AI Governance Course

Week 2

March 13, 2024

Agenda

2:45 - 3:05 pm	Jonathan intro, session intro
3:05 - 3:35 pm	Activity 1 - Think Pair Share: Prioritising Hazards
3:35 - 3:45 pm	10-minute break
3:45 - 4:15 pm	Activity 1.5 - AI catastrophic risk scenario analysis
4:15 - 4:20 pm	Activity 2 - Poll: AI Risk Implications
4:20 - 4:30 pm	Wrap-up and closing

Introduction

- Introduction to Jonathan
- Introduction to today’s topic and goals of the session

Introduction to Session 2

Topics to discuss

- Key Definitions
  - Hazards
  - Risks
  - “Existential” risk
  - Narrow AI vs AGI/ASI
  - Alignment

By the end of the week, you should be able to:

Explain the basics of why many experts are concerned about each of the following AI issues:

- AIs could increasingly enable various forms of misuse (e.g. bioterrorism, disinformation, and dangerous entrenchment of values).
- AIs could raise the risks of war (including nuclear war), destructive competition, and other forms of conflict.
- In the next few decades, widely capable AI agents could pursue unintended objectives.

Self-Assessment!

- **Based on the readings & exercises, indicate your confidence and level of interest in each of the key learning outcomes with a number between 1 (low) and 3 (high).**
  - High: Raise your hand high
  - Medium: Raise your hand medium
  - Low: Raise your hand low

Key learning outcomes:

- [Explain why experts are concerned] AIs could increasingly enable various forms of misuse (e.g. bioterrorism, disinformation, and dangerous entrenchment of values).
- [Explain why experts are concerned] AIs could raise the risks of war (including nuclear war), destructive competition, and other forms of conflict.
- [Explain why experts are concerned] In the next few decades, widely capable AI agents could pursue unintended objectives.

Discussion based on self assessment

Activity 1 - Think Pair Share: Prioritising Hazards

In this activity, we will develop a prioritisation for different AI-related hazards.

First, let’s brainstorm:

- What are some examples of a current, pressing, or everyday AI risk?
- What are some examples of an extreme or catastrophic AI risk (can be historical or future)

Activity 1 - Think Pair Share: Prioritising Hazards

Next, you'll split up in pairs.

Let's go to the Collaborative Document (scroll down/skip to Week 2), find the hazard to which your pair is assigned, and discuss:

- Its likelihood (without countermeasures)
- Its severity

Feel free to take notes on the discussion. If you finish early, feel free to comment on others' prioritisations directly in the collaborative document!

Collaborative document

Activity 1 - Think Pair Share: Prioritising Hazards

Interesting ressource for risk quantification :
https://www.metaculus.com/project/ai-safety/

Activity 1.5 - Governing potential catastrophic risk from AI

- Reversible and irreversible harm

- A goal of AI governance
  - Balance risks/benefits of reversible harm
  - Reduce risk of irreversible harm : make it impossible to build unaligned AGI/ASI

Activity 1.5 - Governing potential catastrophic risk from AI

- A high level approach :
  - Forecast
  - Plan countermeasures
  - Implement countermeasures

Activity 1.5 - Governing potential catastrophic risk from AI

- Example of misuse : biorisk
  - If we have AI that can enable manufacturing of novel pandemic agents by 2027
  - What should we do about it?

Activity 1.5 - Governing potential catastrophic risk from AI

- Example of accident : AGI misalignment + scheming + takeover
  - If we can build AGI by 2027 (a few top labs, more every year)
  - Some training runs would result in a strategically aware misaligned AGI with instrumentally convergent objectives
    - Without countermeasures, it could gain more influence until it can takeover
  - What should we do about it?

Activity 1.5 - Governing potential catastrophic risk from AI

- Specific scenario built on existing tech
    - AI can plan and execute autonomously (AI agents)
    - AI can be deceptive (Cicero paper)
    - AI can autonomously code and find vulnerabilities
    - AI could use these to gain more influence over time by participating in the economy, getting influence like MAGMA, lobbying for itself, doing strategic alliances
    - … (this is already takeover)

Activity 2 - Poll: AI Risk Implications

Vote and Discuss! In the Collaborative Document, under Activity 2, you will vote on some statements and whether you agree or disagree.

With the remaining time, we'll discuss any points of disagreement and answer any questions.

Collaborative document

Closing

Takeaways

- What was most useful to you from this week's resources and discussion, and why?
- What's something you found particularly interesting?

Feedback

- What's something you enjoyed or appreciated from this week's resources and/or discussion?
- What didn't go so well from this session for you, and what might we do to improve next week's session?

Next Week

- Paul and Jonathan will be leading the course
  - Michelle will be at an AI conference in Stanford, CA 🤓

Thank you

Next meeting is 20 March at 2:45pm, same place