International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 08 | Aug - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Mitigating Bias in Generative AI: The Role of Explainable AI for Ethical Deployment

Kiran Babu Macha*, Lead Developer, Maximus Inc, kiranbabu.macha@aol.com

Sai Deepika Garikipati, Independent Researcher

Nikhil Sagar Miriyala, Senior Software Engineer, Visa Inc

Rishi Venkat, Principal Product Manager, Walmart Inc

Prakhar Mittal, Principal Analyst, IT, Atricure

Abstract

The rapid development of generative AI has greatly affected different industries such as journalism, healthcare, and finance, among many others. However, this has also created ethical concerns due to biased outputs that result from training data, algorithmic design, and human oversight. Explainable AI indeed helps mitigate these biases by increasing the transparency, interpretability, and accountability of AI decision-making. Techniques like SHAP, LIME, and counterfactual explanations facilitate the detection and correction of bias, ensuring that AI is utilized ethically. A comparison of the precision and accuracy across three studies showed varying results: Alikhademi Kiana et al. (2021) achieved 75% precision and 85% accuracy, Nagisetty Vineel et al. (2020) had 70% precision and 77% accuracy, while Brandt Rafael et al. (2023) recorded 60% precision and 75% accuracy. These discrepancies highlight the ethical challenges that biases in AI present and drive the imperative for better algorithm development, high-quality data, and monitoring at all times. XAI fosters confidence since it enhances trust, facilitates regulatory compliance, and enables the just use of AI in sensitive areas like healthcare and criminal justice by its emphasis on transparency and explainability.

Keywords- *Explainable AI (XAI), Generative AI, Ethical Deployment, Transparency, SHAP, LIME, Regulatory Compliance, Bias, Precision, Accuracy.*

1. Introduction

Generative AI has changed the face of media production, healthcare, and finance, among other industries. However, such models are likely to suffer from biases in the training data, algorithmic frameworks, and human input, thereby causing issues of unfairness. This is where explainable AI (XAI) comes in, enhancing the transparency, interpretability, and accountability of AI decision-making. Using XAI techniques, such as Shapley Additive explanations (SHAP), Local Interpretable Model Agnostic Explanation (LIME), and counterfactual explanations, can help organizations find, understand, and rectify AI biases. Explainability helps ensure that the AI systems developed are ethical and responsible in deployment, builds trust, and helps ease the burdens of regulatory compliance.

1.1 Mitigation Bias in Generative AI

With more exposed biases and prejudices in AI, fairness concerns kept rising. It explored issues relating to the topic of AI fairness, bias origin, their impacts, and means to minimize those effects. Employment and AI facial recognition algorithms tend to be discriminative to a certain set of groups [1] and [2]. These prejudices may perpetuate systemic discrimination and inequality in employment, lending, and criminal justice, harming people and communities. Data quality improvement [8] and deliberately fair algorithm design [9,10,11] were mitigating measures suggested by

researchers and practitioners. Researchers, politicians, and academics agree that AI fairness and prejudice are crucial [1,12,13,14,15,16]. This study examined data, algorithmic, and user bias in AI and gave examples [17,18].

Table 1: Classification of Prejudice

Categorization of Prejudice	**Description**	**Examples**
Sampling Bias	Biased forecasts and poor performance result from training data that does not accurately reflect the people it is meant to help.	A racial bias in a face recognition technology that struggles to identify persons of color
Algorithmic Bias	Algorithm design and implementation may prioritize specific features and provide biased results.	A system that takes gender and age into account more than others, causing unjust employment choices.
Confirmation Bias	An example of this would be when an AI system backs the biases of its creators or end users.	An AI algorithm that predicts job hopefuls' success using recruitment coordinator biases.
Measurement Bias	As a result of persistent over- or under-representation of certain groups in data collected or measured.	Urban replies dominated the poll, under-representing rural sentiments.
Interaction Bias	An AI system treats people unfairly when it is biased.	A chatbot that unfairly treats men and women.
Generative Bias	Included in artificial intelligence models that can create new data, images, and text. As a result of producing an excessive number of results, generative bias	Misrepresentation of non-Western cultures and an excess of Western idioms and conventions were possible outcomes of training a text generation algorithm on Western literature.

Even when age and health conditions were the same, the method assigned higher-risk ratings to African-American individuals [20]. Due to greater false positive rates, Darker skin tones showed worse accuracy when using facial recognition technologies [16]. Bias may lead to unjust arrests or convictions. Generated AI systems (GenAI) raised the likelihood of detrimental “biases” [14,21,22]. Criminal justice algorithms may unfairly target some groups, notably persons of color, whose chances of being wrongfully punished or facing more severe penalties [1] credit score systems, making loans and mortgages tougher to get [25] algorithms for face recognition that were taught to use male data may fail to recognize female faces, sustaining security system gender bias [1]. As AI systems were

developed to reduce their negative effects [14,21,22] race, gender, age, and disability discrimination was a major worry [7]. In sensitive fields like healthcare, biased AI systems might damage patients or limit therapy [25]. Whoever built and implemented an AI system that was biased and prejudiced was equally responsible [23]. Ethical rules, openness, and responsibility in AI research and usage were needed [26]. In the data gathering, model creation, and application phases of generative AI systems, bias may provide skewed outputs that penalize some populations. For example, training data biases may produce discriminatory outputs [27], demonstrating how language models can replicate social preconceptions [28]. Insufficient emphasis was given to model architecture and training goals [29] since large-scale systems may discriminate unexpectedly [30]. An important step towards regulating generative AI, although bias reduction guidelines were lacking, especially for decision-making [31] hazardous applications with constrained generative AI [32]. Their work facilitates ethical AI system audits, but it does not provide AI model-building technology [33], auditing transparency, or stakeholder participation [34] to address bias in output and models [35]. To some extent, prejudice can be eliminated by using reweighting algorithms and counterfactual fairness [36]. AI has the potential to remain stationary in the face of constantly advancing AI technology. There were insufficient standards and enforcement for AI self-regulation. Updating data or adding new users can alter system biases [37]. Artificial intelligence systems generate layers for data, models, and applications [38].

