Ethical Flaws in AI: Bias and Privacy Contemplation in the Digital Era

[1]Natasha Tomar

[1]Lecturer Computer Science and Engineering
[1]Government Polytechnic, Roorki, Baheri, Bareilly, Uttar Pradesh, India
[1]natashadept@gmail.com

Abstract— Artificial Intelligence (AI) is remodeling the area of education, transportation, commerce, health care, entertainment etc. but its pervasive adoption pioneers the serious ethical concerns. This paper analyses the two major concerns: Algorithmic Bias and Privacy Contemplation. Artificial intelligence (AI) systems are an important part of sociotechnical bionetworks that incorporate biases, societal morals and purposes reflected in multifaceted interactions between humans (like developers, users and affected participants) and technology (including algorithms, data and figuring structure). AI models trained on multilingual datasets, often inclusive of English, French, German, and Italian languages, risk unauthorized access, theft, and regulatory noncompliance with sensitive levels of personally identifiable information (PII) [1]. This research uses secondary data from peer-reviewed studies, industry reports, and academic literature to analyze these challenges. After the analysis of dataset how AI models produce different outcomes, potential solutions such as adjusting dataset sampling, incorporating fairness constraints in model design are discussed. Addressing the ethical issues is essential for ensuring fairness, accountability, and trust in AI systems.

***Index Terms*— Artificial Intelligence, Artificial Intelligence, Bias, Algorithmic fairness, Societal Bias, Privacy, Ethics.**

I. INTRODUCTION (HEADING 1)

Artificial Intelligence (AI) is one of the most fascinating technological advancements of the digital era. Various applications of AI like virtual assistants, recommendation systems, and medical diagnostics etc. AI was begun as a vision of thinking machines leading to the age of rule based expert systems and finally arising at the powerful deep learning model. On the security side, we will discuss through the various ways such as AI can be attacked from data poisoning to adverse exploits to threat against entire AI infrastructure. We will also examine how organizations can secure their network patch critical vulnerability and adapt best practices then we will set gear to deal data privacy and ethics confronting the important questions that arise when AI systems learn from and sometimes expose massive volumes for personal data. We will discuss possibility of biases, privacy risks, responsible data collection, AI driven decision making fairness and the emerging global regulations design to keep pace with their fast-evolving technology.

II. ETHICAL CONCERNS

This paper will guide through the most important ethical concerns in AI and how AI systems impact our daily lives, the risks they bring, and the principles needed to ensure that they are used ethically. Ethical flaws threaten the core principles for ethical AI such as – accountability, fairness, privacy protection, transparency, promoting positive impact.

II.A Bias in AI

AI learns from data—just as humans learn from experiences—but if the data is incomplete, unbalanced, or influenced by societal prejudices, the AI may produce unfair outcomes. AI systems are powerful but not perfect. They can reflect human biases, perpetuate inequality, or make mistakes if trained on incomplete or unbalanced data. Bias in AI can arise in several ways:

1. **Data bias:** Occurs when the training data does not represent everyone equally.
2. **Algorithmic bias:** When the AI system emphasizes certain patterns over others.
3. **Societal bias:** Embedded human prejudices influence AI outcomes.
4. **Model bias in AI systems:** Advanced models, like neural networks, may inherit biases from unrepresentative training data.

II.B Privacy apprehensions in AI

This particular concern- privacy risk, is especially prevalent in the age of artificial intelligence (AI), as sensitive information is collected and used to create and fine-tune AI and machine learning systems. Such AI privacy risks include [2]:

1. Collection of sensitive data
2. Collection of data without consent
3. Use of data without permission
4. Unchecked surveillance and bias
5. Data exfiltration
6. Data leakage

III. Literature Review

Researches done already highlights that AI systems are not fundamentally neutral. Instead, they reflect the biases exist in their training data. Researches have shown that algorithmic bias can appear in fields like as hiring, law enforcement, and healthcare [3].

While debates about privacy usually treat human observers (e.g., a nosy neighbor) or institutions (e.g., corporations) as the primary threats, the possibility that AI systems themselves might function like observers with inferential or interpretive capacities suggests new ethical terrain [4].

When generative AI (GenAI) models are prompted to create images of CEOs, they tend to reinforce stereotypes by depicting CEOs predominantly as men [5].

