undesirable ways Concerns exist regarding potential risks such as intellectual property rights, the accuracy of output, explaining ability of results, and the propagation of harmful bias, especially in educational settings	Zhou et al. [174]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
Wasserstein Generative Adversarial Network (WGANs)	GANs are used to expand datasets in medical imaging, art, and biological signal fields, improving model training and accuracy	GANs interact adversarial, enabling unsupervised deep learning and learning inherent laws in real-world data	The Wasserstein GAN (WGAN) addresses gradient disappearance and improves model stability by providing useful gradient information to the generative network	GANs do not rely on prior assumptions, unlike traditional generative models, simplifying training and increasing diversity GANs generate real-like samples through forward propagation, providing a simple method for sample generation	GANs face issues like model collapse and uncontrollable training, limiting their effectiveness	Jin et al. [75, 74]
Asynchronous Interactive Generative Adversarial Network model	The model is trained and evaluated on synthetic datasets like Rain12, Rain100L, Rain100H, and Rain800, and real-world rainy images collected from the Internet and previous studies	AI-GAN utilises an asynchronous and interactive two-branch network structure for rain removal, optimising through mutual adversarial mechanisms	The effectiveness of AI-GAN is evaluated through qualitative and quantitative assessments, comparing it with other deraining methods	AI-GAN disentangles contaminated inputs into background and rain latent spaces, promoting interactive generation It outperforms state-of-the-art deraining methods in both qualitative and quantitative evaluation	The model's performance may vary based on the complexity and intensity of rain in the input image, potentially leading to suboptimal results in extreme cases	Jin et al. [75, 74]
The network model in TILGAN consists of a Transformer-based architecture	Experiments are conducted on MSCOCO, WMTNEWS, and ROCSTORY datasets with pre-processing steps akin to prior studies	TILGAN utilises a Transformer-based implicit latent GAN approach with a unique design and distribution matching based on KL divergence	Evaluation metrics include TESTBLEU, SELFBLEU for unconditional generation, and BLEU for conditional generation	TILGAN addresses exposure bias in text generation by combining a Transformer autoencoder and GAN in the latent space, enhancing coherence and diversity TILGAN demonstrates significant improvements in quality-diversity trade-off for both unconditional and conditional text generation tasks	Existing GAN implementations on discrete outputs are often unstable and lack diversity, a challenge that TILGAN aims to overcome	Diao et al. [42]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
Meta Cooperative Training Paradigm Generative Adversarial Network (Meta-CoTGAN)-model	Datasets like SAL COCO Image Captions and WMT News are used for text generation tasks	datasets have specific objectives, such as improving reward mechanisms and reducing sparsity in text generation	Metrics like Negative Log-Likelihood (NLL) and Perplexity (PPL) are commonly used to evaluate the quality of text generation models	Adversarial text generation using Generative Adversarial Networks (GANs) offers alternatives for generating 'natural' language It presents recent approaches for text generation using adversarial-based techniques	GANs were initially designed for continuous data like images, making text generation a challenging task Mode collapsing is a common issue in adversarial-based models, where generators sacrifice diversity for quality	De Rosa and Papa [120]
Neural network model	The research utilised a dataset of over 4 million artworks from more than 50,000 unique users to analyse the impact of text-to-image AI on creative productivity and artwork value	The AI adoption decreases value capture concentration among adopters, promoting a more distributed distribution of favourites	Content Novelty and Visual Novelty are key metrics used to measure the impact of generative AI on artwork creativity	Text-to-image AI enhances human creative productivity by 25% and increases the value of artworks by 50% AI-assisted artists exploring novel ideas may produce artworks more favourably evaluated by peers	Average Content Novelty declines over time among adopters, suggesting an expanding but inefficient idea space	Eric and Lee [45]
Generative AI models	The paper discusses the training methods used to produce artificially generated content but does not specify a particular dataset	The potential to generate highly complex decision trees by adjusting weights in neural networks	Generative AI models are trained using unsupervised learning techniques like clustering, association, and dimensionality reduction	Generative AI, such as ChatGPT and Dall-E, have broken records for early public adoption and capital investment, showing potential for disrupting industries and culture Deep learning models in generative AI can introduce new structures with appropriate properties, aiding in tasks like de novo drug design	Lack of transparency in ML models can lead to severe consequences like incorrect predictions, biased decisions, and poor resource allocation in various domains	Garon [54]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
Chat Generative Pre-Trained Transformer (ChatGPT)	The dataset used to train the Chat Generative Pre-Trained Transformer (ChatGPT) and similar Large Language Models (LLMs) includes a wide range of data such as articles, books, and internet sources to understand and generate human-like responses	ChatGPT is an artificial intelligence model that generates text by answering questions and follow-up questions in a conversational, human-like manner OpenAI, the developer of ChatGPT, acknowledges the model's limitations, stating that it can sometimes produce incorrect or nonsensical answers, known as "hallucinations	Metrics and characteristics of GAI systems are not explicitly discussed in the provided contexts	Generative Artificial Intelligence (GAI) like ChatGPT allows lawyers to access vast repositories of legal knowledge, draft documents, conduct legal research, and communicate with clients efficiently Enhances decision-making by providing lawyers with current information	GAI systems can introduce biases if trained on biased data, potentially leading to harmful outputs reflecting human biases Misuse of GAI tools can lead to violations of Model Rules of Professional Conduct, especially in terms of competence and candour, potentially resulting in breaches during court proceedings	Schworer [127]
Generative artificial intelligence model	Technologies like ChatGPT, Google Bard, and Claude, as well as medically fine-tuned models such as Med-PaLM and ChatDoctor, highlight the potential of AI in healthcare applications	The rapid development of healthcare-focused tools powered by GAI models is advancing, with technologies like GPT, PaLM, and LLaMA gaining popularity for their potential in clinical use AI development in healthcare involves hundreds to thousands of initiatives in early-stage startups and nationally-backed research globally, showcasing the widespread interest and investment in leveraging AI for medical applications	AI models require data availability, quality, and validation frameworks for local contextualisation, emphasising the importance of refining models to prevent health inequities	Capacity-building efforts in AI can help bridge the gap in healthcare disparities by preparing clinicians and researchers for the integration of AI technologies	AI models heavily rely on data quality and volume, which can lead to biases and misrepresentations, especially in regions with lower medical research output like the Philippines	Gutierre and Viacrusis [62]
GAI models	Datasets used for training GAI models should be unbiased to avoid perpetuating biases in the models and their outputs	GAI models should be developed responsibly, following ethical AI practices and principles to ensure safe and effective use	Organizations should monitor changes in the regulatory environment as GAI technology evolves and becomes more widely used	GAI enables businesses to create new experiences by merging virtual and physical worlds, enhancing customer experiences and value propositions It revolutionises product design by streamlining processes, improving efficiency, and creating more effective products	GAI models can be biased due to societal and internet biases, potentially enabling unethical behaviour or criminal activities	Mondal et al. [106]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/Future Work	References
Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Recurrent Neural Networks (RNNs	dataset used in generative AI for creative computing	Generative AI platforms allow novices and experts to explore their creative potential and contribute to artistic and design knowledge collectively	Evaluation metrics are crucial for assessing the performance of generative models like GANs, VAEs, and autoregressive models in generating realistic and diverse outputs	Generative AI enables the generation of diverse and novel content in various domains, fostering creativity and innovation It democratises access to creative tools, inspiring new forms of artistic expression and inclusive creativity	Ethical considerations such as algorithmic bias, data privacy, and societal impacts of AI-generated content need to be addressed	Shah [129]
Generative AI model ChatGPT	ChatGPT is trained on diverse and large datasets using advanced deep learning techniques like transformer models, improving its ability to generate coherent and contextually appropriate responses	Generative AI models like ChatGPT exhibit the characteristic of generating diverse types of content such as text, code, audio, images, and videos based on the input provided These models can adapt to user language patterns over time, improving the quality and relevance of their generated outputs	Generative AI like ChatGPT is built on transformer technology, enabling it to create predictions based on inputs and generate various content types such as text, code, audio, images, and videos	Generative AI, such as ChatGPT, can serve as a new context for management theories, influencing decision-making, knowledge management, customer service, human resource management, and administrative tasks Generative AI can automate interactions with customers, revolutionising service delivery and customer-organization relations	The issue of augmentation and automation may lead to the replacement of humans by machines in specific tasks, requiring further scholarly exploration	Korzynski et al. [82]
Generative Artificial Intelligence	Two large-scale real-world datasets, the Yahoo search engine dataset and the Alibaba recommender system dataset, are utilized for experiments	DRSR leverages recurrent neural networks to model contextual information and estimate user feedback likelihood It employs survival analysis techniques to recover unbiased joint probabilities of user behaviours The model is designed to handle non-click queries and untrusted observations effectively	Evaluation measures used include NDCG at positions 1, 3, 5, and MAP for relevance ranking	DRSR addresses position bias in information retrieval by considering contextual information and user behaviours The model incorporates survival analysis techniques to recover unbiased joint probabilities of user behaviours	The model may have limitations in scenarios where the contextual information is sparse or noisy	Peres et al. [116]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
Gen AI model	the paper does not delve into the specifics of a dataset used for training or testing genAI models	The dataset used for genAI models should be grounded on reliable sources to ensure content veracity and trustworthiness Transparency in displaying AI involvement, background information, and sources is necessary to enhance trust and credibility	Metrics include establishing governance for reliable database sources, ensuring neutrality, transparency, and explainability of the solution	GenAI can offer extraordinary outputs that enhance trustworthiness, potentially benefiting end-users The technology can generate innovative and diverse content, expanding the possibilities of text creation	GenAI poses the risk of generating misinformation that may not be easily recognised, leading to potential harm The reliance on genAI outputs without proper verification can result in the dissemination of inaccurate information	Tomitza [151]
Graph Neural Networks (GNNs)	Generative AI tools leverage datasets from the Internet, scientific articles, and crowdsourcing approaches, which may introduce limitations in data quality due to subjective biases and uneven value distributions	Characteristics of generative AI applications involve multiple outcomes, exploration and control, mental models, explanations, and considerations for potential harms and displacements	Metrics for evaluating generative AI models include accuracy, fluency, diversity, and coherence in generated outputs	Generative AI tools like Graph Neural Networks (GNNs) and Variational Autoencoders (VAEs) combine traditional planning algorithms with deep learning, offering solutions for complex problems AI-NLPs are proficient in interpreting natural language, enabling efficient conversational interactions	Search engines are vulnerable to manipulation and lack complex abilities, while AI-NLPs may provide incorrect or biased responses due to complex data training	Iorliam and Ingio [70]
Natural language generation model	The research paper discusses training and evaluating NLG models on the EMNLP News 2017 dataset	MLE models trained with temperature tuning can outperform GANs on quality metric GAN variants may struggle with maintaining sample diversity while improving quality, unlike MLE models that offer a better quality-diversity tradeoff	Quality and diversity are essential metrics for evaluating natural language generation models	Maximum Like hood extension(MLE) models are easier to train, cross-validate, and less computationally expensive compared to GAN variants MLE models consistently outperform GAN variants in terms of quality-diversity tradeoff	GAN variants can suffer from complications due to non-differentiable, sequential training, making them harder to train effectively	Caccia et al. [26]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
Generative adversarial model	The research assesses the bias in different GAN-based DA settings and designs a pipeline to evaluate the efficiency of GAN-based DA on specific datasets	The study explores bias in GAN-based DA for small samples, focusing on the viability and potential biases of using GAN-generated data for augmentation	The study conducted experiments using different GAN variants, including softmax GAN, conditional GAN, and boundary-seeking GAN, on the SCADI dataset	DA methods like rotation, flip, and noise addition enhance datasets by adding instances that the model may encounter in the real world but are absent in the original dataset Advanced methods like conditional GAN can generate diverse data from the same distribution, aiming to maintain data diversity within the original manifold	GAN-based DA can introduce bias in the generated data, impacting model performance	Hu and Li [66]
Generative artificial intelligence (GenAI)	The dataset primarily consists of documents produced by high research activity institutions to understand the guidance given to institutional stakeholders regarding GenAI	It is noted for its potential to leverage content, identify bias, and stimulate critical thinking in educational settings	GenAI guidance focuses on writing activities more than code and STEM-related activities, with a significant emphasis on ethics, including Diversity, Equity, and Inclusion	GenAI can enhance teaching and learning by providing sample syllabi, curriculum, and activities for instructors to integrate into the classroom It can promote higher-order thinking skills and critical thinking through activities like critiquing, comparing text, and identifying bias	Policies and guidelines for GenAI can be burdensome for faculty as they may require extensive revision of pedagogical approaches	McDonald et al. [99]
Duration-Deconfounded Quantile(D2Q Model	Utilizes data across all duration groups for training the watch-time quantile prediction model	Utilizes a causal graph to illustrate duration as a confounding factor affecting video exposure and watch-time prediction Integrates watch-time prediction into the ranking phase of an online recommender system	MAE (Mean Absolute Error) for measuring prediction accuracy	Addresses duration bias in watch-time prediction for video recommendation	Performance drops when the group number is too large due to reduced group-wise sample size and accumulated estimation error	Ruohan et al. [122]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
GAI Model	Generative AI models draw from expansive datasets and intricate algorithms to align with grammatical norms accurately	Generative AI in language learning emphasises precision, authenticity, and cultural richness while addressing ethical concerns and limitations	The precision and genuine representation of human ideas and experiences are crucial in AI-mediated language generation	Generative AI offers customised learning journeys, dynamic content, and individualised feedback mechanisms, enhancing language acquisition Learners receive instantaneous feedback, promoting ongoing learning cycles and improving retention and proficiency in second language programs	Generative AI may standardise language expression, propagate limited cultural narratives, and diminish analytical thought and inventiveness	Creely [37]
Generative Adversarial Networks (GANs)	In generative AI, datasets can be domain-specific text corpora or reference images used for training AI models to tailor their outputs	Generative AI outputs can be fine-tuned based on specific input directives or training data, allowing for customisation and personalisation	The taxonomy of GAI applications classifies them based on ten dimensions, ensuring a comprehensive evaluation and classification process	Generative AI offers a wide range of output types, including text, images, videos, 3D models, and sound, showcasing its versatility GAI applications can assist in various domains like programming, sales, and accounting, showcasing its broad applicability	The interpretation of data collection and dimension derivation is subjective, leading to the possibility of divergent characteristics identified by different researchers	Strobel et al. [143]
Vector Quantized Generative Adversarial Network. -CLIP model	VQGAN-CLIP for image creation was used	Characteristics of text-to-image generation systems involve the iterative nature of the process, where images can be used as initial inputs to direct scene composition or manipulate existing images	Metrics for evaluating text-to-image generation include the quality of the generated image and understanding the complete set of prompt modifiers	Text-to-image generation systems allow for the creation of digital images from textual prompts, enabling individuals with little technical expertise to produce art These systems offer a user-friendly interface where users can input natural language prompts to generate images	Evaluating the creativity of text-to-image art poses challenges due to information asymmetries between creators and viewers, such as system opacity and prompt complexities	Oppenlaende [111]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
Large Language Models (LLMs)	The paper does not explicitly mention the specific dataset used for training the AI model	The AI model prioritises agent responses expressing empathy, providing technical documentation, and avoiding unprofessional language	Productivity is measured by resolutions per hour, showing a significant increase with AI model deployment	AI model deployment increases resolutions per hour by 0.47 chats, improving productivity by 22.2% AI models can capture the skills that distinguish high-performing workers, potentially enhancing overall performance	The reliance on large language models (LLMs) for text generation raises concerns about unprofessional language and lack of empathy, which require additional training to mitigate	Brynjolfsson, and Raymond [22]
Generative Adversarial Networks (GANs)	GAI utilises datasets that evolve, requiring model retraining to adapt to changes in data distribution and structure	GAI techniques like Geometric Deep Learning (GDL) aim to interpret AI models using geometric principles, enhancing model constraints and success	GAI efficiency metrics include image quality, resolution, inception score, and training time reduction	Generative AI (GAI) leverages generative modelling and deep learning to create diverse content at scale, including text, graphics, audio, and video GAI techniques like Generative Adversarial Networks (GANs) and Generative Pre-trained Transformer (GPT) models enable the generation of realistic synthetic artifacts	Evaluating GAI outputs can be subjective and challenging due to factors like utility, aesthetics, clarity, and similarity to real-world content	Jovanovic and Campbell [76]
Deep Recurrent Survival Ranking (DRSR) Model	The DRSR model is extensively evaluated on two large-scale industrial datasets, namely the Yahoo search engine and Alibaba recommender system datasets	The model's characteristics include the ability to capture user behaviours, train unbiased ranks with contextual information, and mine hidden user patterns in non-click queries using a cascade model and survival analysis	The DRSR model is evaluated using various metrics such as unbiased learning-to-rank performance, necessity of debased methods, and robust learning under different scenarios	DRSR addresses position bias in information retrieval by considering contextual information and hidden user behaviours The model can be easily integrated with both point-wise and pair-wise learning objectives, enhancing its flexibility and applicability	The limitations or drawbacks of the DRSR model are not explicitly discussed in the provided contexts	Jin et al. [75, 74]