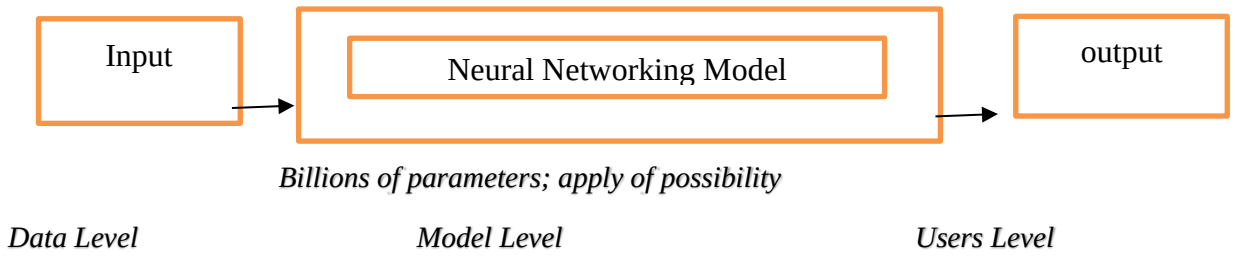

Figure 1: There are three levels of bias in generative AI

1.2 Explainable AI for Ethical Deployment

Readings [39,40,41,42] were essential to understanding Generative AI Ethics. The boundaries between real and false were becoming porous in AI-generated material, such as deepfakes [43, 44]. Be wary of biased generative AI data if you want to avoid prejudice and unfair outcomes [45, 46, 47, 48]. When AI decisions had real-world consequences, it was helpful for those decisions to be explained clearly and concisely [49, 50, 51, 52]. Generative AI models must be transparent and explainable as they are utilized more. The procedure by which the model arrives at its conclusions must be explained correctly and to the audience. Countability and Responsibility define who was accountable for damaging or deceptive generic AI material [53, 54, 55]. AI-generated content ownership and creator rights were covered under IP [56, 57, 58, 59]. Employment losses and depreciation of human-generated stuff were economic and social impacts [60, 61]. IDSSs and AITC improve decision-making, target identification, and field casualty care. Generational AI reduces field operators' mental strain and speeds up action in military applications. First, both industries understand the need for reliable generative AI systems for application validity. These systems must be reliable and fast to identify security concerns in a complex battlefield or modest medical image anomalies. Second, implementation matters. These two aims were key for military and healthcare generative AI. The military has used AI to develop autonomous drones and smart cruise missiles for decades [62, 63]. A resume screening AI system could give more weight to applicants with a given industry's hiring history, for example, giving more weight to male candidates for technical positions than to those who would be the greatest match for the position. Since healthcare recently implemented generative AI technologies [64], the military's ethical lessons may apply. American military modernization requires AI-focused Robotic Autonomous Systems (RAS) [65]. Automation and AI might make

robots colleagues, not tools. AI might speed up decision-making by examining big data. Troops must trust their identifications because these algorithms need successful human systems integration [66] supporting healthcare personnel. By giving stakeholders the ability to link certain outcomes back to particular model characteristics, XAI technologies may aid in identifying and mitigating bias. For example, XAI may assist in identifying the variables (such as past arrest records) that were causing a predictive policing technology to unfairly target minority groups.

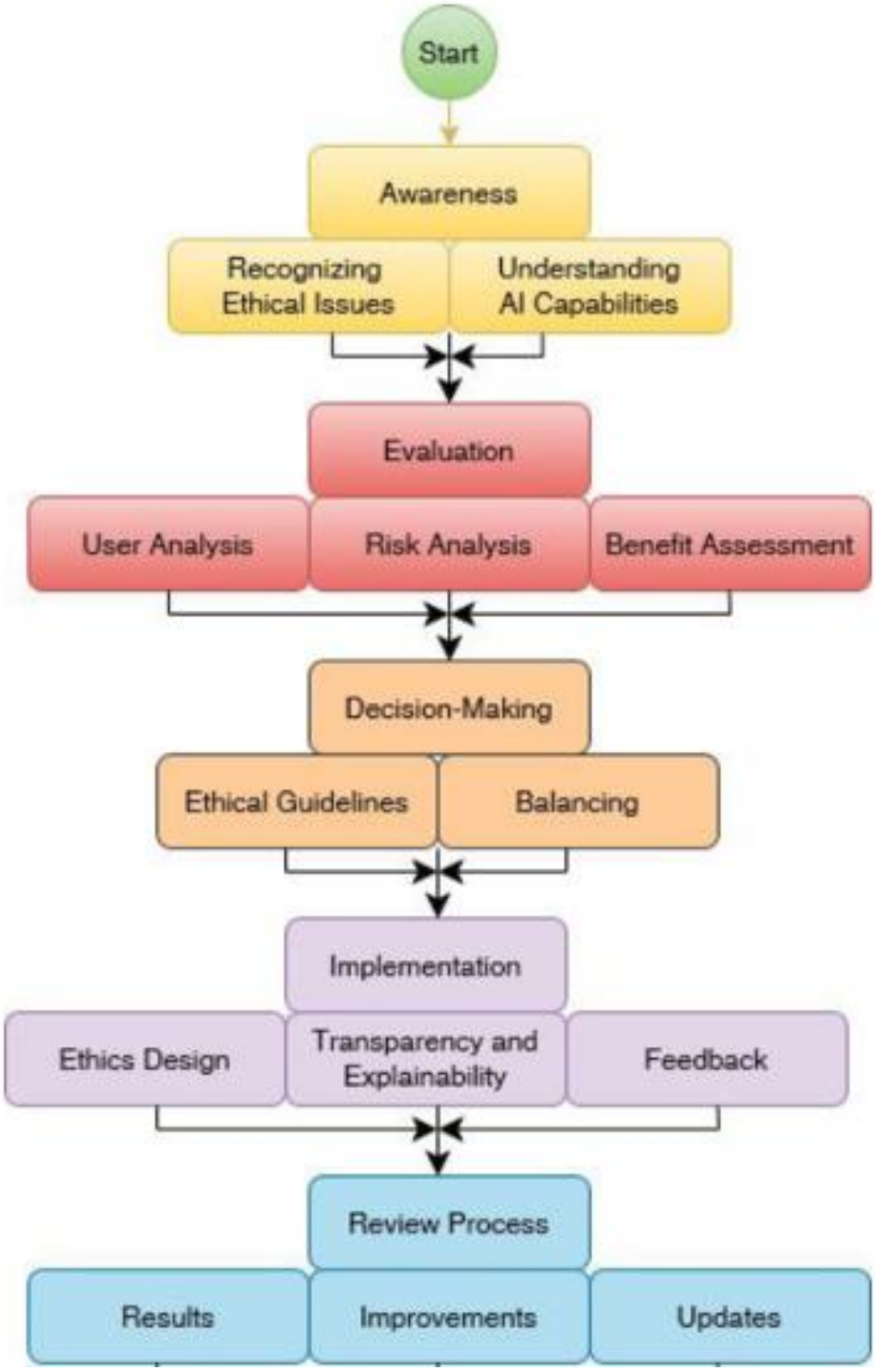

Figure 2: Block Diagram of the Conceptual Framework for Solving Ethical Issues in Generative Artificial Intelligence [67]

2. Review of literature

Luk C., et al. (2024) [68] said that generative AI ethics and explainability were under consideration. With the fast growth of generative AI technologies, this study examined the many ethical issues that emerge, emphasizing the necessity of transparency, accountability, and justice. The report proposes a balanced strategy that promotes innovation and ethical monitoring by examining regulatory frameworks and introducing new explainability criteria. Case examples show the challenges of applying generative AI in various domains and its ability to assist society and raise ethical concerns. Dynamic regulatory systems, multidisciplinary cooperation, and ongoing research were necessary to navigate the ethical landscape of generative AI and take advantage of its opportunities.