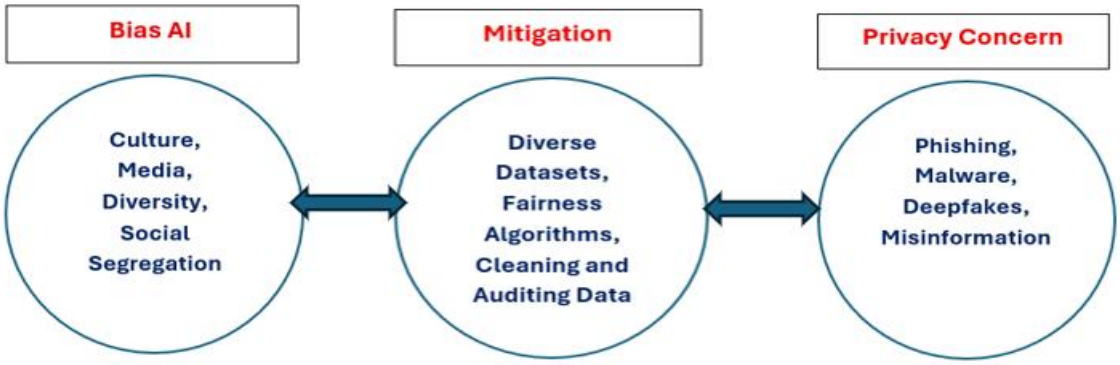
Another example of bias in AI systems is the facial recognition technology employed by law enforcement agencies. This bias can have severe consequences, including wrongful arrests or convictions. [6]

Wajiha Shahid et al. did a survey on detecting fake news spreaders which sheds light on the current state of detecting fake news spreaders. The authors categorize features into four main types: content-based, user-based, network-based, and hybrid features, which combine elements from the previous categories. [7]

IV. Methodology

While AI offers many benefits, it also opens the door to new cybersecurity threats. This research is based on secondary data collected from credible online sources, academic journals and research papers. An interpretive analysis was used to study existing works on AI ethics, with an emphasis on bias and privacy issues. Hackers and cybercriminals are increasingly targeting AI systems and using AI-based tools themselves to launch sophisticated attacks. Hence, relevant studies were selected based on their contribution to understanding ethical challenges in AI. The data was then categorized into themes such as reasons of bias (Societal), privacy risks and proposed ethical mitigations.

Figure 1: Analysis of Ethical Flaws in AI



V. Societal bias in AI

Societal bias is the root cause. A very example to deep down is - If historically fewer women were hired for tech roles, an AI trained on that data may assume: “Men are better suited for tech jobs.” This assumption comes from society, not the algorithm itself. [8] Societal bias in AI refers to the way artificial intelligence systems can reflect, reinforce, or even amplify the existing prejudices and inequalities present in human society.

V.A Sources of Bias

Bias is also a learned behavior, as humans learn from their surroundings, including family, ecological, culture, social segregation, media, diversity advantage and education. we can understand a complex system as a “system in which large networks of components with no central control and simple rules of operation give rise to complex collective behavior, sophisticated information processing, and adaptation via learning or evolution" (Mitchell 2009). Gender bias, stereotypes, least representation are the major source of history bias, culture norms, data gaps respectively.

V.B Areas affected by the Bias

Different areas get affected from basic roots like – AI in hiring process automatically by resume filtration, AI in finance get bias by loan filtration, AI in legal side can be manipulated by crime bias. Resultantly, psychological harm, economic inequality, reduced legal efficiency broadened.

V.C Impact of AI Bias

The profound impact of societal bias leads to systemic discrimination, unfair decisions, criminal justice disparities, reinforcement of inequality, healthcare inaccuracies, loss of trust and reputation, legal and ethical issues, reinforcement of stereotypes.

V.D AI BIAS MITIGATION

Solutions to societal bias in AI need a multidimensional approach compounding organizational, technical and ethical strategies. Significant mitigation approaches are –

- **Use diverse and representative data –** Data fed into deep learning systems and machine learning models must be balanced and comprehensive, representative of all the groups of people and reflective of the actual demographics of society.
- **Apply fairness techniques in Machine Learning -** Implement algorithms designed for fairness, utilizing techniques such as adversarial training to remove bias during training.
- **Build inclusive development teams -** Assemble multidisciplinary teams (ethicists, developers, social scientists) to identify biases early.
- **Increase transparency and accountability -** Conduct continuous monitoring using tools like Fairness Indicators to detect bias across various demographic groups.
- **Conduct regular bias audits and testing -** Audit data for gaps or imbalances. Use techniques like data balancing, re-weighting, and oversampling underrepresented groups to correct skewed data.

VI. PRIVACY RISKS SURROUNDING AI

As artificial intelligence endures to restructure industries and become more intensely integrated into our lives, protecting that data has become extra critical and challenging. Whereas AI provides many benefits, it also opens the door to different cybersecurity threats. Some countries have blocked the AI model Deep Seek due to privacy concerns and have launched investigations into its data collection methods.

