AI Governance Course

Week 5

April 10, 2024

Introduction to Session 5

Topics to discuss

- Nonproliferation
- Compute governance

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

- Explain why nonproliferation is considered important for achieving international guardrails on AI.
- Explain how several policy tools could be used by governments with (future) adequate AI regulations to contribute to nonproliferation.
- Identify two types of actions by which leading AI labs can leverage their technology to advance AI safety. E.g.: developing safety methods, assessing and demonstrating risks.

Discussion on the resources

- **Given your knowledge of the risks and some approaches to AI governance, where do you feel the biggest gaps are between existing regulations and ideal AI governance?**

Activity 1 - Compute governance

Purpose: in this activity we will create a shared understanding of 'compute governance', the strategies that could be used for implementation, and issues or challenges with the idea

In the collaborative document, answer the following questions:

- **What are some of the other inputs for advanced AI systems?**
- **According to Lennart Heim (in Introduction to Compute Governance), what are some of its 'fundamental properties'?**

Collaborative document

Activity 2 - Debate on compute governance

Proposition: "Compute governance is actually the only effective way to do AI governance."

When preparing your arguments, you might reflect on the following:

What are some of the issues or challenges with 'compute governance'?

- Why might 'compute governance' be a useful way to reduce catastrophic risks from advanced AI?
    - What are some of the strategies?
    - How could the strategies be implemented, verified, or enforced?
- How practical is the implementation of compute governance strategies?
- What are some of the second-order consequences of these strategies? (e.g., how might actors react if these strategies are put in place?)

Activity 2 - Debate on compute governance

Proposition: "Compute governance is actually the only effective way to do AI governance."

→ Two teams: **"For"** and **"Against"**; each team will have a **2 minute opening statement** and the chance to make rebuttals (15 minutes total for activity)

→ Spend 10 minutes discussing and researching within your team, formulating your argument, predicting any counter arguments, and developing possible rebuttals to what you think the other team will argue. You may use the Collaborative Document as a brainstorming space / notepad if you wish.

Debate!
(20 min)

Discussion on the debate

Coming out of the role of ‘debater’, discuss your true beliefs on this proposition, what evidence supports this view, and what you changed your mind on during this debate.

Activity 3 - Writing exercise of the week

The below writing exercise was the exercise of the week in the curriculum.

Select one of the following strategic directions from “racing through the minefield”:

- **Investing in alignment:** ensuring that AI is developed safely, if it’s going to be developed.
- **Threat assessment:** assessing the risk of misaligned AI, and potentially demonstrating it (to other actors) as well.
- **Avoiding races** between actors aiming to deploy powerful AI systems.
- **Selective information sharing:** sharing some information widely, some selectively, and some not at all.

Explain:

- Somewhat more concretely, what might this look like?
- Explain concrete reasons this may be feasible or infeasible to implement. You may wish to focus on a particular jurisdiction of your choice.
- What are the potential costs to the approach, or why might it be harmful?
- How well does it address the risks it addresses? E.g. there is a significant difference between mitigating a little, mitigating a lot and eliminating a risk - which may inform whether this measure is sufficient for handling this risk.

Activity 4 (if time) - Policy Tools for Advancing AI

As in last week, in this exercise we try to separate the notion of “policies that benefit AI safety globally” and “policies that boost a country’s lead”.

Let’s discuss each of the following policy proposals, and consider:

- **Will this policy benefit AI safety globally?**
- **Will this policy boost a country’s economic / competitive / security lead?**

Proposals

- EU requirement that new models with GPT-4 or above capabilities be made available for third party auditors before deployment
- G7 pausing giant AI experiments for 2 years (Open Letter)
- US banning ‘open source’ models more capable than GPT-4
- EU assigning the burden of designing and carrying out evaluations to independent (for-profit or non-profit) evaluation organisations, rather than auditors that work for the state

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

- Session 6 will cover the topic of **Closing regulatory gaps through international agreements.**
- Please come prepared and read the resources in advance!

Open Questions

- **The last session is 1 May - prefer virtual session?**
- A guest speaker will be invited to speak on the topic of AI careers

Thank you

Next class is 17 April, same time and place