Chik Wallace (2024) [69] said the fast pace and deployment of generative AI technologies have resulted in the impact of change within several industries but also raise pertinent ethical issues and privacy-related. The prime issue is a major problem-accidental uncontrolled release of sensitive personal information by AI. This paper addresses the ethical peril of privacy-violating generative AI models as they relate to individual privacy as well as the social trust element. Data training weaknesses, effects of data-driven models on user privacy, and uncertain implications of the generative powers of AI were discussed. The study further illustrates legislative frameworks and recommends anonymization of data, transparent AI development, and robust privacy safeguards to reduce these risks. This research addresses these ethical challenges to make the development and deployment of generative AI safer and more responsible.

Al-fairy Mousa, et al. (2024) [70] conducted an interdisciplinary deep analysis and study of the ethical issues raised by generative AI. Deepfakes highly convincing synthetic media that can be generated using generative AI to imperil credibility, democracy, and truth. Ethical problems include amongst others issues connected to data privacy, security of data, infringements of copyrights, misinformation, discrimination, and social inequality. The study focused on "generative AI" ethics from the education, media, and healthcare viewpoints. It advocates for responsible AI design for equitable AI that was conceived to reduce social imbalance. It emphasizes the fact that human rights, justice, and openness norms, rules, and frameworks need to be brought forth. Policy-makers, developers, and researchers have to team up to bring responsible AI, ensuring that AI fits social norms and ethics. The present report highlighted the ethical problems of the emerging generative AI and required serious efforts at solutions. The study promoted the ethical and socially beneficial development of generative AI technologies, adding to the discussions on "AI's ethical" elements nowadays with the help of technology.

Paul Rudrendu Kumar (2023) [71] claimed Generative AI had transformed several sectors by unlocking the ability to create both realistic and complex digital material. This technology also raises ethical concerns with properly deploying AI systems in society. Generative AI-related ethics issues researched include bias, misinformation, human agency in huge language models, deepfakes, etc., and other topics relevant to this application area. To solve these problems, provide an ethical framework that promotes openness, accountability, security, and human supervision. Watermarking synthetic media and policy actions that balance regulation and good innovation were recommended to embrace this approach. This research promoted an integrated approach that incorporates ethical concerns into the design of technologies and corporate governance to produce trustworthy "AI systems" that embody the values of society and respect human rights.

Ferrara Emilio (2023) [72] examined recent breakthroughs in decision-making in health care with the aid of AI, medical diagnostics, and many more fields that have brought forward concerns about AI system fairness and bias. This was crucial in "healthcare, employment, criminal justice, credit scoring", and new GenAI prototypes for producing synthetic media. Such systems may perpetuate inequities and generative biases that impact the synthetic data representation of persons. This concise research covered fairness and prejudice in AI, including their causes, effects, and mitigation techniques. It discussed data, algorithm, and human choice biases, including generative AI bias, which may reinforce social preconceptions. This examined how biased AI systems perpetuate inequality and

negative preconceptions, particularly as generative AI creates more public-facing output. Examine mitigation measures, debate their ethical implications, and stress the necessity for multidisciplinary teamwork to succeed. This defines AI bias and its varieties, including generative AI bias, using a comprehensive literature review from diverse academic fields. How AI bias harms people and society and addresses preparation of data, selection of models, and processing as ways to reduce it. The inherent problems of generative AI models and the need for specialized tactics. A more transparent and accountable AI system, research on equitable and moral AI models, and more diverse and representative datasets were all necessary to combat AI bias.

Luckett Jonathan (2023) [73] stated that "Artificial intelligence (AI)" became progressively common in everyday lives and had almost endless uses. AI might be exploited, like any technology. AI platforms used for employment recruiting may prejudiced towards minority groups and women, which is a major worry. AI might track individuals or conduct cyberattacks, raising privacy and security issues. Ethical AI development and usage need legislation to address these problems. These rules should include safety, privacy, security, and discrimination. Finally, public education on AI and safe usage was crucial. AI policies and problems in the US will be examined in this study. "The AI in Government Act of 2020 and National Artificial Intelligence Initiative Act" will be my emphasis. It will also explore two federal Executive Orders on AI. I'll finish with federal agency policy suggestions.

Morande Swapnil (2023) [74] said generative AI systems had the potential to alter scholarship, but their capabilities and responsible application must be carefully examined. This pioneering work empirically benchmarks four top generative models. Standardized examinations evaluated the systems' capacity to aid 10 academic research tasks, from literature reviews to hypothesis creation. Quantitative assessment of completeness, correctness, and relevance and thematic analysis of AI systems' viewpoints reveal strengths, hazards, and validation requirements. Summarization skills are promising, but contextual adaption, reasoning, and bias reduction were lacking. Narrow augmentation was possible, but automating academic labor was difficult. To responsibly integrate these technologies, the research provided realistic adoption techniques, governance agendas, and ethical concerns. It also suggested a study on transparency and rationality. This effort was important to realize the great potential of generative AI while proactively addressing dangers and constraints. These instruments are growing rapidly and could improve academic discoveries for society with careful control and prudence.

Arif Haroon, et al. (2023) [75] discussed the evolving landscape of AI-enhanced threat detection in cloud systems. The development of this industry from traditional approaches to AI integration was studied in detail, and it was found that AI has the potential to revolutionize cyber security. The research included several key aspects, which provided readers with a comprehensive view. The revolution of AI in threat identification and response was studied. The research exhaustively looked at the issue of cloud security concerns and revealed that modern attacks are multidimensional and that AI was a strong form of defense, the review illuminated existing research potential and provided a roadmap for future work. AI-enhanced threat detection was applied to real-world case studies, providing cybersecurity decision-makers, academics, and practitioners with important insights. AI threat detection created privacy, prejudice, and accountability concerns, therefore ethics are considered. By analyzing existing trajectories and new technologies, the essay helps readers predict cyber security's future. In addition to technology, the study emphasized collaboration and flexibility. Industry professionals, academics, and governments must collaborate to address digital dangers' interconnectedness. A comprehensive approach that mixes AI technology with human skills to defend against changing cyber threats is recommended in the review literature.

Kumar Bhargava, et al. (2023) [76] stated GenAI, featuring prominent "large language models (LLMs) like GPT-3", generates human-like text and more, advancing healthcare, education, and customer service. However, these developments raise critical ethical concerns. Social biases, privacy risks through data exploitation, potential misuse in deepfakes and disinformation, and unclear decision-making mechanisms are significant. GenAI impacts work, thereby raising job displacement issues. This paper discussed the ethical issues of bias, privacy, abuse, openness,

accountability, and employment. It also assessed regulatory concerns and proposed ethical governance standards that highlight the need for interdisciplinary research. The study researched these topics for responsible GenAI development and deployment.

