Designing Responsible and Fair AI Systems

AI Ethics Assignment Report

Author: Amanuel Alemu Zewdu

Group Members: [Add group member names here]

1. Introduction

This report examines ethical principles for designing responsible and fair AI systems. The focus is on bias identification, fairness metrics, and mitigation strategies. We analyze the COMPAS case study and provide a practical fairness audit and recommendations.

2. Case Study: COMPAS

The COMPAS risk assessment tool has been shown to produce disparate outcomes across racial groups. Key issues include overprediction for some groups and differences in false positive rates. We recommend auditing models, increasing transparency, and using fairness-aware methods.

3. Fairness Audit Plan

We propose a reproducible audit using AI Fairness 360 and complementary metrics. Steps include loading data, computing disparate impact and equal opportunity difference, visualizing group-level metrics, applying mitigation, and documenting results for stakeholders.

4. Ethical Frameworks & Governance

Adopt frameworks such as the EU Ethics Guidelines for Trustworthy AI, emphasizing human oversight, robustness, privacy, transparency, and accountability. Implement model cards and datasheets for datasets.

5. Proposed Solutions

Mitigation strategies: pre-processing (reweighing), in-processing (fairness constraints), post-processing (threshold adjustments), and operational controls (monitoring, human-in-the-loop).

6. Reflection & Peer Review Plan

Peer reviews will focus on replicability and metric interpretation. Assign roles for data, modeling, and documentation. Maintain logs of experiments and decisions to ensure accountability.

References

- Bellamy et al., AI Fairness 360 (2018).
- European Commission, Ethics Guidelines for Trustworthy AI (2019).
- ProPublica, Machine Bias (2016).