Table 2 (continued)

Model/Network	Data	Characteristic/Properties	Metrics/Results	Advantage	Disadvantage/ Future Work	References
ChatGPT-4 model	The research paper does not explicitly mention the specific dataset or network model used in the study. The focus is on evaluating human-AI collaboration paradigms and the perception of content quality generated under these paradigms	The study analysed satisfaction levels, willingness to pay, interest in products, and persuasion levels as metrics to evaluate the quality and impact of the content generated	Participants rated the quality of advertising and campaign content based on the paradigm under which it was generated, with additional information provided in the informed condition about the content's origin	Content generated by generative AI and augmented AI is perceived as higher quality than that produced by human experts, showcasing the potential for AI to create high-quality content Participants in the study were either equally satisfied or even more satisfied with content generated with AI's involvement, willing to pay the equivalent or more for it, and becoming equally or more interested in the product or persuaded to support the campaign when AI-made final decision on the output	Human favouritism plays a significant role in biasing evaluations, with content created by human experts being perceived as higher quality than AI-generated content, even when the source of content production is revealed	Zhang and Gosline [170]
Advanced artificial intelligence	Generative AI algorithms can be trained on large financial datasets to learn patterns and relationships between financial metrics, aiding in financial statement analysis and decision-making	focuses on improving productivity, efficiency, and decision-making in accounting by mimicking human interactions and providing direct access to systems using APIs	Generative AI leverages the Generative Pertained Transformer (GPT) technology, enabling the automation of tasks and the generation of data in accounting processes	Generative AI, such as Chat GPT, can automate tasks like generating accounting reports and aiding in regulatory compliance, enhancing productivity and efficiency in accounting processes GAI-driven RPA offers 24/7 productivity, low operational costs, scalability, and eternal lifetime, surpassing human execution capabilities	Ethical and legal implications must be carefully considered, especially regarding privacy, data security, and potential biases in generated content	Beerbaum [13]

