Assignment 2: Thematic Review On:

AI Bias, Fairness, and Privacy

Submitted
by:

Sasika Bista

Level 5

Group 5

Submitted to:

Faculty of Concept And Technologies

Of AI

Bachelor Of Computer Science

Herald College Kathmandu

January 17, 2026

Abstract

Use of AI applications across various domains for making decisions, automating tasks, and solving complex problems highlights both its potential benefits as wells as ethical challenges. This report examines three aspects of ethical AI, bias, transparency and accountability, and privacy and data governance. It explores how biases can arise from data, algorithms, and user interactions, leading to unfair or discriminatory outcomes, how transparency and accountability are essential for fostering trust in AI system, and finally analyses privacy risks and framework to develop AI systems such that they are equitably safe, and socially responsible to use. In short, this study presents the need of ethical AI for promoting trust, inclusivity, and long-term societal benefit.

Table of Contents

Introduction

Use of AI applications across various domains for making decisions, automating tasks, and solving complex problems such as employment hiring, medical diagnosis, has been on a steep incline. However, it has raised concerns about the fairness and bias of AI system and synthetic media produced by such systems. They risk reinforcing social inequalities and shaping cultural norms, making it essential to address biases in AI models.

Research by Caliskan et. Al, 2017 shows that the dataset used for training can reflect human-like biases embedded within society. Although initial privacy issues were mainly concerned with data collection and storage, with advancement in AI especially in generative AI which can reveal or mimic private information, privacy concerns now focus on full data lifecycle, from data processing, analyzing, sharing, to even deletion (Billiris et Al, 2025). While many frameworks that address privacy concerns across the full AI cycle, emphasize relationship between privacy, security, and trust, or identifying emerging ai threats exist, a study by Billiris at al.(2025) highlights how existing frameworks fail to fully capture unique and complex privacy challenges introduced by AI technologies.

The European Commission's High-Level Expert Group on AI (EU-HLEG-AI), 2019 has established guidelines regarding transparency, accountability, and social benefit, to address ethical challenges. In this report, we will examine impacts of bias, significance of these principles, and later discuss a framework based on these principles to address ethical concerns of AI development as guide a for developers and policymakers.

Algorithmic Bias and Its Impact

In AI, bias can arise from data collection, algorithm design, and interpretation made by humans. AI models can result in unfair outcomes as they can learn and mimic patterns of bias present in the data used to train them. In this section, we will explore the different sources of bias in AI and examine real-world examples of their impact.

Data bias emerges from incomplete and enormous data, algorithmic bias from biased assumption or criteria in model's algorithms, user bias from users consciously or unconsciously prejudices while interacting with the system. Fazil, A.W & Hakimi, M. & Shahidzay, A. ,2024, defines bias as a systematic error in decision-making processes that results in unfair outcomes and emphasizes on importance of careful data preprocessing is essential to ensure output quality reflects input quality.

Real-world examples highlight impacts like, COMPAS system overestimated the likelihood of a defendant reoffending among African-American defendants, even without prior convictions, while an AI system used to predict patient mortality rates was found to more likely assign higher-risk scores to African-American patients, neglecting other health factors. Facial recognition systems have misidentified minorities and generative AI bias was reported with more racial and gender bias. As such, biased algorithms can led to unfair treatment, wrong conviction, denying healthcare access, due to existing social prejudices, that permeate complex decision-making processes.

Transparency, Explainability, and Accountability

In machine learning, transparency and explainability can influence user trust, legal compliance, cultural values, and corporate ethics and are key quality requirements in AI systems (Balasubramaniam et al., 2023).

AI systems lack clarity in data inputs, decision rules, and outputs, often labeling them as "black-boxes" raising accountability concerns, especially in public sector (Saldanha et al., 2022). Transparency enables stakeholders, including citizens, regulators, and civil society to monitor AI systems, evaluate outcomes, and ensure ethical behavior. Feng & Chandra (2025), studies show that providing technical transparency, such as sharing source code or model architecture, can increase system adoption by 65%, while ecosystem transparency whereas, covering development, management, and operational processes can raise adoption by 157%

Explainability defines how a system communicates its decisions. Köhl et al. (2023) defines a system is explainable if an entity (explainer) provides information enabling an addressee to understand a specific aspect of the system in context. Chazette et al. (2023), further expands this definition by highlighting the need to clearly identify who needs explanations, what should be explained, and how explanations should be delivered. Effective explainability is essential as it supports trust, accountability, and informed decision-making.

Similarly, accountability policies act as trust-building mechanisms, especially to groups that are less fluent about technology. Accountability is important to address both moral responsibility and operational responsibility. For this reason, it is important for policymakers to ensure that AI algorithms and their applications are accountable.

A study by Feng and Chandra, 2025, compares China and Western countries approach for AI accountability. The authors found that in China, where AI accountability focuses on hierarchical oversight, performance targets, and regulatory compliance, its citizens viewed AI more optimistically than citizens from Westerns countries, which focused more on public participation, and access to information. Their study explains how transparency and accountability differ across different countries due to cultural and governance environment. Hence, we can conclude that

as much as accountability and transparency in AI are essential, their implementation must differ based on local governance and cultural expectations

Privacy and Data Governance

AI systems pose complex privacy challenges across the entire data lifecycle, including collection, storage, processing, analysis, sharing, and deletion. Vulnerabilities arise from model architecture, technical infrastructure, human factors, and data handling practices, making traditional security measures insufficient (Billiris, Gill & Bandara, 2025). Generative AI adds further risks by creating synthetic data that may reveal private information.