VI.A ROLL OF DATA IN AI

Data is the foundation of AI systems. Large language Models (LLMs) rely on massive datasets, raising privacy concerns about how this data is collected and processed. Many AI firms do not disclose exactly what data they used, which leads to concern such as unauthorised data collection, lack of explicit user consent and potential exposure to sensitive information. Now a days, many regulators are imposing strict Data Protection Act to fully reveal their training datasets and implement risk assessments for AI generated content.

VI.B PRIVACY CONCERNS

Many AI models are connected to sensitive data, which makes them a target for cyberattacks. This might happen accidentally due to misconfigured systems or a deliberate cyberattack. Either way, the effects can be devastating, damaging reputations, violating privacy, and causing financial loss. Common cyberthreats in AI can be classified as below.

- **Phishing:** Tricks such as sending fake emails, messages, or even creating fake websites that look real are common practices that are used by cybercriminals to steal your personal data like your password or bank details. Attackers can use AI to craft natural-sounding and convincing messages, bypassing built-in protections in chatbots.
- **Malware:** Malicious software are harmful programs like spyware or viruses that secretly get into your computer or phone via accidentally opening suspicious emails, clicking on ads, or downloading apps from unknown websites. Generative AI can help in writing code quickly and efficiently. But that also means it can be used to create malware or insert backdoors into software.
- **Misinformation:** When AI models can hallucinate and generate made-up content. For example, an AI chatbot might assuredly provide incorrect information or cite fake sources. Attackers might also perform prompt injection or corpus poisoning, feeding false data into models to skew results. In data poisoning, attackers tamper with the training data, teaching incorrect or harmful actions to the AI model.
- **Deepfakes:** AI can replicate a person's persona, voice and appearance to produce indistinguishable videos from the real behaviors. This makes it tough to determine legitimacy, and authenticating. This has significant consequences for elections, personal safety and corporate reputation.

VI.C ETHICAL CHALLENGES IN AI

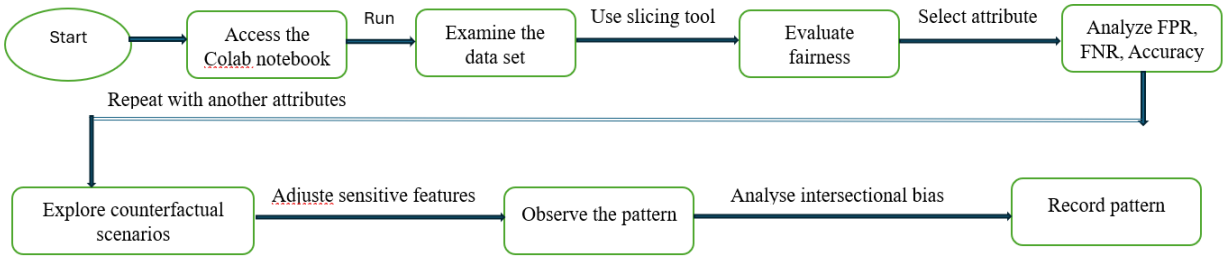
AI-powered surveillance systems increase worries about individual freedom and privacy. According to Ferrara (2023), "It is often unclear who is responsible for AI decisions because these systems involve multiple stakeholders, making accountability difficult to establish when errors or harm occur." [9]. Ministry of Electronics and IT (MeitY) is drafting a new AI governance framework that will require AI models to register and disclose data sets. Privacy violations can lead to loss of trust and potential harm to individuals and society.

VII. SOLUTIONS AND RECOMMENDATIONS

Bias in AI can be reduced by using diverse and representative datasets, applying fairness-aware algorithms, cleaning and auditing data, involving diverse development teams, conducting regular testing and ensuring transparency and ethical standards.

VII. A EXAMINING THE DATASET

FIGURE 2: STEPS TO ANALYSE THE DATASET



Artificial intelligence systems make predictions based on the data and algorithms they are trained on. If the data or modeling decisions are biased, the AI's predictions can unfairly disadvantage certain groups. To analyze, we will use the COMPAS recidivism dataset with the What-If Tool to investigate how AI models produce different outcomes across groups and explore strategies to improve fairness [10]. The sequential phases are as follows-

1. Accessed the Colab notebook: COMPAS What-If Tool Demo.
2. Run all by granting permission.
3. After execution, examined the data set and observed features such as race, gender, age, prior convictions etc.
4. Evaluated fairness by using the slicing tool to compare model predictions for different groups.
5. Selected a sensitive attribute, such as race and analyzed outcomes for Black vs. white defendants. Studied key metrics with False positive rate, False negative rate and Accuracy.
6. Repeated with another attributes such as gender or age and compared the model.
7. Explored counterfactual scenarios such as modified attributes such as age or prior convictions and note changes in predictions. Then adjusted sensitive features such as race or gender. Then observed the pattern in terms that the prediction changed reasonably? Did it reveal potential biases?
8. Analyzed intersectional bias and found- subtler biases that single-attribute evaluation may miss. Record patterns and discuss their ethical implications.