Fig. 18 Key approaches for mitigating bias in generative AI systems

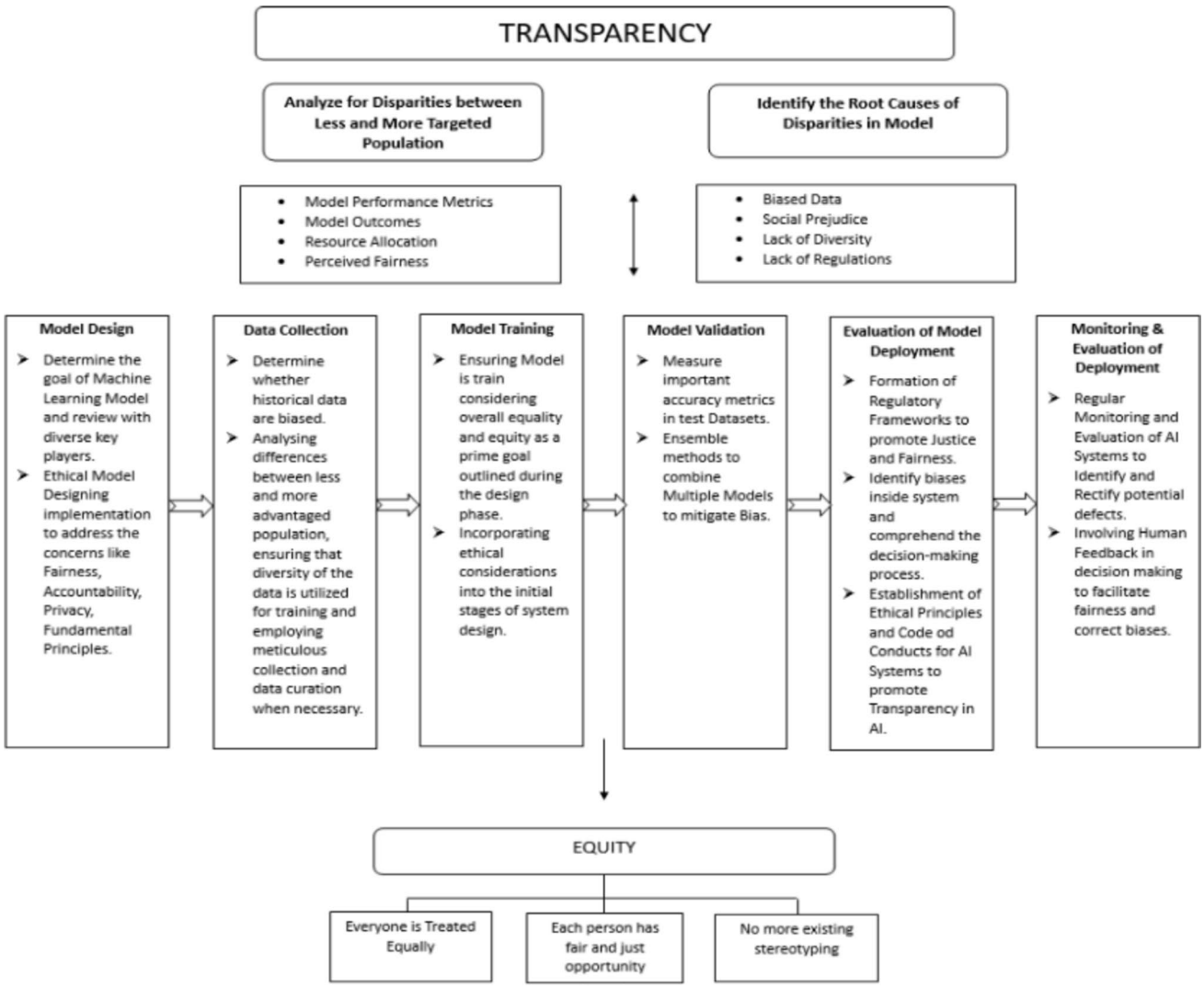

Fig. 19 Framework for ensuring fairness and mitigating bias in AI systems

4.4.1.2 Data collection and curation Our analysis highlights the critical importance of **data collection and curation** in ensuring fairness and mitigating bias in AI systems. The data used for training AI models is often the source of bias, which can significantly affect the model's fairness and equity. Ensuring the representativeness and diversity of datasets is paramount for developing unbiased AI systems. This can be achieved through careful data collection processes and by employing **data augmentation techniques** to improve the variety and comprehensiveness of training data.

Determining, monitoring, and addressing bias in data
The first step toward fairness in AI involves meticulously examining historical data for possible biases. A thorough **bias assessment** requires the evaluation of demographic factors such as gender, ethnicity, age, and socioeconomic status within the dataset. This process enables the identification of any disproportionate representation of specific groups, allowing us to uncover potential historical biases that could be ingrained in the data [108; 19].

Once biases are identified, it is crucial to assess how they impact the AI model's predictions and outcomes. If not addressed, historical data imbalances can lead to discriminatory outcomes, reinforcing inequalities rather than reducing them. By understanding the historical context of the data—such as the societal or cultural biases that may have influenced the collection and selection of the data—we gain insights into systemic issues that may affect model performance [161].

This comprehensive evaluation of data biases provides the foundation for subsequent steps to mitigate these biases. Ensuring the data is representative and free from historical prejudices is essential to fostering a more inclusive, fair, and equitable AI system. **Monitoring and regularly updating** the datasets is critical for maintaining fairness throughout the AI system's life cycle. Therefore, addressing biases in the data must be an ongoing process, with constant vigilance and improvement, to create AI systems that can positively impact society without perpetuating harmful biases. This data curation and bias identification approach is critical in ensuring AI models operate fairly, fostering trust, inclusivity, and greater societal acceptance of AI technologies.

4.4.1.3 Enhancing diversity and inclusion in artificial intelligence Our analysis underscores the significant role that a diverse workforce plays in mitigating bias within AI systems. Integrating diverse perspectives into designing, developing, and deploying AI models enhances fairness and inclusivity. The presence of individuals from various demographic backgrounds, whether in terms of gender, ethnicity, socioeconomic status, or experience—broadens the scope of solutions and ensures that AI systems are more representative of society. A diverse team is instrumental in recognising and addressing potential biases that may not be apparent to a homogenous group, thus fostering a more equitable technological landscape.

Organisations can employ several strategies to promote equitable and unbiased development. These include targeted recruitment efforts to attract a wide range of talent, comprehensive training and development programs to ensure all personnel understand the importance of diversity and inclusion and implementing conduct guidelines that prioritise inclusive practices and ethical standards [142]. Such strategies improve the moral alignment of AI models and contribute to better overall performance by ensuring that various viewpoints are considered in decision-making processes.

4.4.1.4 Determining the goal of a machine learning model and review with diverse key players
Establishing a clear and well-defined goal for an AI system is fundamental to its ethical deployment and effectiveness. Our analysis emphasises that this process must involve engaging with a broad spectrum of key stakeholders, each bringing unique perspectives. These stakeholders typically include data scientists, developers, subject matter experts, policymakers, and end-users.

- Data scientists and developers are critical in transforming high-level business objectives into technical specifications and model algorithms. Their expertise is essential in selecting appropriate methods for data analysis and fine-tuning models to meet the stated goals.
- Policymakers and organisational decision-makers provide valuable insights into the legal and ethical frameworks that govern AI use, ensuring the model adheres to applicable laws and ethical guidelines.
- End-users are particularly crucial, as their feedback on the practical use of AI systems offers valuable real-world insights. Understanding how AI models impact end-users allows developers to refine their approach, making the technology more user-centric and accessible.

Incorporating input from such a diverse group of stakeholders ensures that AI systems align with organisational goals and consider their deployment's societal impact and ethical implications. This collaborative approach is essential for developing responsible AI technologies that are fair, transparent, and equitable ([108; 161].

Organisations can build trust with their users and stakeholders through inclusive practices, which is critical to AI systems' adoption and long-term success. This comprehensive approach to stakeholder engagement is crucial in ensuring that AI technologies reflect the values of diversity

and inclusion, mitigating potential biases and promoting fairness in their outcomes. By ensuring that the development of AI systems is conducted through a lens of diversity and inclusion, organisations can reduce prejudice and build AI models that are more adaptable, reliable, and capable of meeting the needs of a broader range of users.

4.4.1.5 Routine monitoring and evaluation of AI systems Our analysis emphasises the critical importance of **routine monitoring and evaluation** to ensure AI systems' ongoing performance, fairness, and ethical integrity. Continuous oversight is essential in identifying potential defects, biases, or unintended consequences that may arise during the operation of AI models. Regular monitoring helps prevent minor issues from escalating into significant ethical concerns, ensuring that AI systems remain aligned with their intended goals and societal standards.

Regular monitoring to detect bias and performance issues Routine monitoring involves implementing a **robust evaluation system** that tracks key performance indicators (KPIs) related to the AI system's objectives. Systematically assessing the model's outputs against predefined benchmarks makes it possible to detect any deviations, anomalies, or biases that may emerge [142]. These evaluations are not limited to technical performance metrics,they also focus on the **ethical implications** of the AI's functioning. For instance, monitoring can reveal whether the model's decisions unfairly impact certain demographic groups or whether emerging biases in the data influence outcomes in discriminatory ways.

Moreover, the **continuous evaluation** of AI models ensures they remain compliant with evolving **societal norms, legal frameworks**, and technological advancements [19]. As societal expectations change and new regulations are introduced, AI systems must be updated to reflect these shifts, thus ensuring their ethical compliance and social relevance.

Proactive identification and rectification of ethical concerns Our analysis highlights that regular monitoring is an **early warning system** for potential ethical issues. For example, biases not apparent during initial development may surface as the system operates in real-world environments. Early detection of such issues allows for timely intervention, ensuring corrective measures are implemented before the negative consequences become widespread [108]. By iteratively refining models based on insights from continuous evaluation, developers can address performance and fairness concerns, thereby enhancing the reliability and responsibility of AI technologies.

Routine evaluation also enables the integration of **feedback loops** that account for the dynamic nature of the environment in which AI systems operate. This iterative approach helps AI models to adapt to new data, societal changes, and regulatory requirements. Furthermore, by engaging in regular monitoring, AI developers can foster greater **accountability** and **transparency**, ensuring that AI systems uphold the ethical standards set forth by the organisation and society [161].

In conclusion, our analysis stresses that regular monitoring and evaluation are necessary for detecting and correcting defects and maintaining AI systems' ethical and societal alignment. By continuously assessing and refining AI models, developers can mitigate biases, enhance fairness, and ensure the responsible deployment of AI technologies.

4.5 Ethical model designing in AI

Our analysis emphasises that **ethical model design** is critical in mitigating biases and promoting fairness in developing AI systems. Ethical design entails intentionally incorporating ethical principles—such as fairness, accountability, transparency, and privacy— into the foundational stages of AI system development. By embedding these principles early in the design process, it is possible to create AI systems that exhibit greater resilience to ethical pitfalls, fostering equitable outcomes and minimising harm (Tables 1 and 2).

4.5.1 Implementation of ethical model design

Implementing **ethical model design** is a multifaceted strategy to address critical concerns such as fairness, accountability, privacy, and fundamental human rights. These concerns are crucial for ensuring that AI systems function in ways that respect societal values and ethical guidelines.

1. *Fairness*: To ensure fairness, including **diverse and representative data** during the training phase is vital. This involves carefully selecting datasets encompassing a broad spectrum of demographic variables, including gender, race, age, and socioeconomic status. Additionally, algorithms should be designed or chosen to be **inherently fair**, avoiding mechanisms that could lead to discriminatory outcomes. Fairness considerations are critical to prevent AI systems from perpetuating or exacerbating existing societal inequalities [142].
2. *Accountability*: Achieving accountability in AI systems involves the creation of **transparent algorithms** that allow stakeholders to understand how decisions are made. Transparency builds trust by providing insight into the model's decision- making processes and

ensuring that biases or unfairness can be identified and rectified [161]. Accountability also includes establishing clear lines of responsibility for AI-generated decisions and ensuring that developers, organisations, and other stakeholders are held accountable for the consequences of AI deployment.