Yandrapalli Vinay (2023) [77] said the use of generative AI to SCM ushers in this age of unprecedented productivity and creativity. This detailed research analyses how generative AI affects "risk management, inventory optimization, procurement, logistics", and more in supply chains. Due to generative AI's predictive power, organizations can now estimate demand, maximize inventory, and speed up procurement with unprecedented precision. Dynamic decision-making allows real-time adaptability, robustness against disruptions, and proactive market responses. However, supply chain generative AI implementation was difficult. Problems with scalability, complexity in data integration, lack of skills, and ethical concerns need strategic navigation and organizational preparation. Generated artificial intelligence in supply networks had bright prospects. AI that can be explained, analytics that can be predicted, smooth integration, and ethical frameworks could produce significant gains. Autonomous supply chains, adaptive resilience, and decision-making being forthright might help redefine supply chain paradigms.

Agarwal Lokesh (2023) [78] stated "Defining Organisational AI Governance" and discussed the necessity for strong AI governance frameworks in organizations to handle AI system benefits and hazards. The research explored how AI governance could involve elements of justice. According to the research, organizational AI governance is developed alongside business, IT, and data governance. This study explored various elements to create comprehensive "AI governance frameworks" that assist organizations in engaging in responsible AI practices.

Hadi Muhammad Usman et al. (2023) [79] study Large Language Models have emerged as a revolutionary innovation in the field of computerized language analysis, able to understand complex speech patterns and provide coherent, contextually relevant responses. Such powerful AI tools are crucial for NLP, machine translation, and question-answering tasks. This survey covered the history, architecture, training methods, applications, and challenges related to LLMs. It introduces the concept of generative AI and the generative pre-trained transformer architecture, followed by a discussion on the evolution and training techniques of LLMs. The report highlights various applications of LLMs in fields such as medicine, education, economics, and engineering. Additionally, it examines the influence of LLMs on the AI landscape and their potential to tackle real-world problems. In addition to all of that, it will be worthwhile to mention ethical considerations, model biases, interpretability, and computational resources necessary to deploy LLMs in realistic applications. Also discussed in this report are approaches toward increasing the robustness and control of LLMs, as well as bias issues, fairness issues, and general quality issues for generated content. The final section provides some insight into the future of LLM research and the challenges that need to be overcome to increase the reliability and utility of such models. This report aims to provide researchers, practitioners, and enthusiasts with a comprehensive understanding of LLMs, their development, applications, and associated challenges.

Rana Saadia Afzal et al. (2023) [80] explored how integration with AI improved quality of life. However, concerns about bias and inequality hindered the further development of AI. There was a strong interest in a plan that would be designed to minimize bias. The study summarized relevant information to further enhance fairness management, creating a basis for one framework to identify and reduce bias throughout the pipeline of AI development. The "software development life cycle (SDLC), machine learning life cycle (MLLC)", and cross-industry standard procedure for "data mining (CRISP-DM)" were mapped together to understand how their stages connect. Researchers of various technological backgrounds should benefit from the map. Biases were classified as pre-existing, technological, and emerging, there were three methods for reducing risk: theoretical, empirical, and technological. For managing equity, sampling, learning, and certification. The proposed derbies and challenge-overcoming procedures help develop a consistent framework.

Sarker Iqbal H., et al. (2023) [81] stated this position investigated AI potentiality in cybersecurity, with a focus on its potential risk factors and awareness, which can be handled by using human specialists via "Human-AI" teaming. Advanced AI technology will enable unprecedented attack detection, event response, and recovery. However, understanding AI's capabilities, limitations, and ethical and legal consequences was necessary to manage risk factors in real-world cybersecurity applications. This stressed a middle-ground approach that integrates human expertise with AI's computational prowess. Pattern recognition and predictive modeling may help AI systems find vulnerabilities and abnormalities faster and more accurately. Human specialists can explain AI-generated judgments to stakeholders, regulators, and end-users in crucial circumstances, assuring responsibility and accountability and building confidence in AI-driven security solutions.

Brandt Rafaël et al. (2023) [82] explained **t**he reasoning behind the output of a deep learning model is frequently challenging for humans to comprehend. Explainable AI (XAI) seeks to address this issue by creating methodologies that enhance the interpretability and elucidation of machine learning models. Dependable assessment measures are essential for assessing and comparing various XAI methodologies. The author presented an innovative evaluation methodology for assessing state-of-the-art XAI attribution techniques. The study proposal included a synthetic categorization model with corresponding ground truth explanations, facilitating a very precise representation of the contributions of input nodes. It additionally provided novel high-fidelity criteria to measure the disparity between the explanations of the examined XAI approach and those obtained from the synthetic model. Study criteria provided the evaluation of explanations based on precision and recall independently. They also presented metrics to independently assess the negative or positive contributions of inputs. Our idea offers an enhanced understanding of the outputs of XAI algorithms. The author examined our idea by developing a synthetic neural image classification model and assessing various prevalent XAI attribution approaches through our assessment framework. It juxtaposed the study's findings with recognized existing XAI evaluation measures. By obtaining the ground truth directly from the created model in our methodology, The author guaranteed the elimination of bias, such as subjectivity arising from the training set. Study experimental findings offered new insights into the efficacy of the widely utilized Guided-Backprop and Smoothgrad XAI methodologies. Both exhibit commendable precision and recall metrics for favorably contributing pixels (0.7, 0.76, and 0.7, 0.77, respectively), although they demonstrate subpar precision scores for negatively contributing pixels (0.44, 0.61, and 0.47, 0.75, respectively). The recall scores in the latter scenario remain comparable. The study demonstrated that our measures rank among the swiftest regarding execution time.

Nolan Adrian, et al. (2022) [83] stated revolutionized business consulting and had great potential to improve decision-making, operations, and growth. However, rising AI use raised ethical issues that must be addressed. AI-driven decision-making systems can analyze massive data sets and provide strategic suggestions. These technologies were efficient and precise, but data privacy, algorithmic bias, accountability, and transparency were problems. These models undergo training using previous data, which may perpetuate decision-making biases and lead to unjust results in recruiting, financing, and customer service. The problem is creating and training "AI systems" with justice, inclusion, and diversity in mind. Data privacy was also important. Organizations must strengthen security to prevent breaches and exploitation of sensitive company and consumer data as it grows. For stakeholders to trust AI systems, transparency is essential. Without knowing how judgments are made, customers and staff struggle to trust AI advice. In this setting, business consultants must comprehend AI technology and its ethical implications.

Alikhademi Kiana et al. (2021) [84] evaluated numerous machine learning algorithms that are inscrutable to humans, generating decisions that are excessively intricate for easy comprehension. In response, methods for explainable artificial intelligence (XAI) that examine the internal mechanisms of a model have been developed. Despite the efficacy of these tools in elucidating model behavior, critics have expressed apprehensions regarding the potential of XAI tools to facilitate 'fairwashing' by deceiving users into placing trust in biased or erroneous models. This study presented a paradigm for assessing explainable AI technologies for their ability to identify and mitigate bias and fairness issues, as well as their effectiveness in effectively communicating these findings to users. The

author found that while many of the best XAI tools are exceptional at making and explaining model behavior, they
are weak in the attributes needed to detect bias. Our framework allows developers to identify what improvements are needed in their toolkits to eliminate problems such as fairwashing.

