AAI301 Student Study Guide

Seminar 3: Misinformation, Transparency, and Trustworthiness in Generative AI

Overview

This guide is designed to help you understand and apply the core concepts covered in **AAI301 Seminar 3**. 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 challenges of trust and truth in the age of Generative AI.

Learning Objectives

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

- Analyze how Generative AI can be used to create and disseminate **misinformation** and **disinformation**.
- Evaluate the importance of **transparency** in Generative AI systems.
- Explain the role of transparency and explainability in building **trust** and ensuring accountability.
- Understand how **trustworthiness** ensures AI tools operate reliably, safely, and ethically.

Key Concepts

1. Misinformation vs. Disinformation

- **Misinformation:** False or misleading information created or shared *without* intent to deceive. (e.g., A user unintentionally shares an AI hallucination believing it to be true).
- **Disinformation:** False or misleading information created or shared *intentionally* to deceive. (e.g., Weaponized deepfakes used for political propaganda).
- **Role of Gen AI:** The scale and speed of AI generation amplify the potential for harm, making it easier to flood the internet with believable falsehoods.

2. Deepfakes: The Era of Synthetic Reality

Deepfakes use deep learning (GANs/Autoencoders) to create realistic synthetic media.

- **Positive Uses:** De-aging actors in films (e.g., Harrison Ford in *Indiana Jones*), restoring voices for those with speech impairments (e.g., *JFK Unsilenced* project).
- **Negative Uses:** Non-consensual pornography, political disinformation (e.g., fake video of a leader surrendering), and CEO voice cloning scams.

3. The Psychology of Belief

Why do we fall for AI fakes?

- **Confirmation Bias:** We are more likely to believe fakes that align with our existing beliefs.

- **Visual Supremacy:** We are evolutionarily wired to trust what we see ("Seeing is believing"), a trust that deepfakes exploit.
- **The "Liar's Dividend":** The mere existence of deepfakes allows bad actors to dismiss *real* evidence as fake, eroding trust in all media.

4. Transparency in AI

Transparency means being upfront about how an AI system works.

- **Explainability:** Can we understand *why* the AI made a specific decision?
- **Data Transparency:** Do we know what data the model was trained on?
- **Model Transparency:** Is the architecture and logic of the model disclosed?
- *Risk:* "Black Box" models (opaque decision-making) lead to erosion of trust and accountability.

5. Trustworthiness

Trust is earned when an AI system is:

- **Reliable:** Consistently accurate.
- **Safe:** Prevents harm (physical or digital).
- **Fair:** Free from harmful bias.
- **Accountable:** Clear lines of responsibility for errors.

Practical Exercises (Catch-Up Activities)

Exercise 1: Spot the Fake (Visual Literacy)

- **Context:** AI image generators often leave subtle artifacts (e.g., unnatural hands, asymmetrical eyes, weird text).
- **Activity:** Test your skills at Google's "Odd One Out". Can you distinguish the AI-generated artifact from the real ones?
- **Reflection:** How confident were you? Did you find yourself doubting the real images?

Exercise 2: "JFK Unsilenced" (Audio Synthesis)

- **Case Study:** Review the project where AI recreated JFK's voice to deliver the speech he never gave in Dallas.
- **Discussion:**
  - **Pros:** Historical preservation, emotional connection.
  - **Cons:** Ethical concerns about consent (using a deceased person's voice) and the potential for misuse in rewriting history.

Exercise 3: The "Morgan Freeman" Deepfake (Puppet Master)

- **Case Study:** Watch the viral deepfake video where "Morgan Freeman" (actually a Dutch actor with an AI face swap) warns about "synthetic reality."
- **Critical Thinking:** If a deepfake can explicitly tell you it is fake and *still* look convincing, what does that mean for video evidence in courts or news?

Exercise 4: Detecting Deepfakes

- **Techniques discussed:**
  - **Forensic Analysis:** Looking for pixel-level inconsistencies.
  - **Biometric Glitches:** Unnatural blinking patterns or lack of micro-expressions.
  - **Contextual Analysis:** Cross-referencing the content with trusted sources (Does the weather in the video match the weather on that day?).

Preparation for Next Class (Seminar 4)

In Seminar 4, we will tackle the legal and safety boundaries of AI.

- **Topic:** Nonmaleficence, Privacy, Liability, and Intellectual Property.
- **Preview:** We will explore:
  - **Privacy:** How AI can infer sensitive data about you from non-sensitive inputs.
  - **Liability:** If an autonomous AI causes harm, who pays? The developer, the user, or the AI itself?
  - **IP Rights:** Who owns the copyright to AI art? The "Zarya of the Dawn" comic book case study.

Self-Assessment Questions

1. What is the key difference between misinformation and disinformation in the context of AI?
2. Name one positive and one negative application of Deepfake technology.
3. Why is the "Black Box" problem a threat to AI transparency?
4. How does the "Liar's Dividend" affect public trust in genuine news media?

Need Help?

If you are struggling with the concepts of GANs or the ethical nuances of transparency:

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