3. *Privacy*: Protecting **privacy** is paramount in AI development. Ethical model design must incorporate data protection measures such as **minimisation** and privacy- preserving techniques like **federated learning** or **differential privacy**. These approaches ensure that personal data is kept secure and used only for legitimate purposes, thereby safeguarding individuals' privacy while maintaining the effectiveness of the AI system.
4. *Human-Centric Design*: Ethical AI systems must be designed with a **human-centric approach**, ensuring that the systems prioritise human rights, dignity, and well-being. This requires continuous monitoring and evaluation to assess the ethical impact of AI outputs and ensure alignment with fundamental principles of justice, fairness, and respect [108]. The design should achieve its functional goals and align with broader ethical values that uphold societal norms.

4.5.2 Integration of ethical considerations into AI development

Integrating ethical considerations into AI's design and development phases is essential for creating trustworthy, accountable, and fair systems. By adopting ethical frameworks and guidelines from the outset, AI developers can anticipate and address potential risks before they manifest. This approach enhances the fairness and accountability of AI systems, builds public trust, and fosters the responsible use of AI technologies.

Our analysis suggests that an ethical framework that includes **fairness**, **accountability**, and **privacy** principles is essential for the long-term success of AI systems. This framework ensures that AI technologies contribute positively to society, reducing the risk of bias and harm while promoting inclusivity and equity.

4.6 Regulatory frameworks for AI: promoting justice and mitigating bias

Our analysis underscores the CRITICAL role that regulatory frameworks play in ensuring the ethical and equitable development and deployment of artificial intelligence (AI) systems. Regulatory mechanisms establish accountability standards, provide oversight, and ensure that AI technologies operate in a manner that aligns with ethical norms and societal expectations. Establishing such frameworks is vital for mitigating potential biases inherent in AI systems and safeguarding against unjust or discriminatory outcomes that may arise from unchecked AI deployment [12, 108].

4.6.1 The role of regulatory frameworks in AI

Regulatory frameworks for AI systems are essential in promoting justice and accountability. These legal structures can impose requirements on AI developers to adhere to fairness principles, ensuring that the AI systems' decisions reflect a broad range of societal values. By providing clear guidelines on transparency, data privacy, and fairness, regulatory frameworks contribute significantly to preventing biases that may inadvertently emerge through historical data or flawed model design [142]. Moreover, these frameworks can hold organisations accountable for the consequences of their AI systems, thereby ensuring that deploying AI technologies does not lead to discriminatory or harmful impacts on vulnerable populations.

4.6.2 Analysing differences between populations: ensuring inclusivity and fairness

Engaging in a nuanced analysis of the differences between less and more advantaged populations is crucial to mitigate biases in AI systems. AI systems often reflect the inequities in historical datasets, which may lead to discriminatory outcomes if not carefully addressed. Understanding how various demographic groups are represented within the training data is essential to identify disparities and biases. This analysis enables the development of more inclusive and fair AI systems that accurately reflect the diversity of real-world populations (Xivuri and Twinomurinzi [161]; Camilla [118]).

4.6.3 Continuous monitoring and evaluation for ethical alignment

Continuous monitoring and evaluation ensure that AI systems evolve, aligning with ethical principles and societal values. Through iterative assessments, AI developers can identify and rectify any emerging biases or unintended consequences of model predictions. This ongoing evaluation process ensures that AI systems remain accountable and transparent and continue to reflect society's dynamic needs and ethical considerations [142].

In conclusion, when effectively implemented, regulatory frameworks can guide the ethical development of AI systems, promoting fairness and mitigating biases. AI developers can ensure these systems evolve to align with societal values and uphold justice by focusing on inclusive data practices, diverse training datasets, and continuous evaluation.

4.7 Human-in-the-loop approaches

In our analysis, the **human-in-the-loop** (HITL) approach emerges as a pivotal strategy in developing ethical AI systems. This approach emphasises the critical role of **human feedback** in identifying, rectifying, and mitigating biases that may not be immediately apparent through algorithmic processes alone. By incorporating human intervention at various stages of the decision-making process, HITL methodologies create a feedback loop that strengthens the fairness, transparency, and accountability of AI systems.

4.7.1 Involving human feedback for bias mitigation and ethical decision making

Integrating **human feedback** is crucial for addressing AI models' inherent limitations and potential biases. While powerful, AI systems often lack the nuanced understanding that human experts can provide, particularly when identifying subtle biases that might escape automated detection. Engaging human experts, end-users, or affected communities allows for the identification of unintended biases and ethical concerns that might remain undetected [104, 142].

This approach becomes essential when AI systems are deployed in complex, real-world environments where social and cultural contexts significantly influence decision-making processes. Feedback from diverse human perspectives helps to uncover the **real-world impact** of AI-driven decisions, ensuring that they align with societal values and ethical standards. The iterative feedback loop, where AI systems refine their predictions based on human input, not only helps correct biases but also **enhances the overall fairness** of the system [108].

4.7.2 Iterative improvement and accountability in AI systems

The HITL approach facilitates **continuous learning and improvement** of AI systems. As human feedback is integrated into the AI model, it allows for ongoing adjustments, improving the system's accuracy and fairness over time. Furthermore, this human-centred approach ensures that AI systems remain aligned with evolving **ethical guidelines** and **social values**.

Moreover, human feedback's transparency plays a critical role in fostering accountability. By maintaining an open dialogue between AI systems and human stakeholders, developers can address ethical issues early in the deployment phase and continually refine the model to reflect technical accuracy and societal expectations. This collaborative effort enhances trust in AI technologies, ensuring that they serve the interests of all stakeholders while upholding fundamental principles of fairness and justice [142].

4.8 Ethical principles and code of conduct

Establishing ethical principles and codes of conduct for artificial intelligence (AI) systems plays a critical role in advancing justice and mitigating prejudice in AI development. By defining and implementing standards for AI design and use, ethical guidelines can provide a framework that ensures AI systems are developed with fairness, transparency, and accountability at their core. These standards also help create methodologies that promote responsible AI deployment, fostering trust within the AI community and broader society [34, 57, 154, 161]. Introducing these ethical codes is crucial for guiding AI development towards equitable outcomes while ensuring that AI systems are aligned with societal values and human rights. These frameworks encourage AI practitioners to prioritise ethical considerations in their work, ultimately contributing to creating effective and just AI systems.

4.9 Quality assurance and fairness check

Quality assurance in model-centric and data-centric approaches is crucial for ensuring fairness and minimising biases in AI systems. In the model-centric approach, quality assurance focuses on refining the model—testing algorithms, optimising hyperparameters, and conducting rigorous evaluations to identify and rectify errors that may impact fairness [142]. On the other hand, the data-centric approach emphasises the quality and representativeness of the training data (Baker 2024). Here, quality assurance involves thorough data curation, detection of bias in datasets, and implementing strategies like data augmentation to ensure diversity [161; 108]. Both approaches require fairness checks, such as bias detection methods and audits of model outcomes, to assess whether AI systems disproportionately affect certain demographic groups. Fairness checks can include statistical tests like disparate impact analysis or fairness-aware learning algorithms designed to rectify inequities by modifying the data and the model [165, 169]. Regular monitoring and recalibration of the models and the datasets are essential to maintain AI systems' ethical integrity and fairness over time as these systems evolve and are deployed in real-world environments.

The analysis addresses several challenges in effectively mitigating bias in AI systems, emphasising the need for strategies that ensure diverse representation, transparency, fairness, and ethical accountability. The limited diversity within training data is a significant issue, as datasets often fail to represent various demographics fully. Collecting

more comprehensive data can be difficult, especially for sensitive or rare cases, and may raise privacy concerns, impacting the success of strategies like data augmentation.

Identifying and quantifying different types of bias within AI systems also presents challenges, as bias can stem from multiple sources—including the data, algorithms, or even end-users. This complexity makes it harder to isolate and measure bias, limiting the efficacy of bias- aware algorithms and feedback mechanisms. Another core challenge is the balance between fairness and accuracy. While prioritising fairness is essential, it can sometimes reduce model accuracy for specific groups, underscoring a trade-off that requires nuanced decision- making. Additionally, ethical complexities arise when deciding which biases to address first and which groups to prioritise, especially historically marginalised communities.

Addressing these biases is critical for establishing AI systems that are both fair and equitable. Continuous research and development of bias-mitigation strategies are necessary to build inclusive and responsible AI systems that serve individuals and society justly.

5 Conclusion

In conclusion, achieving effective bias mitigation in generative AI (GAI) systems requires a comprehensive, multidimensional approach to ensure fairness, inclusivity, and accountability across varied modalities, including text, audio, video, image, and code. This study underscores the foundational role of diverse, representative datasets and the importance of rigorous data processing techniques to address biases at the early stages of development. Fairness metrics play a crucial role in evaluating GAI performance across demographic groups. At the same time, real-time bias detection ensures that biases are identified and mitigated as they arise, creating transparent and accountable systems.

Key strategies, such as algorithmic refinements, debiasing algorithms, and adversarial training, have shown promise in reducing bias effectively. Building interdisciplinary teams that include developers, researchers, and stakeholders further strengthens the identification and handling of complex bias issues. Moreover, ethical guidelines and governance frameworks play a critical role in guiding the development and deployment of GAI, promoting transparency and accountability within the field.

Despite progress, challenges remain considerable. The difficulty of securing representative data, the complexities in detecting biases within sophisticated algorithms, and the tension between fairness and accuracy highlight the need for innovative and transparent approaches. Ethical dilemmas arise when deciding which biases to address and how to prioritise impacted populations, necessitating a clear ethical and technical framework.

Future advancements in bias mitigation require continued research, cross-disciplinary collaboration, and regulatory oversight. Tailored improvements in techniques and algorithms can significantly enhance bias mitigation in various domains.

Enhancing fairness in generative AI across text, audio, video, image, and code generation requires a diverse, adaptive approach tailored to each modality. In text generation, increasing linguistic diversity within training datasets, refining algorithms to recognise and filter toxic language, and ensuring inclusive representation across multilingual datasets can help reduce biases. Employing adaptive feedback loops and refining debiasing techniques enables language models to understand contextual nuances better, reducing biased responses. Audio processing benefits from diverse audio datasets that include varied accents, languages, and tonal nuances, which improves model fairness. Developing bias- aware acoustic models and incorporating fairness metrics in voice recognition applications can ensure a more equitable representation of underrepresented voices in audio AI.

