AAI301 Student Study Guide

Seminar 4: Nonmaleficence, Beneficence, Privacy, Liability, and Intellectual Property

Overview

This guide is designed to help you understand and apply the core concepts covered in **AAI301 Seminar 4**. If you missed the seminar or need a refresher, this document provides the essential theoretical foundations, legal frameworks, and practical discussions to ensure you are up to speed on the complex ethical landscape of AI development.

Learning Objectives

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

- Identify how **Nonmaleficence** (minimizing harm) and **Beneficence** (maximizing benefits) are core ethical mandates for Gen AI.
- Evaluate the **privacy implications** of training and using Generative AI models.
- Analyze the challenges Generative AI poses to existing **intellectual property (IP)** frameworks.
- Discuss potential solutions for protecting privacy and IP in the age of Generative AI.

Key Concepts

1. Nonmaleficence (Do No Harm) vs. Beneficence (Do Good)

- **Nonmaleficence:** The obligation to avoid causing harm (intentionally or unintentionally).
    - *Risks:* Misinformation, bias, job displacement, privacy violations, and environmental impact.
    - *Example:* An AI diagnostic tool trained on biased data misdiagnosing a patient (Harm).
- **Beneficence:** The obligation to actively promote well-being and positive outcomes.
    - *Benefits:* Drug discovery, personalized education, accessibility tools for the disabled.
    - *Example:* AI accelerating the discovery of drugs for rare diseases (Benefit).

2. Privacy in the Age of AI

Privacy is the right to control personal information. Gen AI creates new risks:

- **Identity Theft:** AI generating fake IDs or synthetic identities.
- **Deepfake Scams:** Using cloned voices to trick family members or employees (e.g., the Arup finance case).
- **Data Leakage:**
    - *Training Data Memorization:* Models inadvertently outputting PII (Personally Identifiable Information) they were trained on.

    - *Inference Attacks:* Attackers reverse-engineering a model to extract sensitive training data.

3. Liability: Who is Responsible?

Determining who to blame when AI causes harm is difficult due to:

- **The "Black Box" Problem:** We often don't know *why* an AI made a specific decision.
- **Distributed Responsibility:** Is it the developer (code flaw), the data provider (bad data), or the user (malicious prompt)?
- **Emergent Properties:** AI systems developing capabilities they were not explicitly programmed for.

4. Intellectual Property (IP) Challenges

- **Copyright:** Does training an AI on copyrighted books/art constitute "Fair Use"?
    - *Case Study:* **Zarya of the Dawn**. The US Copyright Office ruled that while the text and arrangement were human-created and protected, the individual AI-generated images (made with Midjourney) were **not** copyrightable because they lacked human authorship.
- **Trademarks:** AI can accidentally generate logos that infringe on existing brands (e.g., a fake Adidas logo).
- **Patents:** Can an AI be an inventor? (Currently, most courts say "No," inventors must be human).

Practical Exercises (Catch-Up Activities)

Exercise 1: Balancing Harm and Good

- **Scenario:** An AI tool can write perfect essays.
- **Nonmaleficence Check:** It encourages plagiarism and devalues original work.
- **Beneficence Check:** It can help non-native speakers improve their writing and overcome language barriers.
- **Discussion:** How do we design the tool to maximize the benefit while minimizing the harm? (e.g., Watermarking AI text).

Exercise 2: The "Zarya of the Dawn" Case Study

- **Context:** A comic book created using Midjourney.
- **Key Takeaway:** The "human authorship" requirement is strict. Prompting an AI is not currently considered sufficient "creative control" for copyright in the US.
- **Reflection:** If you use AI to generate assets for a project, do you own those assets? (Likely not, under current US/Singapore interpretations).

Exercise 3: Privacy by Design

- **Concept:** Integrating privacy into the core design of AI systems.
- **Techniques:**
    - *Data Minimization:* Only collecting what is strictly necessary.

    - *Federated Learning:* Training models on user devices without sending raw data to a central server.
    - *Differential Privacy:* Adding "noise" to datasets so individual users cannot be identified.

Preparation for Next Class (Seminar 5)

In Seminar 5, we will explore the human side of the equation—how AI affects our ability to make choices and the broader impact on society.

- **Topic:** Human Autonomy and Societal Impact.
- **Preview:** We will discuss:
    - **Human Autonomy:** Does relying on AI for decisions (e.g., GPS, movie recommendations) erode our ability to think for ourselves?
    - **Human-in-the-Loop (HITL):** Why keeping a human in the decision process is critical for ethics.
    - **Societal Impact:** The long-term effects on employment, democracy (filter bubbles), and culture.

Self-Assessment Questions

1. Explain the difference between Nonmaleficence and Beneficence with an AI example for each.
2. Why did the US Copyright Office deny copyright protection for the images in "Zarya of the Dawn"?
3. What is an "Inference Attack" in the context of AI privacy?
4. Why is the "Black Box" nature of AI a challenge for assigning legal liability?

Need Help?

If you are struggling with the legal nuances of IP or the technical aspects of privacy preservation:

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