VII.B PROPOSED ETHICAL MITIGATIONS

After exploring how AI models can produce different outcomes across groups and why fairness is critical. Small changes, such as reweighting features or adjusting decision thresholds, can improve fairness but may impact accuracy. Based on your analysis, suggest actionable strategies to improve fairness:

- Adjusting dataset sampling or reweighting features
- Incorporating fairness constraints in model design
- Ongoing monitoring and model updates
- Policy recommendations for responsible AI deployment

VIII. CONCLUSION

With evolution, entire world is working to strike a balance between innovation and regulations. AI not only have technical issue but it has legal and human rights issues. AI is transforming the world in exciting ways, but it's not without its dangers. As we embrace intelligent machines in business, government, and daily life, we must also prepare for the evolving threats they bring. Phishing, malware, misinformation, and deepfakes are just the beginning.

AI helps us solve complex problems, but can also be used against us. That's why it's more important to identify cyberthreats and focus on AI security, including:

- Regular audits of AI systems to identify vulnerabilities
- Monitoring data sources for signs of poisoning or tampering
- Training employees on recognizing phishing attacks and social engineering tactics
- Implementing access controls, strong authentication, and encryption
- Using AI-based security tools to defend against AI-powered threats

We need a competent and careful approach to staying protected, using the right tools, learning to spot dangers, and following strong security rules. As future innovators, professionals, and informed citizens, we can shape AI use by advocating for fair and transparent AI systems.

REFERENCES

[1] Viswanath V, M T, Naganandh S (April 24, 2025). Artificial Intelligence and Privacy Concerns: Balancing Innovation With Security. Cureus J Comput Sci 2 : es44389-025-03689-z. doi: https://doi.org/10.7759/s44389-025-03689-z

[2] Alexandra Jonker , Alice Gomstyn, 2026. Exploring privacy issues in the age of AI. https://www.ibm.com/think/insights/ai-privacy

[3] Janet Delgado 1, Alicia de Manuel 2, Iris Parra 2, Cristian Moyano 2, Jon Rueda 3, Ariel Guersenzvaig 4, Txetxu Ausin 5, Maite Cruz 6, David Casacuberta 2, Angel Puyol 2 (2022 Jul 20). Bias in algorithms of AI systems developed for COVID-19: A scoping review. Journal of Bioethical Inquiry 19(3):407–419. doi: 10.1007/s11673-022-10200-z

[4] Christopher Register1 · Maryam Ali Khan1 · Alberto Giubilini1 · Brian David Earp1,2 · Julian Savulescu1,2 (18 October 2025). Privacy and Human-AI Relationships. Philosophy & Technology (2025) 38:147 https://doi.org/10.1007/s13347-025-00978-2

[5] Nicoletti, L., & Bass, D. (2023). Humans Are Biased: Generative AI Is Even Worse. Bloomberg, Technology and Equality. June, 23, 2023.

[6] Jeff Shuford, Vol.3, Issue 01, March 2024. Examining Ethical Aspects of AI: Addressing Bias and Equity in the Discipline. Journal of Artificial Intelligence General Science ISSN:3006-4023.

[7] W. Shahid, Y. Li, D. Staples, G. Amin, S. Hakak, and A. Ghorbani, "Are you a cyborg, bot or human? —a survey on detecting fake news spreaders," IEEE Access, vol. 10, pp. 27069–27 083, 2022.

[8] Technical series of book: AI Ethics: Bias, Responsibility, and Who Controls the Future? By Denis D. (2025) Chapter 2 page 247-268.

[9] Emilio Ferrara, (2023), Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. Sci 2024, 6, 3, https://doi.org/10.3390/sci6010003

[10] G. Yu, L. Ma, X. Wang, W. Du, W. Du, and Y. Jin, "Towards fairness-aware multi-objective optimization," Complex & Intelligent Systems 2024 11:1, vol. 11, no. 1, pp. 1–20, Nov. 2024, doi: 10.1007/S40747-024-01668-W