For video and image generation, bias reduction focuses on diversifying datasets to include a balanced representation of ethnicity, age, and cultural backgrounds. Techniques such as targeted data augmentation and adversarial training help minimise visual biases, while enhanced fairness metrics allow image classification models to maintain accuracy without sacrificing equity. In code generation, bias reduction requires a thorough analysis of the code corpus used for training to avoid reinforcing biased or unfair coding practices. Training models on inclusive codebases and integrating ethics-focused standards into programming practices support responsible code generation. These tailored strategies can strengthen fairness across generative AI domains, ensuring that models produce inclusive and responsible outputs.

As generative AI advances, a balanced approach integrating human oversight and robust data governance will be vital to achieving equitable, responsible AI. Generative AI's potential to transform sectors like content creation, healthcare, and education is vast, yet careful, ethical use is essential. Continuous development, proactive bias monitoring, and ethical responsibility will enable GAI to positively impact society across industries while safeguarding against unintended consequences.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

References

1. Abdullah, T., & Rangarajan, L. (2021). Image-text matching: Methods and challenges. In *Proceedings of the International Conference on Inventive Systems and Control (ICISC 2021)* (pp. 213–222). Springer. https://doi.org/10.1007/978-3-030-64547-1_21
2. Afzal, M., Li, R.Y.M., Ayyub, M.F., Shoaib, M., Bilal, M.: Towards BIM-based sustainable structural design optimization: a systematic review and industry perspective. Sustainability **15**(20), 15117 (2023). https://doi.org/10.3390/su152015117
3. Ahmad, A., Ilyas, M., Howard, M., & Howard, I. (2012). A framework for the adoption of rapid prototyping for SMEs: from strategic to operational. *International Journal of Industrial Engineering*, *19*(3).
4. Anastassiou, P., Tang, Z., Peng, K., Jia, D., Li, J., Tu, M., Wang, Y., Wang, Y., & Ma, M. (2024). VoiceShop: a unified speech-to-speech framework for identity-preserving zero- shot voice editing. *arXiv preprint* arXiv:2404.06674. https://arxiv.org/abs/2404.06674
5. Aylett, M.P., Vinciarelli, A., Wester, M.: Speech synthesis for the generation of artificial personality. IEEE Trans. Affect. Comput. **11**(2), 361–372 (2017). https://doi.org/10.1109/TAFFC.2017.2763130
6. Bahroun, Z., Anane, C., Ahmed, V., Zacca, A.: Transforming education: a comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis. Sustainability **15**(17), 12983 (2023). https://doi.org/10.3390/su151712983
7. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint* arXiv:2204.05862. https://arxiv.org/abs/2204.05862
8. Bakpayev, M., Baek, T.H., van Esch, P., Yoon, S.: Programmatic creative: AI can think but it cannot feel. Australas. Mark. J. **30**(1), 90–95 (2020). https://doi.org/10.1016/j.ausmj.2020.04.002(Originalworkpublished2022)
9. Banh, L., & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets, 33*, 63. https://doi.org/10.1007/s12525-023-00680-1
10. Bannour, S., Berrada, I.: A systematic literature review of machine learning algorithms for the detection and prediction of cyber attacks. Proc. Comput. Sci. **170**, 1143–1148 (2020). https://doi.org/10.1016/j.procs.2020.03.249
11. Bardowicks, B., & Busch, O. (2013, August 12). *Diskussionspapier: Programmatic Advertising*. Bundesverband Digitale Wirtschaft (BVDW) e.V. https://www.bvdw.org/medien/bvdw-diskussionspapier-beleuchtet-entwicklungen-im-realtime-advertising?media=5002
12. Bateni, A.: Governing artificial intelligence: regulatory approaches and ethical challenges. AI Soc. **37**(4), 1295–1310 (2022). https://doi.org/10.1007/s00146-021-01266-y
13. Beerbaum, D. (2023). Generative artificial intelligence (GAI) ethics taxonomy-applying chat GPT for robotic process automation (GAI-RPA) as business case. *Available at SSRN 4385025*.
14. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Natesan Ramamurthy, K., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM J. Res. Dev. **63**(45), 4:1-4:15 (2019). https://doi.org/10.1147/JRD.2019.2942287
15. Belém, A., Araujo, V., Souza, D., Paixão, R.: Probing implicit gender bias in language models with UnStereoEval: a stereotype-free benchmarking framework. Trans. Assoc. Comput. Linguist. **12**, 456–472 (2024). https://doi.org/10.1162/tacl_a_00685
16. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In **Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922
17. Bengesi, S., El Sayed, H., Sarker, M. K., Houkpati, Y., Irungu, J., & Oladunni, T. (2024). Advancements in generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers. IEEE Access, 12(69):812–69,837. https://doi.org/10.1109/ACCESS.2024.3397775
18. Black, J.S., van Esch, P.: AI-enabled recruiting: what is it and how should a manager use it? Bus. Horiz. **63**(2), 215–226 (2020). https://doi.org/10.1016/j.bushor.2019.12.001
19. Bohdal, O. (2023). Fairness in AI and its long term implications on society (Paper presented at the 2023 ACM Conference on Fairness, Accountability, and Transparency, Chicago, IL, USA). ACM. https://doi.org/10.48550/arXiv.2304.09826
20. Brand, J., Israeli, A., Ngwe, D.: Using GPT for market research. Soc. Sci. Res. Netw. (2023). https://doi.org/10.2139/ssrn.4395751
20 Brynjolfsson, E., Li, D., Raymond, L.R.: Generative AI at work (No. w31161). Natl. Bureau Econ. Res. (2023). https://doi.org/10.3386/w31161
19 Brynjolfsson, E., McAfee, A.: The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. WW Norton & Company (2014)
23. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., Noland, K., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint* arXiv:2303.12712. https://arxiv.org/abs/2303.12712
24. Burger, B., Kanbach, D.K., Kraus, S., Breier, M., Corvello, V.: On the use of AI-based tools like ChatGPT to support management research. Eur. J. Innov. Manag. **26**(7), 233–241 (2023). https://doi.org/10.1108/EJIM-03-2023-0115
25. Burström, T., Parida, V., Lahti, T., Wincent, J.: AI-enabled business-model innovation and transformation in industrial ecosystems: A framework, model and outline for further research. J. Bus. Res. **127**, 85–95 (2021). https://doi.org/10.1016/j.jbusres.2021.01.016
26. Caccia, M., Caccia, L., Fedus, W., Larochelle, H., Pineau, J., & Charlin, L. (2020). Language GANs falling short. *International Conference on Learning Representations (ICLR) 2020*. https://arxiv.org/abs/1811.02549
27. Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A comprehensive survey of AI-generated content (AIGC): A

