AI Governance Course

Week 1

March 6, 2024

Agenda

2:45 - 3:15 pm	Icebreaker & Overview of the course
3:15 - 3:45 pm	Introduction to the first session
3:45 - 3:55 pm	10-minute break
3:55 - 4:05 pm	Activity - Poll and Debate: Understanding the 'AI Triad'
4:05 - 4:15 pm	Closing

Who are we?

- Introduction to facilitators
- Introduction to Effective Altruism & AI Safety Collab

Introductions

Introduce yourself:

- Name
- Pronouns
- Program / policy stream you're in
- Motivation for joining the course
- Fun fact or hobbies/interests

Welcome to the first course

Resources
to prepare for the session

Exercises
to complete before the session

Activities
during the session

Engaging with the class / community

Course hub

Participant hub

- Information about the courses and about the project.

Curriculum hub

- Contains info about your next session, meetings and the curriculum.

Slack

- Invitation in the email you signed up with
- Use Slack to have discussions about the course content, or broader field of AI governance, with other participants.
- The course organisers will occasionally make announcements or suggestions there, too.
- Try to check it at least few times a week, for updates.

Discussion Norms

Brainstorm:

- How would you feel best supported by your cohort during this course?
- How should you indicate you want to speak?
- What are the expectations around having done the reading and pre-class exercises before the sessions?
- How can you approach disagreements productively?
- Any other norms for this space?

Introduction to Session 1

Goals of the session

This week introduces the technical basics of machine learning, which is the dominant approach to AI.

The overall goal is to gain a high-level understanding of the technology itself, before we move to understanding the risks and governance solutions in the next parts of the course.

What we'll discuss

- The potential impacts we could see from continued AI progress over the next decade. This will help inform what challenges policy will need to anticipate.
- The factors that are driving progress in machine learning. Understanding the technology sets the groundwork for how to go about setting standards and regulation.

Introduction to Session 1

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

- **Explain the basics of what a neural network is, how they are trained, and how they do inference.** As a result, you should be reasonably comfortable having non-technical discussions about machine learning.
- **Describe some key developments in AI capabilities** over the past decade, with examples. Use this knowledge to be able to make initial predictions about what developments could occur in the next decade.
- **Describe the significance of algorithms, computing power, and data**, for AI development:
  - **Understand the difference between supervised learning, unsupervised learning, and reinforcement learning;**
  - **Describe how compute power has changed** over the last decade, and why that has been important for ML progress.
- **Describe the role data has played** in ML progress over the last decade.

Activity - Poll and Debate: Understanding the ‘AI Triad’

In this activity, we want to ensure everyone understands the definitions and importance of:

- Data
- Algorithms
- Computing power (also known as ‘Compute’).

These are 3 key intervention points for safety standards and regulations, which we’ll discuss in later weeks.

Activity - Poll and Debate: Understanding the 'AI Triad'

Step 1: Go to the Collaborative Document, make a copy of the template, and answer the questions! Don't forget to add your name.

Step 2: Everyone will read each other's responses and add comments where they disagree, or if they are confused about anything that's been written.

Collaborative document

Activity - Poll and Debate: Understanding the ‘AI Triad’

Debrief and discuss!

- What did you learn from others’ writing?
- What did you learn from other’s feedback and comments?

Closing

Discuss

- Any lingering questions?
- Feedback on how the pre-work was
- Any specific skills or topics you want to learn more about?

Next Week

- Jonathan Claybrough from EffiSciences will join as a guest speaker / facilitator

Thank you

Next meeting is 13 March at 2:45pm, same place