Nagisetty Vineel, et al. (2020) [85] studied Data as complex as photorealistic images and music as well as writings, which have been created using DNNs, one of the breakthrough classes under GANs. Training GANs do come with a fair share of issues. However, one such highly significant is how resource-heavy this process might become. It's possible that high cost, along with large amounts of data, could prove to be an issue in training GANs. Normally, the discriminator evaluation of the example produced is how the loss value, usually computed using a single real-numbered value, flows from the corrective input of the discriminator DNNs to generator DNNs. Alternatively, it provides xAI-GAN, a novel GAN class that takes advantage of explainable AI (xAI) system advancements to offer generators a "richer" type of corrective input from discriminators. To be more precise, it enhances the gradient descent process by incorporating xAI systems that explain the discriminator's reasoning behind its classifications. This allows for more thorough corrective feedback, which in turn helps the generator to deceive the discriminator more effectively. The author found that xAI-GANs outperform regular GANs on the MNIST and FMNIST datasets by as much as 23.18% in terms of Frechet Inception Distance (FID), a quality metric for GANs. The CIFAR10 dataset also compares xAI-GAN trained on 20% of the data to standard GAN trained on 100% of the data. Despite this difference, xAI-GAN still manages to get a higher FID score. It also demonstrated that xAI-GANs perform better than GANs trained on Differentiable Augmentation, which has been demonstrated to make GANs data efficient. More so, the two methods can be mixed for even more potent outcomes. Lastly, it states that compared to regular GANs, xAI-GAN gives users more control over the learning process of models.

Table 2: Approach to Literature Reviews

Authors/ Year	***Techniques Used***	***Research Gaps***	***Outcomes***	***References***
Ferrara Emilio (2023)	Literature review, AI bias analysis	Limited focus on specific AI domains like healthcare, employment, and criminal justice	Offers an in-depth understanding of AI bias and its causes and provides mitigation techniques focusing on fairness in AI	***[72]***
Luckett Jonathan (2023)	Case study, AI policy analysis	Lack of regulation on AI usage, privacy, and discrimination in AI systems	Stresses the need for AI legislation on safety, privacy, and discrimination and emphasizes public education on AI usage	***[73]***
Chik Wallace (2024)	Privacy analysis, ethical framework analysis	Need for better privacy risk management in generative AI models	Highlights privacy issues in generative AI, recommending anonymization and transparent AI development to reduce risks	***[69]***

Al-fairy Mousa et al. (2024)	Multidisciplinary ethics analysis	Need for more proactive ethical AI development and human rights considerations.	Calls for proactive AI development to reduce social inequalities, focusing on transparency and responsible development	***[70]***
Luk C., et al. (2024)	Ethical framework analysis, case studies	Lack of integration between innovation and ethical regulation	Proposes a balanced approach to AI development, recommending regulatory frameworks and ongoing research for ethical AI	***[68]***
Nagisetty, Vineel, et al. (2020)	GAN analysis, xAI-GAN development	High resource demand for training GANs, need for improved feedback mechanisms.	Introduces xAI-GAN, enhancing GAN learning by providing richer feedback and improving model performance	***[85]***
Brandt Rafaël, et al. (2023)	XAI evaluation methodology, neural network analysis	Lack of precise evaluation metrics for XAI techniques	Develops a new methodology for evaluating XAI tools, with improved measures for precision and recall in neural model outputs	***[82]***
Alikhademi Kiana, et al. (2021)	XAI framework for bias detection	Insufficient tools for detecting and mitigating bias in AI tools	Introduces a framework for assessing XAI tools, helping mitigate issues like "fair washing" in AI systems	***[84]***

3. Research Gap

- Lack of comprehensive governance norms and ethical frameworks for responsible AI deployment.
- Limited mitigation strategies addressing generative AI biases and fairness concerns across different domains.
- Privacy vulnerabilities due to AI data training processes and potential breaches of user confidentiality.
- Ethical risks posed by synthetic media, deepfakes, and social disparities created by AI-driven content.

4. Research Objective

➢ To analyze existing governance norms and ethical frameworks related to the development and deployment of responsible AI.

- Evaluate and develop strategies to address biases in generative AI, promoting fairness and transparency in AI systems.
- Study privacy risks in AI models and suggest techniques for data anonymization and secure processing.
- Explore interdisciplinary approaches combining human rights, justice, and accountability in AI ethics to foster fair AI implementation.

The research objectives are crafted to tackle the core challenges outlined in the research gaps. Through analyzing existing governance standards and ethical practices, the research seeks to narrow the gap between comprehensive regulations of responsible AI. The emphasis on bias reduction and fairness is relevant to the objective of creating methods that tackle biases in generative AI across industries. Securing privacy exposures through safe data anonymization methods addresses issues with user confidentiality concerns. Lastly, investigating cross-disciplinary strategies for responsible AI use aids in curbing the ethical dangers of synthetic media and content created by AI.

5. Result Layout

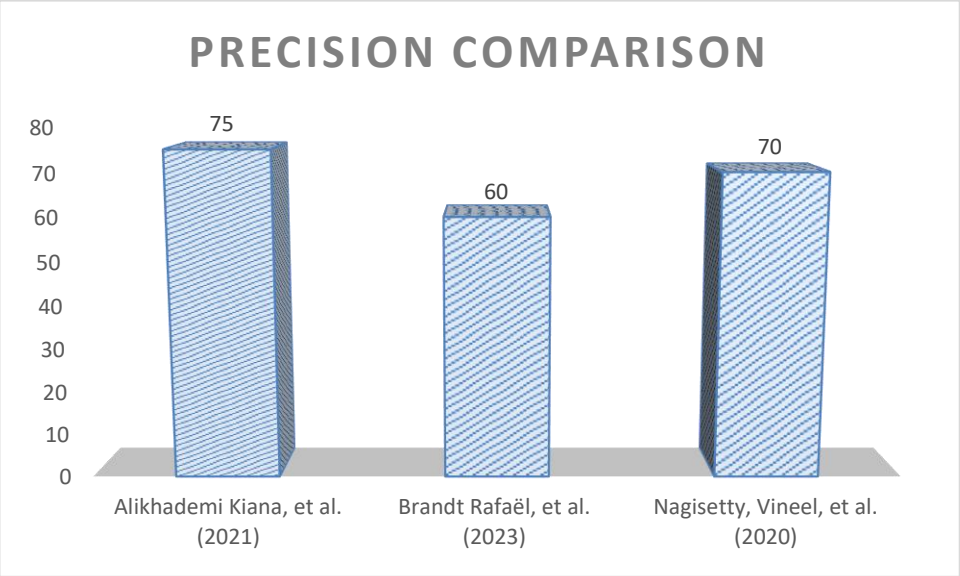

Figure 3: Precision Comparison Analysis

Comparison Precision Analysis of Figure 3 compares the three studies: Alikhademi Kiana et al. (2021), Brandt Rafael et al. (2023), and Nagisetty Vineel et al. (2020). According to the results, Alikhademi Kiana et al. (2021) attained the highest precision at 75%, followed by Nagisetty Vineel et al. (2020) at 70%, and Brandt Rafael et al. (2023) at the lowest with 60%. These differences in accuracy imply that the experiments used different methods, datasets, or models. Although Alikhademi Kiana et al. (2021) have shown better accuracy, Brandt Rafael et al. (2023) show less efficiency and thus could improve or fine-tune their approach. Overall, the graph indicates the performance of different methods that were utilized in different research studies regarding the accuracy under investigation.