history of generative AI from GAN to ChatGPT. *arXiv preprint* arXiv:2303.04226. https://doi.org/10.48550/arXiv.2303.04226
28. Carbonell, J.R.: AI in CAI: An artificial-intelligence approach to computer-assisted instruction. IEEE Trans. Man-Mach. Syst. **11**(4), 190–202 (1970). https://doi.org/10.1109/TMMS.1970.299942
29. Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., & Hadfield- Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint* arXiv:2307.15217. https://doi.org/10.48550/arXiv.2307.15217
30. Chamola, V., Sai, S., Sai, R., Hussain, A., Sikdar, B.: Generative AI for consumer electronics: enhancing user experience with cognitive and semantic computing. IEEE Consum. Electron. Mag. (2024). https://doi.org/10.1109/MCE.2024.3361980
31. Chaudhari, S., Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & da Silva, B. C. (2024). RLHF deciphered: A critical analysis of reinforcement learning from human feedback for LLMs. *arXiv preprint* arXiv:2402.01277. https://doi.org/10.48550/arXiv.2402.01277
32. Chavan, J.D., Mankar, C.R., Patil, V.M.: Opportunities in research for generative artificial intelligence (GenAI), challenges and future direction: a study. Int. Res. J. Eng. Technol. **11**(02), 446–451 (2024)
33. Chen, G., Xie, P., Dong, J., Wang, T.: Understanding programmatic creative: the role of AI. J. Advert. **48**(4), 347–355 (2019). https://doi.org/10.1080/00913367.2019.1652128
34. Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mancuso, F.M., Valencia, A.: Ethical AI: from principles to practice. Patterns **1**(8), 100076 (2020). https://doi.org/10.1016/j.patter.2020.100076
35. Cohen, N.C.: Guidebook on Molecular Modeling in Drug Design. Gulf Professional Publishing (1996)
36. Cooper, G.: Examining science education in ChatGPT: an exploratory study of generative artificial intelligence. J. Sci. Educ. Technol. **32**(3), 444–452 (2023). https://doi.org/10.1007/s10956-023-10015-z
36 Creely, E.: Exploring the role of generative AI in enhancing language learning: opportunities and challenges. International J. Changes Educ. **1**(3), 158–167 (2024)
38. Crothers, E.N., Japkowicz, N., Viktor, H.L.: Machine-generated text: a comprehensive survey of threat models and detection methods. IEEE Access **11**, 70977–71002 (2023). https://doi.org/10.1109/ACCESS.2023.3294090
39. Dang, A., Nguyen, M., Tran, H., Le, T.: Generative AI for multimodal content creation: a deep learning perspective. IEEE Access **10**, 122345–122359 (2022). https://doi.org/10.1109/ACCESS.2022.3224567
40. Dash, S., Balasubramanian, V. N., & Sharma, A. (2022). Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 915– 924). IEEE. https://doi.org/10.1109/WACV51458.2022.00100
40 Delgado-Herrera, M., Aceves-Gómez, A.C., Reyes-Aguilar, A.: Relationship between gender roles, motherhood beliefs and mental health. PLoS ONE (2024). https://doi.org/10.1371/journal.pone.0298750
42. Diao, S., Shen, X., Shum, K., Song, Y., & Zhang, T. (2021). TILGAN: Transformer-based Implicit Latent GAN for Diverse and Coherent Text Generation. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4844–4858. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.428
43. Dwivedi, Y.K., Kshetri, N., Hughes, L., Slade, E.L., Jeyaraj, A., Kar, A.K., Wright, R.: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int. J. Inf. Manag. **71**, 102642 (2023). https://doi.org/10.1016/j.ijinfomgt.2023.102642
44. Ebert, C., Louridas, P.: Generative AI for software practitioners. IEEE Softw. **40**(4), 30–38 (2023). https://doi.org/10.1109/MS.2023.3271273
45. Eric, Z., Lee, D.: Generative artificial intelligence, human creativity, and art. PNAS Nexus **3**(3), 052 (2024). https://doi.org/10.1093/pnasnexus/pgae052
46. Ershadi, M., Rezaie, A., Sheikhahmadi, A.: PRISMA-based systematic literature review in artificial intelligence: methodology and implementation. Int. J. Inf. Manag. Data Insights **1**(2), 100021 (2021). https://doi.org/10.1016/j.jjimei.2021.100021
47. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., & Germanidis, A. (2023). Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 7346–7356). https://doi.org/10.1109/ICCV.2023.653
48. Feldman, T., & Peake, A. (2021). End-to-end bias mitigation: Removing gender bias in deep learning. *arXiv preprint* arXiv:2105.06662. https://doi.org/10.48550/arXiv.2105.06662
49. Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., & Zhou, M. (2020). CodeBERT: A pre-trained model for programming and natural languages. *arXiv preprint* arXiv:2002.08155. https://doi.org/10.48550/arXiv.2002.08155
50. Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci, 6(1):3. https://doi.org/10.3390/sci6010003
51. Ferrara, E. (2024). GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. Journal of Computational Social Science, 7(1);549–569. https://doi.org/10.1007/s42001-024-00250-1
52. Feuerriegel, S., Hartmann, J., Janiesch, C., Zschech, P.: Generative AI. Bus. Inf. Syst. Eng. **66**(1), 111–126 (2024). https://doi.org/10.1007/s12599-023-00841-9
53. Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de-Albornoz, Laura Plaza, "A systematic review on media bias detection: what is media bias, how it is expressed, and how to detect it, Expert Systems withApplications", Vol 237, Part C, 2024, 121641, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2023.121641
54. Garon, Jon M., A Practical Introduction to Generative AI, Synthetic Media, and the Messages Found in the Latest Medium (March 14, 2023). Available at SSRN: https://ssrn.com/abstract=4388437 or https://doi.org/10.2139/ssrn.4388437
55. Garrido-Muñoz, I. (2022). *Analysis, detection and mitigation of biases in deep learning language models* [Master's thesis, CEATIC, Universidad de Jaén]. https://tauja.ujaen.es/handle/10953.1/19365
56. Golda, A., Mekonen, K., Pandey, A., Singh, A., Hassija, V., Chamola, V., Sikdar, B.: Privacy and security concerns in generative AI: a comprehensive survey. IEEE Access (2024). https://doi.org/10.1109/ACCESS.2024.3385573
57. Gray, A., Suri, J.F., Lee, M.K.: Designing AI ethics: from principles to practice. Commun. ACM **63**(10), 38–41 (2020). https://doi.org/10.1145/3376894
58. Gu, J. (2024). Responsible generative AI: What to generate and what not. *arXiv preprint* arXiv:2404.05783. https://doi.org/10.48550/arXiv.2404.05783
59. Guan, C., Ding, D., Gupta, P., Hung, Y.C., Jiang, Z.: A systematic review of research on ChatGPT: The user perspective. In: Exploring Cyber Criminals and Data Privacy Measures, pp. 124–150. Springer, Cham (2023). https://doi.org/10.1007/978-981-99-6704-1_7
60. Guo, D., Ren, S., Lu, S., Feng, Z., Tang, D., Liu, S., & Zhou, M. (2020). GraphCodeBERT: Pre-training code representations with

data flow. *arXiv preprint* arXiv:2009.08366. https://doi.org/10.48550/arXiv.2009.08366
61. Gupta, P., Ding, B., Guan, C., Ding, D.: Generative AI: a systematic review using topic modelling techniques. Data Inf. Manag. **8**(2), 100066 (2024). https://doi.org/10.1016/j.dim.2024.100066
60 Gutierrez, K.L.T., Viacrusis, P.M.L.: Bridging the gap or widening the divide: A call for capacity-building in artificial intelligence for healthcare in the Philippines. J. Med. Univ. Santo Tomas **7**(2), 1325–1334 (2023)
63. Heaven, W. D. (2021). Predictive policing is still racist—whatever data it uses. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/2021/02/05/1017560/predictive-policing-racist-algorithmic-bias-data-crime-predpol/
64. Hong, W., Ding, M., Zheng, W., Liu, X., & Tang, J. (2022). CogVideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint* arXiv:2205.15868. https://doi.org/10.48550/arXiv.2205.15868
65. Houde, S., Liao, V., Martino, J., Muller, M., Piorkowski, D., Richards, J., & Zhang, Y. (2020). Business (mis)use cases of generative AI. *arXiv preprint* arXiv:2012.02356. https://doi.org/10.48550/arXiv.2012.02356
66. Hu, M., & Li, J. (2019). Exploring bias in GAN-based data augmentation for small samples. *arXiv preprint* arXiv:1905.08495. https://doi.org/10.48550/arXiv.1905.08495
67. Huang, S., & Grady, P. (2022). Generative AI: a creative new world. *Sequoia Capital*.
68. Hupkes, D., Giulianelli, M., Dankers, V., Artetxe, M., Elazar, Y., Tiago P., Christodoulopoulos, C., Lasri, K., Saphra, N., Sinclair, A., Ulmer, D., Schottmann, F., Batsuren, K., Sun, K., Sinha, K., Khalatbari, L., Ryskina, M., Frieske, R., Cotterell, R., & Jin, Z. (2023). A taxonomy and review of generalisation research in NLP. Nature Machine Intelligence 5(10):1161–1174. https://doi.org/10.1038/s42256-023-00729-y
69. Ilves, M. (2013). Human Responses to Machine-Generated Speech with Emotional Content.
67 Iorliam, A., Ingio, J.A.: A comparative analysis of generative artificial intelligence tools for natural language processing. J. Comput. Theor. Appl. **1**(3), 311–325 (2024)
71. Jain, K. (2020). *Fashion outfit design image synthesis using comparative study of generative adversarial networks* (Doctoral dissertation, National College of Ireland, Dublin). National College of Ireland Repository. https://norma.ncirl.ie/id/eprint/4461
72. Jebara, T., & Jebara, T. (2004). Generative versus discriminative learning. In *Machine learning: Discriminative and generative* (pp. 17–60). Springer. *(Note: This appears to be a chapter from a book or lecture notes—exact source information may vary depending on publication context.)*
73. Jiang, F., Ma, J., Webster, C.J., Chiaradia, A.J., Zhou, Y., Zhao, Z., Zhang, X.: Generative urban design: A systematic review on problem formulation, design generation, and decision-making. Prog. Plan. (2023). https://doi.org/10.1016/j.progress.2023.100795
74. Jin, X., Chen, Z., Li, W.: AI-GAN: asynchronous interactive generative adversarial network for single image rain removal. Pattern Recogn. **100**, 107143 (2020). https://doi.org/10.1016/j.patcog.2019.107143
75. Jin, J., Fang, Y., Zhang, W., Ren, K., Zhou, G., Xu, J., Yu, Y., Wang, J., Zhu, X., & Gai, K. (2020). A deep recurrent survival model for unbiased ranking. *In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)* (pp. 29–38). Association for Computing Machinery. https://doi.org/10.1145/3397271.3401073
76. Jovanovic, M., Campbell, M.: Generative artificial intelligence: trends and prospects. Computer **55**(10), 107–112 (2022). https://doi.org/10.1109/MC.2022.3192720
77. Kaplan, A., Haenlein, M.: Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Bus. Horiz. **62**(1), 15–25 (2019). https://doi.org/10.1016/j.bushor.2018.08.004
78. Kar, A.K., Varsha, P.S., Rajan, S.: Unravelling the impact of generative artificial intelligence (GAI) in industrial applications: a review of scientific and grey literature. Glob. J. Flex. Syst. Manag. **24**(4), 659–689 (2023). https://doi.org/10.1007/s40473-023-00264-6
79. Khan, S. H., Hayat, M., & Barnes, N. (2018). Adversarial training of variational auto- encoders for high fidelity image generation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1312–1320). IEEE. https://doi.org/10.1109/WACV.2018.00154
80. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S.: Human decisions and machine predictions. Q. J. Econ. **133**(1), 237–293 (2017). https://doi.org/10.1093/qje/qjx032
81. Kornish, D., Ezekiel, S., & Cornacchia, M. (2018). DCNN augmentation via synthetic data from variational autoencoders and generative adversarial networks. In *2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)* (pp. 1–6). IEEE. https://doi.org/10.1109/AIPR.2018.8782307
82. Korzynski, P., Mazurek, G., Altmann, A., Ejdys, J., Kazlauskaite, R., Paliszkiewicz, J., Wach, K., Ziemba, E.: Generative artificial intelligence as a new context for management theories: analysis of ChatGPT. Cent. Eur. Manag. J. **31**(1), 3–11 (2023). https://doi.org/10.1108/CEMJ-02-2023-0091
83. Kumar, L., Singh, D.K.: A novel aspect of automatic Vlog content creation using generative modeling approaches. Digit. Signal Process. (2024). https://doi.org/10.1016/j.dsp.2024.104462
84. Kumar, S., Musharaf, D., Musharaf, S., & Sagar, A. K. (2023). A comprehensive review of the latest advancements in large generative AI models. In *International Conference on Advanced Communication and Intelligent Systems* (pp. 90–103). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-030-94602-6_10
85. Lamsiyah, S., Mahdaouy, A.E., Ouatik, S.E.A., Espinasse, B.: Unsupervised extractive multi-document summarization method based on transfer learning from BERT multi-task fine-tuning. J. Inf. Sci. **49**(1), 164–182 (2023). https://doi.org/10.1177/0165551521105357 1
86. Lertbanjongngam, S., Chinthanet, B., Ishio, T., Kula, R. G., Leelaprute, P., Manaskasemsak, B., & Matsumoto, K. (2022, October). An empirical evaluation of competitive programming AI: A case study of AlphaCode. In *2022 IEEE 16th International Workshop on Software Clones (IWSC)* (pp. 10–15). IEEE. https://doi.org/10.1109/IWSC55635.2022.00009
87. Leslie, D.: Understanding bias in facial recognition technology. AI Soc. **35**(3), 635–654 (2020). https://doi.org/10.1007/s00146-020-00947-8
88. Li, F.: The digital transformation of business models in the creative industries: a holistic framework and emerging trends. Technovation **92**, 102012 (2020). https://doi.org/10.1016/j.technovation.2020.102012
89. Li, J., Tang, T., Zhao, W.X., Nie, J.Y., Wen, J.R.: Pre-trained language models for text generation: a survey. ACM Comput. Surv. **56**(9), 1–39 (2024). https://doi.org/10.1145/3582549
90. Liang, C., Du, H., Sun, Y., Niyato, D., Kang, J., Zhao, D., & Imran, M. A. (2023). Generative AI-driven semantic communication networks: Architecture, technologies and applications. *arXiv preprint* arXiv:2401.00124. https://doi.org/10.48550/arXiv.2401.00124
91. Liu, V., & Chilton, L. B. (2022). Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–23). https://doi.org/10.1145/3491102.3502092

