# AI Governance Course

## Week 3

March 20, 2024

# Agenda

| | |
|---|---|
| **2:45 - 2:50 pm** | Session intro |
| **2:50 - 3:10 pm** | Key questions on the resources |
| **3:10 - 3:30 pm** | Activity 1 - Presenting on challenges in AI development |
| **3:30 - 3:35 pm** | Break |
| **3:35 - 4:40 pm** | Activity 2 - Role Playing: Stakeholders in AI Development |
| **4:40 - 4:45 pm** | Wrap-up and closing |

# Introduction to Session 3

## Topics to discuss

- AI safety
- AI alignment

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

- Explain in non-technical terms why AI systems might behave differently in testing than in deployment. (That would be a challenge for achieving safety through testing.)
- Describe a few mechanisms by which future, widely capable AI agents could hypothetically take power from humans. (If that escalated sufficiently, it could be irreversible.)
- Describe a few challenges faced by research agendas in AI safety.
- Identify several social/political dynamics that may make it harder to achieve AI safety.

# Quick discussion on the resources

**Was there anything from the resources that struck you as weird or interesting?**

# Discuss - Key Questions

- Are there any aspects of AI safety that you believe could be more measurable than others, and if so, why?
- In your opinion, how justified is the concern raised in "Nobody's on the ball on AGI alignment" about the size of the AI safety field and the challenges it faces?
- If you were to advise policy makers on AI safety, what key points from these readings would you emphasize and why?
- How do these readings influence your perspective on the role of competition in AI development?

# Activity 1 - Presenting on challenges in AI development

**The goal of this exercise was to deepen your knowledge of technical and economic challenges in achieving AI safety.**

**Use this exercise to surface areas of confusion, or pointers to broaden your case for AI risk, by receiving feedback from peers.**

Collaborative document

# Activity 1 - Presenting on challenges in AI development

Imagine you are writing a report that introduces a policy recommendation to address *one* of the following risks or incentive structures:

*Technical problems:*

- Power-seeking behaviour
- Deceptive behaviour post-deployment
- Difficulty in correctly specifying quantities humans value in a loss function (e.g. reward misspecification and Goodhart's law)
- Goal misgeneralisation [from optional readings]

*Incentive Structures:*

- A 'Race to the Bottom' on AI safety
- Competitive pressure to delegate power to AI systems
- Acceleration of AI research by AI systems

For this exercise, write an *introduction* to that policy recommendation, which addresses *one* of the above challenges.

# Discuss - Activity 1

- What similarities did you notice between your essay and your peers'?
- Where did you differ on or disagree on?
- What else stood out to you?

# Activity 2 - Role Playing: Stakeholders in AI Development

**We'll address the following learning objective:**

**"Identify several social/political dynamics that may make it harder to achieve AI safety."**

Roleplaying exercises can help to surface political tension between different stakeholders' objectives. There is also the chance to think creatively about how to satisfy all parties' interests without compromising on safety (as far as possible).

# Activity 2 - Role Playing: Stakeholders in AI Development

A company is proposing to train a new model which they claim would be able to do every task humans can do with a computer, at least to the level of a human expert. It would include language (and coding) abilities, and audiovisual content generation and recognition. It would cost approximately $10/hour to run, at the speed of a human.

**They are collectively deciding if and how to initiate this training run.** This discussion takes place over the course of a week, and the company is able to begin training immediately after that. The training will by default take 1 month, after which the model could be deployed.

# Activity 2 - Roles

**A spokesperson for the President of the United States.** The priorities of the President are to harness the benefits of AI for their citizens, but also protect its citizens (and by extension, the world) from its harms and risks. They may also be concerned about the US' international standing and national security threats from AI.

**An AI Safety technical expert** who is being consulted on the risks of the deployment. This role could consist of pushing for particular checks and interventions to be required, or taking the role of explaining which risks will need to be mitigated for other participants to propose solutions for.

**CEO of a big AI lab** (located in the US), which is proposing to initiate the training run. They aim to do this due to: their belief in the benefits of AI; commercial pressure from their shareholders; and competitive pressure from other AI orgs; and profit incentives.
However they are also aware that there may be risks to deploying the technology (and are ultimately subject to US law) so want to be cooperative in this process.

**CEO of an AI startup** that is 3-6 months behind the former company. Their main motivation may be to ensure that their company can remain competitive and do not lose out due to regulatory capture.

**A member of the general public** being consulted (assume they are not strongly politically motivated).
They may be concerned about the impact AI could have on their life and livelihood, but also receptive to the benefits. This role aims to represent that any solution reached by the other stakeholders has to be palatable to voters.

**Spokesperson for EU**
These will have a similar perspective to the US President's spokesperson, but from a slightly different background/perspective.

**Spokesperson for China**
These will have a similar perspective to the US President's spokesperson, but from a slightly different background/perspective.

[Optional] **A spokesperson for the United States' National Security Agency (NSA)**.
Their main priorities are to ensure that their (perceived) adversaries don't gain technical dominance over the US, and

[Optional] **A generic ML engineer** in favour of accelerating AI.
They represent the most optimistic factions around AI's benefits, and are against slowing down AI development.

# 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

- NO CLASS!

# Thank you

No class next week!
Next week is 3 April, same time and place
Michelle & Paul will lead