# AI Governance Course

## Week 6

April 17, 2024

# Introduction to Session 6

## Topics to discuss

- Nonproliferation
- Compute governance

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

- Describe, at a high level, a proposal for privacy-preserving verification of compliance with rules on large-scale AI development.
- Identify policy tools that governments have historically used for nonproliferation of other technologies, as well as the outcomes of these nonproliferation efforts.
- Explain various concepts relevant to international relations, such as the bureaucratic politics model.

# Quick housekeeping

- **Jobs / Fellowships**
  - EU Tech Policy Fellowship (deadline April 21)
  - Pivotal Research (prev. CHERI) fellowship (deadline April 21)
  - CAIDP Fall Policy Clinic (deadline May 10)

- **Events**
  - **[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
  - **[Webinar]** Managing the Risks of Responsible AI, today at 5pm

# Discussion on the resources

- Do you have any questions about the resources?

# Activity 1 - Mini-presentation: Viability of Policy Tools

**Prompt: Choose one policy tool that governments have historically used for nonproliferation of other technologies, and analyze its viability in the context of AI for preventing proliferation of advanced models in the coming decade. Explain any adaptations that would need to be made.**

In groups of two, spend the next 15 minutes formulating a mini-presentation (~5 minutes, plus 2 minutes Q&A). You may use whatever format you want (slides, writing on whiteboard etc) but you don't have to make slides.

**You may reference the List of Nonproliferation Tools from BlueDot (sent via Slack):**

- Treaties and Agreements
- Export controls
- Sanctions
- Interdiction Initiatives
- Research Controls
- Diplomacy and Dialogues
- Cooperative Threat Reduction Programs
- Disarmament Initiatives
- Awareness and Capacity-Building Programs
- Technical Barriers:
- Intelligence Sharing
- Red Teaming and Simulations

# Presentations!
# (5 min pres,
# 2 min Q&A)

# Activity 2 - Climate Change Governance vs. AI Governance

In this exercise, we will try to develop some ideas about what cooperative institutions for nonproliferation of frontier AI could look like.

This activity aims to identify similarities and differences between the international response to climate change and potential approaches to AI risks, thereby understanding the efficacy of different strategies and tools.

The purpose of this exercise isn't necessarily to land on the 'right' answer, but to expose the sorts of considerations that need to be made when making plans like these.

# *International Response to Climate Change: A Brief Overview*

| Initiative/Agreement | Purpose | Key Features |
|---|---|---|
| **United Nations Framework Convention on Climate Change (UNFCCC) - 1992** | Provide a framework for negotiating specific international treaties (called "protocols") that set binding limits on greenhouse gases. | - Recognized the responsibility of developed countries to take the lead in addressing climate change. |
| **Kyoto Protocol - 1997** | Legally bind developed countries to emission reduction targets. | - Introduced carbon trading. If countries emit less than their target, they can sell the excess capacity to others. |
| **Intergovernmental Panel on Climate Change (IPCC)** | Provide policymakers with regular scientific assessments on climate change, its implications, and potential future risks. | - Produces comprehensive Assessment Reports about climate science, technical, and socio-economic knowledge. Forecast climate change impacts.<br>- Influences policies and international negotiations. |
| **Paris Agreement - 2015** | Strengthen the global response to climate change by keeping global temperature rise well below 2°C above pre-industrial levels. | - Introduced Nationally Determined Contributions (NDCs) for countries.<br>- Applies to both developed and developing countries. |
| **Montreal Protocol - 1987** | Address substances responsible for ozone layer depletion, which indirectly helps tackle climate change. | - Considered the most successful international agreement on environmental protection.<br>- Led to the phasing out of several ozone-depleting substances. |
| **Global Climate Strikes and Public Movements** | Mobilize public opinion and demand urgent action on climate change from governments worldwide. | - Spearheaded by activists like Greta Thunberg.<br>- Demonstrated the power of grassroots movements in influencing international discourse. |

# Activity 2 - Climate Change Governance vs. AI Governance

**Together, let's fill out the below table (discuss).**

| Aspect | Example of Climate Change Response | Potential AI Response |
|---|---|---|
| **Global Cooperation** | | |
| **Research & impact assessments** | | |
| **Verification & Compliance** | | |

# Activity 3 - Feasibility of Compute Monitoring Proposal

Discussion Question: **Which aspect of the proposal to achieve privacy-preserving, efficient verification of compliance with AI development regulations do you feel is likely to be the most challenging to implement?**

- Some things to consider:
- Can you think of a similar policy or relevant organisation that might carry these proposals forward into standards or regulations?
- Which interest groups might pushback on particular aspects of the plan?
- Do you have any suggestions for further research, or mitigations to those challenges?

# 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 7 will cover the topic of **Additional Proposals.**
- 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 24 April in the regular room
(31, 27 rue saint-guillaume)