AI Governance Course

Week 7

April 24, 2024

Introduction to Session 7

Topics to discuss

- Voluntary commitments
- Pausing AI

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

- Explain several strategy ideas for addressing AI safety challenges, and compare their strengths and weaknesses
- Describe several ways in which AI companies can be important governance actors (i.e. decisions they make that shape the impacts of AI)
- Identify two reasons why shared positive visions for AI might be very beneficial, and a few challenges with developing such visions

Quick housekeeping

- **Jobs / Fellowships**
  - CAIDP Fall Policy Clinic (deadline May 10)

- **Events**
  - **[CERI/PSIA event]** Science Diplomacy: Exploring New Frontiers in Climate,Technology, and Space, today at 4:30 pm (AI panel at 5pm)
  - **[Save the date]** Glass Room Misinformation event, May 2-3 (all day)
  - **[EA event]** The Most Impactful Action we can Take for Sustainability, April 25, 5-7pm
  - **[AI Safety Collab partner event]** Exclusive Live Q&A with OpenAI: AI and the Future of Humanity, May 8th, 2024, 7-8:30pm

Discussion on the resources

- Do you have any questions about the resources?

Activity 1 - Presenting your findings: Strategies for AI Safety

Prompt: Pick a strategy from this week's resources and make a case as to why or why not you would prioritise putting it into practice in the next 5 years.

Individually, spend the next 15 minutes choosing a strategy, formulating your arguments, and make your case on why or why not you would prioritize it for a specific jurisdiction (you choose).

You may want to include the following in your case: pros and cons, risks and mitigations, feasibility, and implementation.

As a reminder, here are some strategies covered:

- Voluntary (company) commitments
- Software export controls on frontier AI models - Hardware security features on cutting-edge chips
- Track stocks and flows of cutting-edge chips
- Require a license to develop frontier AI models
- Testing and evaluation requirements
- Specific genres of alignment, interpretation, model eval
- Fund defensive information security R&D
- Antitrust safe harbor for AI safety collaboration
- AI incident reporting
- Hold AI developers liable
- Create means of rapid shutdown
- Slowing down AI / pausing AI
- AI Ideal Governance, including Positive Visions

Activity 2 - Your Vision For The Future

So far on this course, we’ve charted a path which tries to navigate the risks of developing advanced AI technology from now into the future. We have spent very little time on the question of the future: **what happens *after* advanced, generally capable AI is (potentially) developed?**

In this activity, we’ll spend some time envisioning a possible future with advanced AI. The purpose of this activity is for you to spend time thinking about an ideal end state, such that you can later work backwards to find interventions that help achieve that vision.

Activity 2 - Your Vision For The Future

In pairs, discuss the following prompt for the next 10-15 minutes. You will be asked to present your pair's thoughts, whether there was consensus or disagreement on anything, and answer any questions.

Prompt: Assume we are able to develop generally capable AI that's 'aligned' (i.e. such that rogue AI is not a threat, and misuse concerns have been quelled somehow).

What is your ideal end-state for a world with advanced AI? Do you think our current financial and political structures put us on course to achieve that end-state, and if not why not?

A few prompts that may help your formulate your vision:

- What are your assumptions about who might develop advanced AI, and how access will be managed?
- Do you think we should develop advanced, general AI at all?
- What versions of advanced AI are you more or less excited about (e.g. advanced, narrow systems, or broad, generally capable systems)?
- How quickly will we get to your proposed end-state, on the current trajectory?
- Specifics:
  - What do you think will happen to employment?
  - How will AI affect medicine and other health interventions?
  - How will AI affect other global problems, like warfare, poverty reduction and climate change?

Activity 3 - Slowing down AI (if time)

Prompt: Do you think pausing development of AI systems larger than GPT-4 for 10 years would be net beneficial or harmful?

Try to consider:

- Advantages of AI to human health, and potential tools and products that can improve our lives
- How would a pause be implemented and enforced in practice?
- What would AI labs do during the pause? What would this mean for when the pause is lifted?

Additional discussion prompts

- What measures can be put in place to ensure AI accountability? Discuss the potential roles of AI companies, governments, and international institutions in this process.
- In what ways do you think AI companies can act as effective governance actors in mitigating AI risks? Discuss both the potential and limits of their influence. Are there other industries where self-regulation has worked well or not so well?
- Evaluate the potential impact and feasibility of the strategies listed in "12 Tentative Ideas for US AI Policy." Which would you pick to try to implement now, and why? Consider both impact and feasibility.
- Evaluate the effectiveness of the measures suggested for AI companies to prioritize alignment research, strong security, and safety standards in “What AI Companies Can do Today…”. What could be the potential drawbacks?
- What would your ideal vision for a world with advanced AI look like? Are we on track to achieve that world, conditional on developing advanced AI? If not, what would need to change to bring it about?
- Can you design a new social norm that could potentially help in coordinating the world's behavior towards a certain issue? Explain how it would work and how it could be implemented.
- Consider the argument presented in "Let's think about slowing down AI." On balance, do you think slowing down AI development might be a viable and beneficial strategy? Going into more detail:
    - Which of these strategies do you think would be most successful for slowing down AI, and why?
    - What are the downsides and risks of adopting such a strategy?

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 8 will be on **careers**! We will have a mini-panel with people who have worked on different aspects of AI safety.
- Please come prepared and read the resources in advance!

Open Questions

- **The last session is 30 April, 10:15-12:15 in 32, 27 rue saint-Guillaume**

Thank you

Next class is 30 April in 32, 27 rue saint-Guillaume