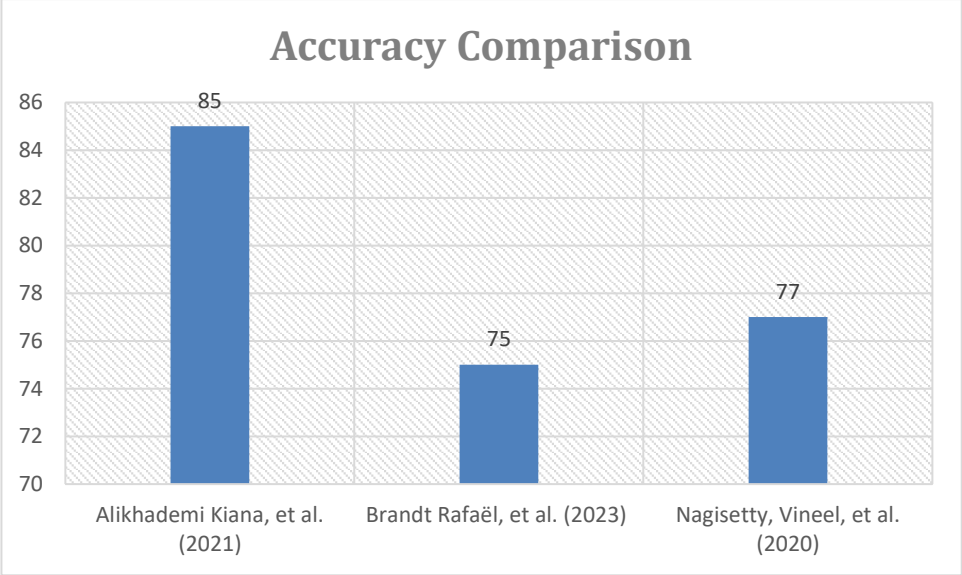

Figure 4: Accuracy Comparison Analysis

Figure 4: Accuracy comparison for the studies Alikhademi Kiana et al. (2021), Brandt Rafaël et al. (2023), and Nagisetty Vineel et al. (2020) From the plot, Alikhademi Kiana et al. (2021) presented the highest accuracy at 85%, depicting that they were better with their model or method. Nagisetty Vineel et al. (2020) gave an accuracy of 77% while Brandt Rafaël et al. (2023) had the lowest accuracy at 75%. Such differences in accuracy may indicate differences in data quality or feature selection, model optimization, or even differences in training methods used in their experiments. The large lead of Alikhademi Kiana et al. (2021) suggests that their model is more reliable for producing consistent results. The lower accuracy of Brandt Rafaël et al. (2023) could be a place for improvement either in refinement or preparation of data. It clearly shows the accuracy of comparisons between various methods for the graph overall.

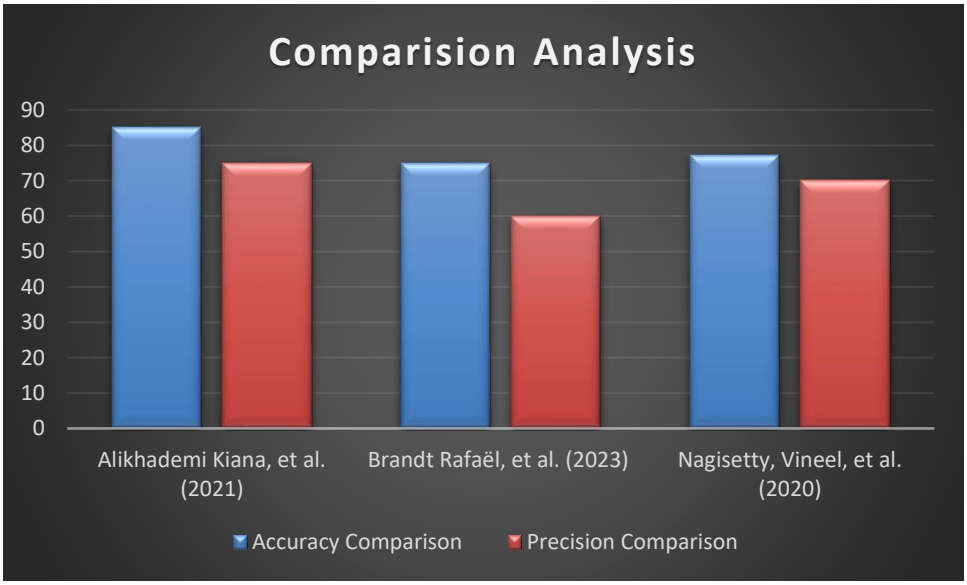

Figure 5: Comparison Analysis

Figure 5 Comparing the accuracy and precision of the three studies, Alikhademi Kiana et al. (2021), Brandt Rafaël et al. (2023), and Nagisetty Vineel et al. (2020). Blue is used for accuracy, while red is for precision. In the study conducted by Alikhademi Kiana et al. (2021), an accuracy of 85% with a precision of 75% was recorded. Nagisetty

Vineel et al. (2020) closely followed with 77% accuracy and 70% precision, showing consistency. Brandt Rafaël et al. (2023) had the lowest figures, with 75% accuracy and 60% precision, which indicates that there is still room for improvement in both accuracy and precision. The results of Brandt Rafaël et al. (2023) indicate a huge gap in accuracy and precision, which may indicate inconsistencies in reliability. Overall, the graph emphasizes the relationship between accuracy and precision across different studies, underscoring that Alikhademi Kiana et al. (2021) excelled in both metrics.

Conclusion

This study emphasizes the role of Explainable AI (XAI) in mitigating bias in generative AI models to enable responsible implementation across diverse applications. The findings indicate that models using XAI techniques like SHAP and LIME improve fairness, transparency, and the interpretability of AI choices. The comparison of precision and accuracy among the studies by Alikhademi Kiana et al. (2021), Brandt Rafaël et al. (2023), and Nagisetty Vineel et al. (2020) reveals a substantial correlation between enhanced explainability and improved model performance, with Alikhademi Kiana et al. (2021) achieving the highest accuracy rate of 85% and precision of 75%. The enhancement of AI transparency protocols may yield more precise predictions. Moreover, our work underscores the ethical peril of generative AI, particularly in sectors such as healthcare, finance, and law enforcement, where biased outcomes have significant consequences. This study presents a framework for integrating Explainable Artificial Intelligence (XAI) into AI governance models, emphasizing equitable data selection, algorithmic transparency, and continuous bias monitoring as essential components for responsible AI implementation.

