AAI301 Student Study Guide

Seminar 6: Building an Ethical AI Future: Governance, Regulation, and Looking Ahead

Overview

This guide is designed to help you consolidate the concepts covered in AAI301 Seminar 6. If you missed the seminar or need a refresher, this document provides the essential theoretical foundations regarding AI governance, regulation, and the future trajectory of Generative AI. It also includes practical thought exercises to test your application of these concepts.

Learning Objectives

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

- **Discuss** different approaches to the governance and regulation of Generative AI (Self-regulation vs. Hard Law).
- **Explore** the future to look at the broader societal implications of widespread Generative AI adoption.
- **Analyze** emerging trends and potential future applications of Generative AI, considering our collective responsibilities.

Key Concepts

1. The Need for Governance and Regulation

As Gen AI advances, oversight is necessary to prevent misuse and protect societal interests. We distinguish between three levels of control:

- **Self-Regulation (Industry Standards):** Companies establish their own internal ethical guidelines and ethics boards.
  - *Example:* OpenAI or Google's internal AI principles.
- **Soft Law (Frameworks & Guidelines):** Non-binding recommendations from international bodies or governments.
  - *Example:* OECD Principles on AI, Singapore's AI Verify.
- **Hard Law (Legislation):** Legally binding laws with enforcement mechanisms.
  - *Example:* The EU AI Act, Singapore's POFMA.

2. Core Focus Areas: Governance vs. Regulation

We analyzed four specific areas where ethical AI practice is critical. Note the difference between *Governance* (internal/corporate) and *Regulation* (external/legal).

Focus Area	Governance Strategies	Regulatory Approaches

	(Corporate)	(Legal)
Bias & Discrimination	**Continuous Monitoring:** Internal audits to detect bias. **Ethical Principles:** Adopting frameworks like "FAST Track" (Fairness, Accountability, Sustainability, Transparency).	**Mandatory Assessments:** Requiring impact assessments for high-risk AI. **Transparency:** Mandating clear explanations of AI decisions (e.g., Brazil's proposed AI Act).
Misinformation	**Responsible Guidelines:** Internal codes to prevent disinformation. **Red Teaming:** Proactively attacking one's own model to find vulnerabilities.	**Watermarking:** Laws requiring AI content to be labeled. **Liability:** Legal responsibility for harmful content (e.g., EU AI Act, POFMA).
IP & Copyright	**Ethical Sourcing:** Licensing training data properly. **Attribution:** Features that credit source material where feasible.	**Human Authorship:** Copyright only for works with significant human input. **Transparency:** Disclosing training data summaries (e.g., EU AI Act Art. 28b).
Privacy	**Privacy by Design:** Integrating safeguards from the start. **Data Minimization:** Only collecting what is necessary.	**Explicit Consent:** Laws requiring permission to use personal data for training (GDPR). **Cross-Border Rules:** Regulating international data transfer.

3. Key AI Trends (2025 and Beyond)

Generative AI is evolving from simple text generation to more complex capabilities:

- **AI Reasoning:** Models performing multistep problem-solving and nuanced analysis (e.g., OpenAI o1).
- **Agentic AI:** Autonomous systems that handle entire workflows (e.g., scheduling, booking) rather than just answering questions.
- **Multimodality:** Systems processing text, audio, image, and video simultaneously (e.g., Gemini 2.0 Flash, Claude 3.5).

4. The Future of AI: Promise vs. Peril

Thought leaders are divided on the long-term trajectory of AI:

- **The Optimists (e.g., Mustafa Suleyman, Demis Hassabis):**
  - **Healthcare Revolution:** Potential to cure all diseases and extend human lifespan.
  - **Abundance:** Solving global challenges like climate change and energy.
  - **Personal Companions:** AI that supports and organizes our daily lives.
- **The Pessimists (e.g., Geoffrey Hinton, Yoshua Bengio):**
  - **Loss of Control:** Risks of "Superintelligent" AI pursuing goals misaligned with human values.
  - **Existential Risk:** Potential for human extinction if AI seizes control.
  - **Concentration of Power:** Widening gap between the wealthy AI-owners and the rest of society.

Practical Exercises (Catch-Up Activities)

Exercise 1: The AI Regulatory Sandbox (Simulation)

- **Context:** Imagine you are launching a startup developing a **"Smart City Traffic Optimizer"** driven by AI.
- **Task:**
  1. **Identify Risks:** List 2 major ethical risks your product might face (e.g., Privacy of pedestrians? Bias against certain neighborhoods?).
  2. **Mitigation:** Propose one specific governance measure you would implement to fix this (e.g., "We will anonymize all video feeds before processing").
- **Reflection:** How does regulation help or hinder your ability to innovate in this scenario?

Exercise 2: Ethical Reflection (Pop Culture)

- **Context:** Recall the clip of **C-3PO meeting Luke Skywalker** in *Star Wars*.
- **Task:** Consider C-3PO's behavior. He is subservient, anxious, and strictly programmed for "human cyborg relations."
- **Reflection:**
  - Does C-3PO represent "Safe AI" because he cannot harm humans?
  - If C-3PO had "Agentic" capabilities (freedom to set his own goals), would he still be safe?

Resources & Tool Access

To explore the "Trends of 2025," you can test the reasoning and multimodal capabilities of these frontier models:

- **Google Gemini (2.0 Flash/Pro):** Test its multimodal (video/text) understanding.
- **OpenAI (o1/GPT-4o):** Test its "reasoning" on complex logic puzzles.
- **Anthropic (Claude 3.5):** Test its coding and long-context understanding.

Self-Assessment Questions

1. **Distinguish:** What is the difference between *Soft Law* and *Hard Law* in AI regulation?
2. **Application:** If a company voluntarily decides to "Red Team" their AI model to find safety flaws, is this an example of Governance or Regulation?
3. **Concept Check:** What is "Agentic AI" and how does it differ from a standard chatbot?
4. **Debate:** Briefly summarize the "Alignment Problem" mentioned by Geoffrey Hinton.

Need Help?

If you have questions about the seminar content or upcoming assessments:

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