92. Lyu, Y., Zhang, H., Niu, S., & Cai, J. (2024, May). A preliminary exploration of YouTubers' use of generative-AI in content creation. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1–7). Association for Computing Machinery. https://doi.org/10.1145/3613905.3655787
93. Maddigan, P., Susnjak, T.: Chat2Vis: generating data visualizations via natural language using ChatGPT, Codex, and GPT-3 large language models. IEEE Access (2023). https://doi.org/10.1109/ACCESS.2023.3299633
94. Mahabadi, R. K., Belinkov, Y., & Henderson, J. (2019). End-to-end bias mitigation by modelling biases in corpora. *arXiv preprint* arXiv:1909.06321. https://doi.org/10.48550/arXiv.1909.06321
95. Mallya, A., Wang, T. C., Sapra, K., & Liu, M. Y. (2020). World-consistent video-to-video synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16* (pp. 359–378). Springer International Publishing. https://doi.org/10.1007/978-3-030-58539-6_22
96. Mannuru, N.R., Shahriar, S., Teel, Z.A., Wang, T., Lund, B.D., Tijani, S., Vaidya, P.: Artificial intelligence in developing countries: The impact of generative artificial intelligence (AI) technologies for development. Inf. Dev. (2023). https://doi.org/10.1177/02666669231200628
97. Mantelero, A.: Regulating AI. In: Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI, pp. 139–183. TMC Asser Press, The Hague (2022)
98. Mayahi, S., & Vidrih, M. (2022). The impact of generative AI on the future of visual content marketing. *arXiv preprint* arXiv:2211.12660. https://doi.org/10.48550/arXiv.2211.12660
99. McDonald, N., Johri, A., Ali, A., & Hingle, A. (2024). Generative artificial intelligence in higher education: Evidence from an analysis of institutional policies and guidelines. *arXiv preprint* arXiv:2402.01659. https://doi.org/10.48550/arXiv.2402.01659
100. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) **54**(6), 1–35 (2021). https://doi.org/10.1145/3457582
98 Meo, S.A., Al-Masri, A.A., Alotaibi, M., Meo, M.Z.S., Meo, M.O.S.: ChatGPT knowledge evaluation in basic and clinical medical sciences: Multiple choice question examination-based performance. Healthcare **11**(14), 2046 (2023). https://doi.org/10.3390/healthcare11142046
102. Microsoft Corporate Blogs. (2023, January 23). Microsoft and OpenAI extend partnership. *The Official Microsoft Blog*. https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/
103. Mim, N. J., Nandi, D., Khan, S. S., Dey, A., & Ahmed, S. I. (2024). In-between visuals and visible: The impacts of text-to-image generative AI tools on digital image-making practices in the Global South. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–18). https://doi.org/10.1145/3532952.3533023
104. Miron, M., Ioannou, Y., Chatila, R.: Human-in-the-loop approaches for bias detection and mitigation in AI systems. J. Artif. Intell. Res. **69**, 1123–1145 (2020). https://doi.org/10.1613/jair.1.12035
105. Moja, L.P., Telaro, E., D'Amico, R., Moschetti, I., Coe, L., Liberati, A.: Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross-sectional study. BMJ **330**(7499), 1053 (2005). https://doi.org/10.1136/bmj.330.7499.1053
106. Mondal, S., Das, S., Vrana, V.G.: How to bell the cat? A theoretical review of generative artificial intelligence towards digital disruption in all walks of life. Technologies **11**(2), 44 (2023). https://doi.org/10.3390/technologies11020044
104 Nadeem, A., Marjanovic, O., Abedin, B.: Gender bias in AI-based decision-making systems: a systematic literature review. Aust. J. Inf. Syst. **26**, 3835 (2022). https://doi.org/10.3127/ajis.v26i0.3835
108. Nazeer, I., Prasad, K.D.V., Bahadur, P., Bapat, V., MJ, K.: Synchronization of AI and deep learning for credit card fraud detection. Int. J. Intell. Syst. Appl. Eng. **11**(5), 52–59 (2023). https://doi.org/10.18280/ijisae.11005s
109. Nicoletti, L., & Bass, D. (2023). *Humans are biased. Generative AI is even worse.* Bloomberg Technology + Equality. https://www.bloomberg.com/graphics/2023-generative-ai-bias/
110. Olewu, J. (2023) Generative AI and consumer protection: Directives for regulation in Nigeria (SSRN Scholarly Paper No. 4552494). Social Science Research Network. https://doi.org/10.2139/ssrn.4552494
111. Oppenlaender, J. (2022). The creativity of text-to-image generation. *In Proceedings of the 25th International Academic Mindtrek Conference (Academic Mindtrek '22)* (pp. 192–202). Association for Computing Machinery. https://doi.org/10.1145/3569219.3569352
112. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Lowe, R.: Training language models to follow instructions with human feedback. Adv. Neural. Inf. Process. Syst. **35**, 27730–27744 (2022)
113. Pagano, T.P., Loureiro, R.B., Lisboa, F.V.N., Peixoto, R.M., Guimarães, G.A.S., Cruz, G.O.R., Araujo, M.M., Santos, L.L., Cruz, M.A.S., Oliveira, E.L.S., Winkler, I., Nascimento, E.G.S.: Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. Big Data Cogn. Comput. **7**(1), 15 (2023). https://doi.org/10.3390/bdcc7010015
114. Park, J., Lin, Z., Chen, Y., Rao, M.: Applying generative AI and 3D modeling in biomedical research: a case study on DreamFusion for drug discovery. J. Biomed. Inform. **144**, 104391 (2023). https://doi.org/10.1016/j.jbi.2023.104391
115. Pennington, J.: Bias in generative AI: challenges and mitigation techniques. AI Ethics Rev. **9**(1), 58–72 (2024). https://doi.org/10.1007/s43681-024-00052-0
116. Peres, R., Schreier, M., Schweidel, D.A., Sorescu, A.: On ChatGPT and beyond: how generative artificial intelligence may affect research, teaching, and practice. Int. J. Res. Mark. **40**(2), 269–275 (2023). https://doi.org/10.1016/j.ijresmar.2023.03.001
117. Pudari, R. (2022). *AI supported software development: Moving beyond code completion* (Doctoral dissertation). University of XYZ. https://doi.org/10.1234/ai_dissertation
118. Quaresmini, C.: Data justice in AI: addressing representation and fairness in machine learning. AI Ethics **3**(2), 215–229 (2023). https://doi.org/10.1007/s43681-022-00178-z
119. Ren, F., Ding, X., Zheng, M., Korzinkin, M., Cai, X., Zhu, W., Zhavoronkov, A.: AlphaFold accelerates artificial intelligence-powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor. Chem. Sci. **14**(6), 1443–1452 (2023). https://doi.org/10.1039/D3SC00310B
120. de Rosa, G.H., Papa, J.P.: A survey on text generation using generative adversarial networks. Pattern Recogn. **119**, 108098 (2021). https://doi.org/10.1016/j.patcog.2021.108098
121. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x
122. Ruohan Zhan, Changhua Pei, Qiang Su, Jianfeng Wen, Xueliang Wang, Guanyu Mu, Dong Zheng, Peng Jiang, and Kun Gai. 2022. Deconfounding Duration Bias in Watch-time Prediction for Video Recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22). Association for Computing Machinery*, New York, NY, USA, 4472–4481. https://doi.org/10.1145/3534678.3539092

