AI Governance Course

Week 4

April 2, 2024

Introduction to Session 4

Topics to discuss

- AI standards
- AI regulations

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

- Articulate some rules that could be set around evaluations of model capabilities and alignment (i.e. when evaluations would be done, and what AI developers would do based on evaluation results), reasons why such rules might advance AI safety, and reasons why they might be limited or harmful
- Compare standards and regulations, including how they are established and what their consequences are
- Compare the state of AI regulation in the UK, EU, China, and US as of approximately mid-2023

Discussion on the resources

- Are there any open questions you have from the resources?
- Was there anything from the resources that struck you as weird or interesting?
- Given the current state of AI development and governance worldwide, what do you think are the most significant challenges in implementing global AI safety measures?

Activity 1 - Debate: Does the US Lead on AI?

Whilst it’s not necessarily an objective to boost the US’ lead, whether the US is “ahead” or not has large implications for the policy actions some are willing to take. In particular, conversations in DC and Silicon Valley about AI regulation often involve a discussion about China’s relative position on AI development to the US.

In this debate, we’ll examine in which ways the US leads on AI development, if advances in the US proliferate to other nations, and examine the likelihood that the US could be ‘overtaken’ in AI development.

The goal of this debate is to surface reasons for and against the proposition, and ultimately help you form your views on the proposition. For the purpose of the debate format you’ll be arguing for one side or another, but we’ll conclude by reflecting on your *true* beliefs.

Activity 1 - Debate: Does the US Lead on AI?

Proposition: **“The United States has a sustainable lead on developing the most advanced AI technology (for at least the next 30 years).”**

→ Two teams: **“For”** and **“Against”**; each team will have a **2 minute opening statement** and the chance to make rebuttals

→ Spend 12 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.

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.

Debate!

(20 min)

Activity 2 - Poll: Policy Tools for Advancing AI

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

Below, the “net beneficial” vote tries to trade off two potential effects:

1. This measure will effectively advance safe development of AI.
2. Any externalities of this measure do not outweigh its benefits for safety.

All of the policies are hypotheticals, excluding the first on US export controls.

They may also be passed for a range of reasons, not just pertaining to AI development, but we’ll discuss them as far as they affect safe development of AI.

Finally, consider them being enacted this year, to set the scope of your discussion.

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

- Session 5 will cover the topic of **Closing regulatory gaps through non-proliferation.**
- Please come prepared and read the resources in advance!

Thank you

Next class is 10 April, same time and place