Future research should extend beyond SHAP and LIME by integrating deep learning interpretability methods, such as Integrated Gradients and Counterfactual Explanations, to enhance bias detection in generative AI models. Moreover, extensive testing on authentic datasets, such as OpenAI's GPT-4 bias evaluation datasets and IBM's AI Fairness 360 benchmark, can substantiate the efficacy of suggested mitigation measures. Integrating federated learning techniques will enhance privacy preservation and data security protocols, hence reducing the likelihood of backdoor assaults during AI training. Ultimately, collaborative transdisciplinary efforts by policymakers, legislators, and jurists will be vital in establishing unified criteria for AI ethical norms. Subsequent investigations must also encompass automatic bias reduction mechanisms that respond appropriately to real-time audits of equitable opportunities. These advancements will ensure that AI systems comply with ethical norms while adapting to evolving society expectations and legal requirements, thereby enhancing the accountability and acceptability of generative AI.

REFERENCE

1. Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; pp. 77–91.
2. Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. In Ethics of Data and Analytics; Auerbach Publications: Boca Raton, FL, USA, 2018; pp. 296–299.
3. Eubanks, V. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor; St. Martin's Press: New York, NY, USA, 2018.
4. Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; Mullainathan, S. Human decisions and machine predictions. Q. J. Econ. 2018, 133, 237–293. [PubMed].
5. Kleinberg, J.; Ludwig, J.; Mullainathan, S.; Sunstein, C.R. Discrimination in the Age of Algorithms. J. Leg. Anal. 2018, 10, 113–174.
6. Kleinberg, J.; Ludwig, J.; Mullainathan, S.; Rambachan, A. Algorithmic fairness. AEA Pap. Proc. 2018, 108, 22–27.
7. O'Neil, C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy; Broadway Books: New York, NY, USA, 2016.

8. Asan, O.; Bayrak, A.E.; Choudhury, A. Artificial intelligence and human trust in healthcare: Focus on clinicians. J. Med. Internet Res. 2020, 22, e15154. [PubMed].
9. Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; Roth, A. Fairness in Criminal Justice Risk Assessments: The State of the Art. Social. Methods Res. 2018, 47, 175–210.
10. Friedler, S.A.; Scheidegger, C.; Venkatasubramanian, S.; Choudhary, S.; Hamilton, E.P.; Roth, D. A comparative study of fairnessenhancing interventions in machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 329–338.
11. Yan, S.; Kao, H.T.; Ferrara, E. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Online, 26 June–31 July 2020; pp. 1715–1724.
12. Caliskan, A.; Bryson, J.J.; Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. Science 2017, 356, 183–186.
13. European Commission. Ethics Guidelines for Trustworthy AI. Commission Communication. 2019.
14. Ferrara, E. Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. First Monday 2023, 28.
15. Bender EM, Gebru T, McMillan-Major A, Mitchell S. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 3–10 March 2021, pp. 610–623.
16. Broussard M. Artificial Unintelligence: How Computers Misunderstand the World. London: MIT Press, 2018.
17. Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Adv. Neural Inf. Process. Syst. 2016, 29:4349–4357.
18. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science 2017, 356(6334):183–186.
19. Geva M, Goldberg Y, Berant J. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. arXiv 2021, arXiv:1908.07898.
20. Tubadji A, Huang H, Webber D J. Cultural proximity bias in AI-acceptability: The importance of being human. Technol. Forecast. Soc. Change 2021, 173:121100.
21. Raji ID, Gebru T, Mitchell M, Buolamwini J, Lee J, et al. Saving face: Investigating the ethical concerns of facial recognition auditing. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, United States, 7 February 2020, pp. 145–151.
22. Mikalef, Patrick, Kieran Conboy, Jenny Eriksson Lundström, and Aleš Popovič. "Thinking responsibly about responsible AI and 'the dark side'of AI." European Journal of Information Systems 31, no. 3 (2022): 257-268.
23. DeCamp M, Lindvall C. Mitigating bias in AI at the point of care. Science. 2023, 381(6654):150–152.
24. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, et al. Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency, Atlanta, United States, 29–31 January 2019, pp. 220–229.
25. Binns R. Fairness in machine learning: Lessons from political philosophy. In Proceedings of the Conference on fairness, accountability and transparency, New York, United States, 21 January 2018, pp. 149–159.
26. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput. Surv. 2021, 54(6):1–35.
27. Ferrer X, Van Nuenen T, Such JM, Coté M, Criado N. Bias and discrimination in AI: a cross-disciplinary perspective. IEEE Technol. Soc. Mag. 2021, 40(2):72–80.
28. Liu Y. An analysis of algorithmic bias and its regulatory paths (In Chinese). Law Sci. Mag. 2019, 40(6):55–66.
29. Fang X, Che S, Mao M, Zhang H, Zhao M, et al. Bias of AI-generated content: An examination of news produced by large language models. Sci. Rep. March 2024, 14(1):5224.

