AI Ethics Casefiles — Report

Augusto Mate | 2025

1. Introduction

Artificial Intelligence is increasingly used to inform decisions in domains such as hiring, healthcare, finance, and criminal justice. While these systems offer efficiency and scale, they also raise critical ethical concerns, particularly around bias, fairness, and explainability.

This report analyzes two real-world AI systems that exhibited ethical issues and discusses them in light of responsible AI principles. A demo notebook and slide deck accompany the analysis for applied illustration.

2. Case Study 1 — Gender Bias in Hiring System

Context

A company developed an automated resume screening system trained on past hiring data. Although gender was not an explicit feature, the model learned to favor male candidates by identifying proxies, such as male-dominated job titles and pronouns.

Ethical Concerns

- **Proxy Bias**: Excluding protected attributes doesn’t prevent discrimination if proxies remain.
- **Opacity**: Stakeholders could not understand or challenge decisions.
- **Disparate Impact**: Qualified female candidates were unfairly penalized.

Discussion

The system reflects historical discrimination embedded in past hiring data. Removing explicit gender features proved insufficient to eliminate bias. Lack of transparency and fairness audits contributed to deployment risks.

3. Case Study 2 — Bias in Medical Risk Prediction

Context

A health analytics system predicted patient risk levels to allocate care management. The model significantly underpredicted risk for Black patients, due to reliance on healthcare cost as a proxy for health need.

Ethical Concerns

- **Historical Bias**: Data reflected systemic underinvestment in Black communities.
- **Measurement Bias**: Cost does not equal care need.
- **Unintended Harm**: Patients most in need received fewer resources.

Discussion

This case highlights how “race-neutral” designs can encode structural inequality. Without community input or contextual validation, algorithms can deepen disparities rather than correct them.

---

4. Practical Demonstration

A simplified demo (`bias-demo.ipynb`) replicates the concept of bias using a toy dataset. The logistic regression model trained on gender and scores illustrates how bias emerges even in minimal settings.

> Educational tools like this help surface ethical questions in tangible ways.

---

5. Transparency and Explainability

Both cases demonstrated limited transparency:

- Models were black-box systems
- Stakeholders lacked interpretability tools
- No model cards or explanations were shared

Efforts such as LIME, SHAP, and model cards can support transparency, but require organizational support and ethical intent to be meaningful.

6. Recommendations

- Perform **bias audits** on datasets and models before deployment
- Ensure **interpretability tools** are accessible to affected stakeholders
- Engage **cross-disciplinary teams**, including ethicists and impacted communities
- Create and publish **model documentation** (e.g. datasheets, cards)

7. Limitations

- The analysis is based on publicly reported cases
- No access to original model code or datasets
- The demo uses a synthetic, small-scale dataset
- Focuses on bias and transparency; other principles (e.g. autonomy, justice) are not covered in depth

8. References

- Obermeyer et al. (2019). *Dissecting racial bias in an algorithm used to manage the health of populations*. Science.
- Crawford, K. (2017). *The Trouble with Bias* — NIPS Conference Keynote
- Barocas, Hardt, & Narayanan. (2023). *Fairness and Machine Learning*

9. Appendix

See accompanying materials:

- `src/bias-demo.ipynb`: toy model illustrating algorithmic bias
- `slides/ai-ethics-casefiles.md`: slide deck summarizing key findings
- `README.md`: overview and setup instructions