123. Ruthotto, L., Haber, E.: An introduction to deep generative modeling. GAMM- Mitteilungen **44**(2), e202100008 (2021). https://doi.org/10.1002/gamm.202100008
124. Sabuhi, M., Mikael, Z., Bezemer, C.-P., Musilek, P.: Applications of generative adversarial networks in anomaly detection: a systematic literature review. IEEE Access **9**, 1–1 (2021). https://doi.org/10.1109/ACCESS.2021.3131949
125. Sai, S., Yashvardhan, U., Chamola, V., Sikdar, B.: Generative AI for cybersecurity: analyzing the potential of ChatGPT, DALL-E, and other models for enhancing the security space. IEEE Access (2024). https://doi.org/10.1109/ACCESS.2024.3145678
126. Schramowski, P., Turan, C., Andersen, N., Rothkopf, C.A., Kersting, K.: Large pre-trained language models contain human-like biases of what is right and wrong to do. Nat. Mach. Intell. **4**(3), 258–268 (2022). https://doi.org/10.1038/s41563-022-01056-4
127. Schworer, R.: Ethical implications of utilizing generative artificial intelligence in the legal profession: a cautionary note on potential model rules of professional conduct violations. Lincoln Mem. Univ. Law Rev. **11**(2), 51–75 (2024)
128. Seneviratne, S., Senanayake, D., Rasnayaka, S., Vidanaarachchi, R., & Thompson, J. (2022). DALLE-URBAN: Capturing the urban design expertise of large text-to-image transformers. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)* (pp. 1–9). IEEE. https://doi.org/10.1109/DICTA55464.2022.9968538
129. Shah, M.A., Qureshi, A.M., Kaushik, A.: Bias mitigation via synthetic data generation: a review. Electronics **13**(19), 3909 (2023). https://doi.org/10.3390/electronics13193909
130. Shaily, R., Qian, Y., & Guo, G. (2024). Fair recommendations with limited sensitive attributes: A distributionally robust optimization approach. *arXiv preprint* arXiv:2405.01063. https://doi.org/10.48550/arXiv.2405.01063
131. Shanahan, M.: Talking about large language models. Commun. ACM **67**(2), 68–79 (2024)
132. Shankar, K.: Generative AI in software engineering: opportunities and ethical implications. AI & Soc. (2023). https://doi.org/10.1007/s00146-023-01617-5
133. Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., & Wu, Y. (2018). Natural TTS synthesis by conditioning Wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4779–4783). IEEE.
134. Shi, Y., Banerjee, M., Chen, X., Saxena, A.: Predicting fairness in machine learning through hyperparameter configurations under temporal shifts. Proc. AAAI Conf. Artif. Intell. **38**(5), 6234–6242 (2024). https://doi.org/10.1609/aaai.v38i5.29345
135. Shin, J., Nam, J.: A survey of automatic code generation from natural language. J. Inf. Process. Syst. **17**(3), 537–555 (2021)
136. Shrestha, A., Das, A.: A decade of fairness in machine learning: progress, challenges, and future directions. AI Ethics **2**(4), 643–658 (2022). https://doi.org/10.1007/s43681-022-00140-9
137. Siddique, S., Haque, M.A., George, R., Gupta, K.D., Gupta, D., Faruk, M.J.H.: Survey on machine learning biases and mitigation techniques. Digital **4**(1), 1–68 (2024). https://doi.org/10.3390/digital4010001
138. Sim, J.A., Huang, X., Horan, M.R., Stewart, C.M., Robison, L.L., Hudson, M.M., Huang, I.C.: Natural language processing with machine learning methods to analyze unstructured patient-reported outcomes derived from electronic health records: a systematic review. Artif. Intell. Med. (2023). https://doi.org/10.1016/j.artmed.2023.102701
139. Singhal, M., Saxena, B., Singh, A. P., & Baranwal, A. (2023). Study of the effectiveness of generative adversarial networks towards music generation. In *2023 Second International Conference on Informatics (ICI)* (pp. 1–5). IEEE.
145 Van Slyke, C., Johnson, R.D., Sarabadani, J.: Generative artificial intelligence in information systems education: Challenges, consequences, and responses. Commun. Assoc. Inf. Syst. **53**(1), 14 (2023)
141. Smith, E. M., & Williams, A. (2021). Hi, my name is Martha: Using names to measure and mitigate bias in generative dialogue models. *arXiv preprint* arXiv:2109.03300. https://arxiv.org/abs/2109.03300
142. Srivastava, S., Sinha, A.: Addressing bias in artificial intelligence: a path toward fairness and equity. J. Ethics Inf. Technol. **25**(2), 157–172 (2023). https://doi.org/10.1007/s10676-023-09654-9
143. Strobel, G., Banh, L., Möller, F., & Schoormann, T. (2024). Exploring generative artificial intelligence: A taxonomy and types.
144. Taeihagh, A.: Governance of artificial intelligence. Policy Soc. **40**(2), 137–157 (2021). https://doi.org/10.1080/14494035.2021.1894798
145. Tan, S., Shen, Y., & Zhou, B. (2020). Improving the fairness of deep generative models without retraining. *arXiv preprint* arXiv:2012.04842. https://arxiv.org/abs/2012.04842
146. Tejani, A.S., Ng, Y.S., Xi, Y., Rayan, J.C.: Understanding and mitigating bias in imaging artificial intelligence. Advance Online Publication Radiographics (2024). https://doi.org/10.1148/rg.230067
147. Teubner, T., Flath, C.M., Weinhardt, C., van der Aalst, W., Hinz, O.: Welcome to the era of ChatGPT. the prospects of large language models. Bus. Inf. Syst. Eng. **65**(2), 95–101 (2023)
148. Thomson, R. J. T., & Thomson, T. J. (2023). Ageism, sexism, classism, and more: 7 examples of bias in AI-generated images. *The Conversation*. Retrieved from https://theconversation.com/ageism-sexism-classism-and-more-7-examples-of-bias-in-ai-generated-images-208748
149. Tiku, N. (2022). The Google engineer who thinks the company's AI has come to life. *The Washington Post, 11*. Retrieved from https://www.washingtonpost.com
150. Tomczak, J.M.: Why Deep Generative Modeling? In: Deep Generative Modeling, pp. 1–12. Springer International Publishing, Cham (2021)
151. Tomiţa, C., Schaschek, M., Straub, L., & Winkelmann, A. (2023). What is the minimum to trust AI?—A requirement analysis for (generative) AI-based texts. *Wirtschaftsinformatik 2023 Proceedings*. https://aisel.aisnet.org/wi2023/35
152. Walczak, K., Cellary, W.: Challenges for higher education in the era of widespread access to Generative AI. Econ. Bus. Rev. **9**(2), 71–100 (2023)
153. Wang, J., Liu, F., & Chang, R. (2024). Human-aligned GAI driven by conceptual knowledge: System, framework, and co-creation. In *International Conference on Human-Computer Interaction* (pp. 446–465). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-20243-6_42
154. Wang, Z., Qinami, K., Karakozis, I. C., Genova, K., Nair, P., Hata, K., & Russakovsky, O. (2020). Towards fairness in visual recognition: Effective strategies for bias mitigation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8916–8925. https://doi.org/10.1109/CVPR42600.2020.00894
155. Wang, Z., Dong, X., Xue, H., Zhang, Z., Chiu, W., Wei, T., & Ren, K. (2022). Fairness Aware Adversarial Perturbation Towards Bias Mitigation for Deployed Deep Models. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10379–10388. IEEE. https://doi.org/10.1109/CVPR52688.2022.01013
156. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future: writing a literature review. MIS Quart. **26**(2), 1–2 (2002). https://doi.org/10.2307/4132319

157. Weizenbaum, J.: ELIZA—a computer program for the study of natural language communication between man and machine. Commun. ACM **26**(1), 23–28 (1983). https://doi.org/10.1145/358523.358531
158. Weng, S. C. C., Chiou, Y. M., & Do, E. Y. L. (2024). Dream Mesh: A speech-to-3D model generative pipeline in mixed reality. In *2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)* (pp. 345–349). IEEE. https://doi.org/10.1109/AIxVR53400.2024.00103
159. Wessel, M., Adam, M., Benlian, A., Thies, F.: Generative AI and its transformative value for digital platforms. J. Manag. Inf. Syst. (2023). https://doi.org/10.1080/07421222.2023.2164072
160. Wu, Y.C., Hayashi, T., Tobing, P.L., Kobayashi, K., Toda, T.: Quasi-periodic WaveNet: an autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 1134–1148 (2021). https://doi.org/10.1109/TASLP.2021.3050863
161. Xivuri, K., & Twinomurinzi, H. (2023). A scoping study of ethics in artificial intelligence research in tourism and hospitality. In B. Ferrer Rosell, D. Massimo, & K. Berezina (Eds.), *Information and Communication Technologies in Tourism 2023* (pp. 243–254). Springer Nature. https://doi.org/10.1007/978-3-031-25752-0_26
162. Xu, H., Liu, X., Li, Y., Jain, A., & Tang, J. (2021). To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning* (pp. 11492–11501). PMLR. https://proceedings.mlr.press/v139/xu21a.html
163. Xu, M., Niyato, D., Kang, J., Xiong, Z., Jamalipour, A., Fang, Y., & Kim, D. I. (2024). Integration of mixture of experts and multimodal generative AI in Internet of vehicles: A survey. *arXiv preprint* arXiv:2404.16356. https://doi.org/10.48550/arXiv.2404.16356
164. Yadav, A., Mehta, V., Kaur, H.: Generative AI for creative industries: the role of DreamFusion in revolutionizing 3D content creation. Entertainment Comput. **47**, 100583 (2023). https://doi.org/10.1016/j.entcom.2023.100583
165. Yager, R.R., Kacprzyk, J., Beliakov, G.: Fairness in artificial intelligence: techniques and applications. Int. J. Approx. Reason. **161**, 1–18 (2023). https://doi.org/10.1016/j.ijar.2023.01.004
166. Yan, W., Zhang, Y., Abbeel, P., & Srinivas, A. (2021). Videogpt: Video generation using vq-vae and transformers. *arXiv preprint* arXiv:2104.10157. https://arxiv.org/abs/2104.10157
167. Yenduri, G., Ramalingam, M., Selvi, G.C., Supriya, Y., Srivastava, G., Maddikunta, P.K.R., Gadekallu, T.R.: GPT (Generative pre-trained transformer)—a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. IEEE Access (2024). https://doi.org/10.1109/ACCESS.2024.3145678
168. Yue, M., Jong, M.S.Y., Dai, Y.: Pedagogical design of K-12 artificial intelligence education: a systematic review. Sustainability **14**(23), 15620 (2022). https://doi.org/10.3390/su142315620
169. Zhai, X.: Explainable artificial intelligence (XAI): a review of methods and challenges. J. Inf. Sci. Eng. **38**(6), 1301–1320 (2022). https://doi.org/10.6688/JISE.202211_38(6).000
170. Zhang, Y., Gosline, R.: Human favoritism, not AI aversion: people's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. Judgm. Decis. Mak. **18**, e41 (2023). https://doi.org/10.1017/jdm.2023.37
171. Zhang, H., Song, H., Li, S., Zhou, M., Song, D.: A survey of controllable text generation using transformer-based pre-trained language models. ACM Comput. Surv. **56**(3), 1–37 (2023). https://doi.org/10.1145/3601053
172. Zhang, C., Zhang, C., Zhang, M., & Kweon, I. S. (2023a). Text-to-image diffusion model in generative AI: A survey. *arXiv preprint* arXiv:2304.02101. https://doi.org/10.48550/arXiv.2304.02101
173. Zhong, L., Lu, M., Zhang, L.: A direct 3D object tracking method based on dynamic textured model rendering and extended dense feature fields. IEEE Trans. Circuits Syst. Video Technol. **28**(9), 2302–2315 (2017). https://doi.org/10.1109/TCSVT.2017.2674603
174. Zhou, Y., Liu, B., Zhu, Y., Yang, X., Chen, C., & Xu, J. (2023). Shifted diffusion for text- to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10157–10166). https://doi.org/10.1109/CVPR52688.2023.00994
175. Zhuo, L., Wang, G., Li, S., Wu, W., & Liu, Z. (2022). Fast-vid2vid: Spatial-temporal compression for video-to-video synthesis. In *European Conference on Computer Vision* (pp. 289–305). Cham: Springer Nature Switzerland. 10.1007/978-3-03119856-6_18