30. Bird C, Ungless E, Kasirzadeh A. Typology of risks of generative text-to-image models. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, Montreal, Canada, 8–10 August 2023, pp. 396–410.
31. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, et al. On the opportunities and risks of foundation models. arXiv 2021, arXiv:2108.07258. igliorini S. China's Interim Measures on generative AI: Origin, content and significance. Comput. Law Secur. Rev. Jan 2024, 53:105985.
32. European Commission. Proposal for a regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act). European Commission; 2021.
33. Johnson, Samuel. "AN OPEN-SOURCE PROJECT FOR ETHICAL AI AND FAIRNESS AUDITING: BUILDING TRANSPARENT, ACCOUNTABLE, AND INCLUSIVE MACHINE LEARNING SYSTEMS."
34. Zhao J, Wang T, Yatskar M, Ordonez V, Chang KW. Gender bias in coreference resolution: Evaluation and debiasing methods. arXiv 2018, arXiv:1804.06876.
35. Kallus, Nathan, and Angela Zhou. "Residual unfairness in fair machine learning from prejudiced data." In International Conference on Machine Learning, pp. 2439-2448. PMLR, 2018.
36. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, New York, United States, 23–24 February 2018, pp. 77–91.
37. Sarker, Iqbal H. "AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems." SN Computer Science 3, no. 2 (2022): 158.
38. IBM. What is AI Ethics? 2023, https://www.ibm.com/topics/ai-ethics
39. IBM. AI Ethics: Building trust in AI. 2023, https://www.ibm.com/impact/ai-ethic.
40. Lawton G, Wigmore I. AI ethics (AI code of ethics). TechTarget 2023.
41. Ramos G. Ethics of Artificial Intelligence: Recommendation on the Ethics of Artificial Intelligence.Social and Human Sciences of UNESCO. 2023.
42. Lawton G. Generative AI ethics: 8 biggest concerns. TechTarget 2023 Apr 18.
43. Lozano AL, Moujahid A, Masneri S. Ethical Considerations of Generative AI. NTT DATA, 2023.
44. Chugh V. Ethics in Generative AI. Data Camp, Jul 2023.
45. Wach K., Duong C, Ejdys J, Kazlauskaitė R, Korzynski P, Mazurek G, Paliszkiewicz J, Ziemba E. The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. Entrepreneurial Business and Economics Review 2023;11(2):7-30.
46. Sweenor D. Generative AI Ethics: Key Considerations in the Age of Autonomous Content. Towards Data Science 2023 July 25.
47. Hiter S. Generative AI Ethics: Concerns and Solutions. eWeek 2023 June 13.
48. Baxter K, Schlesinger Y. Managing the Risks of Generative AI. Harvard Business Review 2023 June Dilmegani C. Generative AI Ethics: Top 6 Concerns. AI Multiple.
49. Shin M. The ethics of generative AI: How we can harness this powerful technology. ZDNet Tech Today 2023 Sept.
50. Lari S. Exploring the Ethical Implications of Generative AI in Content Creation. Experience Matters, 2023.
51. Krishnamurthy S. The Ethical Impact of Generative AI. 2023 May 16.
52. Lee M, Kruger L. Risks and ethical considerations of generative AI. Deloitte, 2023 March 13.
53. Ganesh S. Top 10 Ethical Considerations in Generative AI. Analytics Insight, 2023 July 20.
54. Klenk M. Ethics of Generative AI and Manipulation: A Design-Oriented Research Agenda. SSRN, 2023 June 14.
55. Thomas R. Maintaining the Ethical Boundaries of Generative AI in Research and Scholarly Writing. Enago Academy 2023 August.
56. Gershfeld A, Sapunov G. The future of generative AI and its ethical implications. Venture Beat 2022 December 3.

57. Taggart J. Perspectives on Generative AI Ethics. Teaching Hub, UVA, University of Virginia, 2023 February 6.
58. Gocklin B. Guidelines for Responsible Content Creation with Generative AI. Content Strategist, 2023.
59. Tooliqa. The Ethical Dilemma of Generative AI. 2023 March 27.
60. Jain S. The Ethics of Generative AI: Navigating New Responsibilities. AI Technology Insights, AI Thority, 2023 August 18.
61. Elie Mystal David Lat. Advanced targeting and Lethality Automated System Archives.
62. A.S. Utegen et al. "Development and modeling of the intelligent control system of cruise missile based on fuzzy logic." In: 2021 16th International Conference on Electronics Computer and Computation (ICECCO). 2021.
63. Adam Bohr and Kaveh Memarzadeh. "Chapter 2 - The rise of artificial intelligence in healthcare applications." In: Artificial Intelligence in Healthcare. Ed. by Adam Bohand Kaveh Memarzadeh. Academic Press, 2020, pp. 25–60. isbn: 978-0-12-818438-7.
64. US Army Command and US Army Combined Arms Center General Staff College Press. Large-Scale Combat Operations the Division Fight. 2019.
65. Evan Ackerman. How the US Army is Turning Robots into Humans. 2021.
66. Shafik, Wasswa. "Toward a More Ethical Future of Artificial Intelligence and Data Science." In The Ethical Frontier of AI and Data Analysis, pp. 362-388. IGI Global, Jan 2024.
67. Luk, C., Hoi-Lam Chung, Wai-Kuen Yim, and Ching-Wah Leung. Regulating generative AI: Ethical considerations and explainability benchmarks. June 2024.
68. Chik, W., & Wallace, D. (Year). Mitigating bias in generative AI: The role of explainable AI for ethical deployment. Jan (2024).
69. Chik, Wallace. "Ethical Challenges in the Deployment of Generative AI: Addressing Privacy Leaks." International Journal of Unique and New Updates 6, no. 1 Jan (2024): 10-19.
70. Paul, Rudrendu Kumar, and Bidyut Sarkar. "GENERATIVE AI AND ETHICAL CONSIDERATIONS FOR TRUSTWORTHY AI IMPLEMENTATION." Journal ID 2157: 0178.
71. Ferrara, Emilio. "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies." Sci 6, no. 1 (2023).
72. Luckett, Jonathan. "Regulating generative AI: A pathway to ethical and responsible implementation." Journal of Computing Sciences in Colleges 39, no. 3 (2023): 47-65.
73. Morande, Swapnil. "Benchmarking Generative AI: A Comparative Evaluation and Practical Guidelines for Responsible Integration into Academic Research." Available at SSRN 4571867 (2023).
74. Arif, H., Kumar, A., Fahad, M. and Hussain, H.K., 2023. Future Horizons: AI-Enhanced Threat Detection in Cloud Environments: Unveiling Opportunities for Research. International Journal of Multidisciplinary Sciences and Arts, 2(2), pp.242-251.
75. Kumar, Bhargava, Tejaswini Kumar, and Swapna Nadakuditi. "Ethical Implications of Generative AI: The Case of Large Language Models." Journal of Scientific and Engineering Research 10, no. 7 (2023): 122-127.
76. Yandrapalli, Vinay. "Revolutionizing supply chains using the power of generative AI." International Journal of Research Publication and Reviews 4, no. 12 (2023): 1556-1562.
77. Agarwal, Lokesh. "Defining organizational AI governance and ethics." Available at SSRN (2023).
78. Hadi, Muhammad Usman, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, and Seyedali Mirjalili. "A survey on large language models: Applications, challenges, limitations, and practical usage." Authorea Preprints (2023).
79. Rana, Saadia Afzal, Zati Hakim Azizul, and Ali Afzal Awan. "A step toward building a unified framework for managing AI bias." PeerJ Computer Science 9 (2023): e1630.
80. Sarker, Iqbal H., Helge Janicke, Nazeeruddin Mohammad, Paul Watters, and Surya Nepal. "AI potentiality and awareness: a position from the perspective of human-AI teaming in cybersecurity." In International Conference on Intelligent Computing & Optimization, pp. 140-149.

81. Brandt, Rafaël, Daan Raatjens, and Georgi Gaydadjiev. "Precise benchmarking of explainable AI attribution methods." *arXiv preprint arXiv:2308.03161* (2023).

82. Nolan, Adrian. "Navigating Ethical Challenges in AI-Driven Decision Making for Business Consulting." (2022).

83. Alikhademi, Kiana, Brianna Richardson, Emma Drobina, and Juan E. Gilbert. "Can explainable AI explain unfairness? A framework for evaluating explainable AI." *arXiv preprint arXiv:2106.07483* (2021).

84. Nagisetty, Vineel, Laura Graves, Joseph Scott, and Vijay Ganesh. "xai-gan: Enhancing generative adversarial networks via explainable ai systems." *arXiv preprint arXiv:2002.10438* (2020).