

Ethical Challenges in Using AI for Decision-Making

Table of contents:

Research Summaries2

Critical Analysis3

Connections3

Conflicts between the findings:3

Synthesis Matrix for Literature Review:3

References:5

Research Summaries

1. **Paper 1: Big Data's Disparate Impact**

This study shows how big data can lead to unfair AI decisions by reflecting existing social biases. For example, algorithms might use neutral-seeming data (like zip codes) to discriminate based on race or gender. The paper argues that current anti-discrimination laws are outdated for addressing algorithmic bias and calls for stronger regulations and transparency in AI systems.

2. **Paper 2: The Ethics of Algorithms**

This research categorizes ethical issues into five areas: bias, responsibility, transparency, explainability, and accountability. It criticizes "black box" AI systems, where decisions are unclear, and warns that biased training data can lead to unfair outcomes—even without intentional discrimination. The study stresses the need for collaboration between AI developers, ethicists, and policymakers.

3. **Paper 3: AI4People – An Ethical Framework**

Proposes a framework based on four principles: fairness, transparency, accountability, and sustainability. For example, medical AI should explain how it diagnoses illnesses. However, the framework lacks practical steps for implementation and needs more testing in real-world scenarios.

4. **Paper 4: The Global Landscape of AI Ethics Guidelines**

Analyzes 84 AI ethics guidelines worldwide. Most focus on transparency and fairness but struggle with enforcement. For instance, the EU has strict laws, while China uses flexible policies. The study urges governments and companies to move from theory to action but offers no concrete solutions.

5. **Paper 5: A Survey on Bias and Fairness in Machine Learning**

Explore sources of bias (e.g., unbalanced training data) and solutions like improving data collection. However, measuring fairness remains a challenge—there's no universal standard. For example, facial recognition systems often fail for darker-skinned individuals due to biased datasets.

6. Paper 6: Fairness and Abstraction in Sociotechnical Systems

Argues that oversimplifying data (e.g., using income alone to assess creditworthiness) ignores social factors like historical discrimination. The study calls for AI systems to consider social complexities but lacks practical tools for implementation.

Critical Analysis

The studies agree on key ethical issues (bias, transparency, accountability) but differ in approach:

- Some focus on legal reforms (e.g., Barocas & Selbst), while others propose ethical frameworks (e.g., Floridi et al.).
- Solutions are often theoretical, with few real-world applications.
- Debates exist between ensuring fair *processes* (procedural fairness) vs. fair *outcomes* (results-based fairness).

Connections

- All studies highlight fairness and transparency but disagree on how to achieve them.
- Papers 1, 3, and 5 focus on technical fixes (e.g., better data), while 4 and 6 emphasize social context.

Conflicts between the findings:

- Some studies (e.g., Mehrabi et al.) trust technical solutions, while others (e.g., Selbst et al.) argue that social context must be included.
- Floridi et al. suggest a universal ethical framework, but Jobin et al. show that global policies vary widely.

Synthesis Matrix for Literature Review:

Study	Year	Authors	Method	Key Findings	Limitations
Big Data's Disparate Impact	2016	Barocas & Selbst	Legal analysis	Shows how AI reinforces social biases through data	No practical solutions Strengths: Provides a comprehensive legal analysis of how AI can govern social ideas, highlighting legal and regulatory gaps. Weaknesses: Provides a clear legal analysis and there are practical solutions as it does not rely on

Study	Year	Authors	Method	Key Findings	Limitations
					a lot of data or experiments to inform its conclusions.
The Ethics of Algorithms	2016	Mittelstadt et al.	Literature review	Identifies 5 ethical challenges (e.g., bias)	Lacks quantitative data and practical experiments to support its analysis. Strengths: Provides a thorough classification of the ethical challenges associated with algorithms, and highlights the responsibility of developers and policymakers. Weaknesses: Lacks quantitative data and practical experiments to support its analysis, making its findings more theoretical than applied.
AI4People	2018	Floridi et al.	Framework proposal	Ethical principles for AI (fairness, transparency)	Lacks detailed implementation mechanisms and practical applications. Strengths: Proposes a clear ethical framework that can be applied across multiple domains to ensure responsible use of AI. Weaknesses: Does not provide detailed implementation mechanisms and practical applications, and has not been tested in real-world settings.
Global AI Ethics Guidelines	2019	Jobin et al.	Policy analysis	Global guidelines lack enforcement	It does not address the impact of the principles on real-world applications of AI. Strengths: It provides a comprehensive analysis of global trends in AI ethics, which helps understand gaps in different regulations. Weaknesses: The research does not address the impact of ethical principles on real-world applications, nor does it provide clear recommendations to bridge the gap between policy and practice.
Bias and Fairness Survey	2021	Mehrabi et al.	Literature review	Sources of bias and fixes	It does not discuss how well the solutions work when applied in different settings. Strengths: Provides a detailed analysis of the different types of bias in AI, and reviews the

Study	Year	Authors	Method	Key Findings	Limitations
					solutions available to address them. Weaknesses: It does not discuss how well these solutions work when applied in different settings, nor does it provide a unified criterion for measuring fairness in different models.
Fairness and Abstraction	2019	Selbst et al.	Theoretical analysis	Oversimplified data causes unfair decisions	Lacks empirical evidence to support its conclusions. Strengths: The paper brings out the consequences of abstraction within technical systems on justice and articulates areas that have only been underexplored in extant literature. Weaknesses: Fails to provide any practical solutions to address the problem and it largely lack empirical evidence to back its conclusion.

References:

1. Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, 104(3), 671-732. <https://doi.org/10.15779/Z38BG31>
2. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, 3(2), 1-21. <https://doi.org/10.1177/2053951716679679>
3. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society. *Minds and Machines*, 28(4), 689-707. <https://doi.org/10.1007/s11023-018-9482-5>
4. Jobin, A., Ienca, M., & Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, 1(9), 389-399. <https://doi.org/10.1038/s42256-019-0088-2>
5. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)*, 54(6), 1-35. <https://doi.org/10.1145/3457607>
6. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)*, 59-68. <https://doi.org/10.1145/3287560.3287598>