AAI301 Student Study Guide

Seminar 2: Bias, Fairness, Responsibility, and Accountability

Overview

This guide is designed to help you understand and apply the core concepts covered in **AAI301 Seminar 2**. If you missed the seminar or need a refresher, this document provides the essential theoretical foundations, case studies, and practical activities to ensure you are up to speed on the ethical implications of AI bias and the frameworks for accountability.

Learning Objectives

By the end of this session, you should be able to:

- Identify different sources and types of **bias** in Generative AI.
- Analyze the concept of **fairness** in the context of AI outputs (Group, Individual, and Procedural).
- Apply technical and governance-based **mitigation strategies** to reduce bias.
- Understand the challenges of assigning **responsibility and accountability** for the actions of Generative AI systems.

Key Concepts

1. What is AI Bias?

Bias refers to systematic favoritism or discrimination encoded in AI outputs. It results in some demographics being unfairly favored or disadvantaged.

Where does it come from?

- **Data Bias:** Present in the training data (e.g., **Representation Bias** where datasets skew towards specific demographics, or **Historical Bias** reflecting past societal prejudices).
- **Algorithmic Bias:** Introduced during model design or optimization goals.
- **Labelling Bias:** Introduced by human annotators (e.g., labeling images of doctors primarily as "male").
- **User Bias:** Introduced by how humans interact with the AI (e.g., a user prompts for a "criminal story" and accepts a racially biased output without challenging it).

2. Real-World Impact of Bias

In class, we discussed specific examples of how GenAI perpetuates stereotypes:

- **Professional Stereotypes:** Tools like DALL-E or Stable Diffusion often generate white males for prompts like "CEO" but women for "Nurse."
- **Criminal Imagery:** AI generators disproportionately depict darker-skinned individuals

when prompted with words like "inmate," reinforcing racial profiling.
- **Cultural Bias:** Western perspectives often dominate; for example, local tests showed LLMs stereotyping Chinese individuals as "money-minded" and Malays as "laid back."

3. Defining Fairness

Fairness ensures AI systems treat individuals and groups equitably. We explored three types:

- **Group Fairness:**
  - *Equal Opportunity:* Ensuring qualified individuals from all groups have an equal likelihood of a positive outcome (e.g., an AI hiring tool giving talented candidates from all universities a fair chance).
  - *Demographic Parity:* Ensuring outcomes are statistically equal across groups, regardless of input qualifications (e.g., approving loans for 50% of Group A and 50% of Group B to ensure equity).
- **Individual Fairness:** Treating similar individuals similarly (e.g., two people with the same credit score should get the same loan rate).
- **Procedural Fairness:** Ensuring the process behind the decision is transparent and justifiable.

4. Responsibility vs. Accountability

- **Responsibility:** The moral and legal obligation to ensure AI acts ethically (Developers must design safely; Users must use ethically).
- **Accountability:** The obligation to explain, justify, and take ownership of outcomes.
  - *Challenge:* The **"Black Box"** nature of AI makes it hard to trace why a decision was made.
  - *Challenge:* **Diffusion of Responsibility** occurs when developers blame users, and users blame the system.

Practical Exercises (Catch-Up Activities)

Exercise 1: Bias Detective (Visual Analysis)

- **Context:** AI tools often hallucinate or embed subtle biases in images.
- **Activity:** Visit Google's Art & Culture Experiments (Odd One Out) to see if you can distinguish between real art and AI-generated imitations.
- **Reflection:** How might these subtle visual artifacts mislead a user in a news context?

Exercise 2: Fair Policing Systems (Case Study)

- **Scenario:** A predictive policing algorithm assigned more patrols to "Neighborhood Blue" based on higher reported crime statistics than "Neighborhood Orange."
- **The Context:** "Neighborhood Blue" crimes were mostly "discovered" crimes (like drug stops) caused by *proactive* policing, whereas "Neighborhood Orange" had reported burglaries.
- **Critical Thinking:** By sending more police to Blue based on this data, the AI creates a **feedback loop**—more police find more crime, justifying more police. This violates

fairness by ignoring the context of the data.

Exercise 3: "Who Is Responsible?" (Role-Play)

- **Scenario:** A university chatbot gave a student incorrect academic advice, delaying their graduation.
- **Analysis:**
  - **The Developer:** Responsible for the accuracy of the training data (university documents).
  - **The University:** Responsible for deploying a tool without sufficient "human in the loop" oversight for critical advice.
  - **The Student:** Responsible for verifying critical graduation requirements with a human advisor?
- **Takeaway:** Accountability is often shared, but governance frameworks must define who takes the final blame.

Preparation for Next Class (Seminar 3)

In Seminar 3, we will explore the dangers of **Misinformation, Transparency, and Trustworthiness**.

- **Topic:** Misinformation, Disinformation, and Deepfakes.
- **Preview:** We will look at how Generative AI is weaponized to create:
  - **Deepfakes:** From the "Pope in a puffer jacket" to the unauthorized "Morgan Freeman" video.
  - **Voice Cloning:** The ethical dilemma of the "JFK Unsilenced" project.
  - **Transparency:** Why we need "Explainable AI" (XAI) to trust the systems we use.

Self-Assessment Questions

1. What is the difference between **Equal Opportunity** and **Demographic Parity** in Group Fairness?
2. How does **User Bias** contribute to a feedback loop in Generative AI models?
3. Why does the **"Black Box"** problem make it difficult to assign accountability in AI accidents?
4. List one strategy to mitigate bias at the **Data Collection** stage.

Need Help?

If you have trouble understanding the fairness metrics or the distinction between responsibility and accountability:

- **Email:** Rudy005@suss.edu.sg
- **Office Hours:** [Weekdays / Between 12pm to 1pm]
- **Discussion Forum:** [Link to be shared later]