To address these challenges, responsible data governance and Privacy-by-Design frameworks are essential. Extract, Perform, and Load (ETL) pipelines can embed ethics and governance logic to ensure transparency, accountability, and traceability, while techniques such as anonymization, pseudonymization, aggregation, and differential privacy protect sensitive data during processing and storage. Regulatory compliance, including GDPR principles of transparency, lawfulness, and the right to explanation (Wachter et al., 2017) and HIPAA safeguards in healthcare (Momani, A. , 2025) further reinforce ethical AI deployment.

Discussion

Through this research, I have learnt how AI systems reflect on the data they are trained on and that, bias within the training data, practitioners working on the model, and users can often influence AI outcomes, that can be hidden from both developers and users. Reading through real-life cases like COMPAS criminal justice tool and biased healthcare prediction systems has made me realize how impactful these biases can be to an individuals life. This has helped me realize the need of careful data processing, diverse training data, and proper model evaluation to ensure fairness in AI systems.

Further analysis on transparency, explainability, and accountability has made me aware how crucial these elements are for developing trust in AI systems, especially among those that aren't fluent in technologies. From the study of Feng and Chandra (2025), I learn that it is essential for policies regarding AI to be based on the cultural and governance environment to ensure easy adaptability among their citizens. This shows that ethical AI must be adaptable policies cannot simply be copied across regions.

Finally, privacy and data governance are foundational to ethical AI. AI's power comes from processing vast amounts of data, which creates risks at every stage. Implementing Privacy-by-Design, responsible ETL pipelines, and regulatory compliance with GDPR and HIPAA demonstrates how ethical frameworks can reduce harm, protect individuals, and foster public trust. Overall, I believe that addressing bias, enhancing transparency, and embedding privacy are central to creating AI that is fair, trustworthy, and beneficial to society.

References

Ferrara, E. (2024). 'Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies', Sci, 6(1), p. 3. 10.3390/sci6010003

EU-HLEG-AI (2019) Ethics guidelines for trustworthy AI. European Commission,9-04. Caliskan, A., Bryson, J.J. and Narayanan, A. (2017) 'Semantics derived automatically from language corpora contain human-like biases', Science, 356(6334), pp. 183–186. 10.1126/science.aal4230

Fazil, A.W & Hakimi, M. & Shahidzay, A. (2024). 'A Comprehensive Review of Bias in AI Algorithms. Nusantara Hasana Journal'. 3. 1-11. 10.59003/nhj.v3i8.1052

Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K. and Kujala, S. (2023) Transparency and explainability of AI systems: From ethical guidelines to requirements', Information and Software Technology, 159, p. 107197. 10.1016/j.infsof.2023.107197

Saldanha, D. M. F., C. Nogueira Dias, and S. Guillaumon. 2022. "Transparency and Accountability in Digital Public Services: Learning From the Brazilian Cases." Government Information Quarterly 39, no. 2: 101680.

Feng, N. and Chandra, Y. (2025) 'Accountability in government use of AI: Citizen concerns and preferences', Public Administration.10.1111/padm.70030.

Chazette, L., Brunotte, W. & Speith, T., 2021. Exploring explainability: a definition, a model, and a knowledge catalogue. In: Proceedings of the International Requirements Engineering Conference. pp. 197–208.

Köhl, M.A., Baum, K., Langer, M., Oster, D., Speith, T. & Bohlender, D., 2019. Explainability as a non-functional requirement. In: Proceedings of the International Requirements Engineering Conference. pp. 363–368.

Billiris, Grace & Gill, Asif & Bandara, Madhushi. (2025). Privacy in the Age of AI: A Taxonomy of Data Risks. 10.48550/arXiv.2510.02357.

Youseff, S., Philip, A., Hamzah, F., Barnanas, B., John, B., Alonge, M., Oluremi, D., Adekola, P., Liang, W., Blessing, M., Oluwaferanmi, A., Rajoy, A., Sannareddy, S. & Barnty, B., 2026. Ethical Data Engineering in ETL Pipelines: Aligning Intelligent Governance with Responsible AI and Privacy-by-Design.

Wachter, S., Mittelstadt, B. & Floridi, L., 2017. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. SSRN Electronic Journal. 10.1093/idpl/ipx005.

Momani, A., 2025. Artificial intelligent implications on health data privacy and confidentiality. arXiv. 10.48550/arXiv.2501.01639

Plagiarism report:

Similarity Report

PAPER NAME **2501642_SasikaBista-3.pdf**	AUTHOR -
WORD COUNT **1375 Words**	CHARACTER COUNT **8858 Characters**
PAGE COUNT **8 Pages**	FILE SIZE **201.3KB**
SUBMISSION DATE **Jan 16, 2026 1:59 PM GMT+5:45**	REPORT DATE **Jan 16, 2026 1:59 PM GMT+5:45**

20% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 7% Internet database
- 17% Publications database
- Crossref database
- Crossref Posted Content database
- 10% Submitted Works database