AI Safety Policy Fellowship

AI Safety Initiative At Georgia Tech | S26

aisi.dev | board@aisi.dev

Overview

This course is designed to help participants understand AI alignment and develop a governance mindset, fast-tracking engaged fellows into opportunities in governance and research. This curriculum is heavily inspired by the Wisconsin AI Safety Initiative's AI Policy Safety Fellowship. It has been edited to make it more appropriate for our cohorts.

We organize content into weeks to help you pace your engagement with the curriculum. No technical or governance knowledge is required, though it may be helpful. By the end of this course, you should be able to understand a range of agendas in AI Governance, develop a personal model for forecasting the next few years of AI progress, and have a reviewed policy brief you can put on your portfolio. We encourage you to voice your opinions, push back on any readings, and critically examine your own viewpoints.

For each section, the core readings **will be done before** the meeting. These readings will be discussed during the meeting.

We have further divided the supplementary materials into **Core Readings** and **Highly Recommended Readings** to help more ambitious students prioritize. Reading the

supplementary materials might also be necessary for those interested in doing AI safety research.

Note: We might have some Google Docs comments sprinkled throughout this document, as we've been thinking about how we can improve on the curriculum for Georgia Tech students. Feel free to leave comments!

Contents

Week 1: Understanding Transformative AI
Week 2: Catastrophic Harms
Week 3: Economics of Social Impact
Week 4: National Security
Week 5: Private Governance
Week 6: Regulatory Approaches
Week 7: Further Involvement

Full curriculum

Week 1: Understanding Transformative AI

This week, we explore the technical underpinnings of modern machine learning as a foundation for AI policy. We'll briefly discuss the history of development, the modern paradigm, and work together to forecast potential growth trajectories. We introduce the idea of "Transformative AI," or AI that leads to unprecedented economic change. If you have a technical background in ML, feel free to skip some readings and start on the additional readings.

Core readings:

1. AI 2027 (25 minutes).
    a. A prediction of what the impact of superhuman AI might look like if it were to come quite quickly. Put together by expert forecasters and top researchers.
    b. There is also a video, which you may find more engaging.
2. How Does AI Learn? A Beginner's Guide with Examples (10 minutes)
    a. An excellent technical explanation of AI for a nontechnical audience.
3. Situational Awareness: Intelligence Explosion (15 minutes)
    a. AI progress is unlikely to stop at human-level. This article outlines how AI could be used to automate research, potentially accelerating progress. It discusses the bottlenecks of automated research and models what superintelligence might look like.
4. Video: Large Language Models, Explained Briefly (10 mins)
    a. A primer on how LLMs work. If you are interested (and it is definitely interesting), check out 3b1b's longer form neural network content!

Highly Recommended Readings:

1. Difference between AI, Machine Learning, and Deep Learning (10 minutes)
    a. Understand the different forms of Artificial Intelligence
2. Transformers and the Tech Behind LLMs (27 min)

In-Session Activities:

- Introduction
- (10 min) Explosive Growth from AI: A Review of the Arguments
    - This article examines why we might (or might not) expect growth on the order of ten-fold the growth rates common in today's frontier economies once advanced AI systems are widely deployed.
- (40 min) Explosive Growth Debate
    - Divide into groups of two or three
    - Each group will pick some projection about how AI will grow.
    - We will give you 15 minutes to do research to argue for why your projection is correct. This is very broad, and we are leaving it so intentionally. It is up to you to define what this means and argue why this definition is best.
    - Then, for another 15 minutes you will debate with each other. Try to be fun and playful here. You can, for example, say things that you don't actually

agree with because it supports your side. This doesn't have to be a professional debate. Try to have a good time here!
    - Finally, we'll take 10 minutes to reflect on the activity.

Week 2: Catastrophic Harms

How exactly might AI go wrong? Why are we at AISI concerned about large-impact risks of AI? This week, we explore the possible catastrophic outcomes of rapid AI development.

Core readings:

1. Video: Writing Doom (25 mins)
    a. Short film, Grand Winner of the Future of Life Institute's Superintelligence Imagined Contest. While quite informal, we've found this to be extremely compelling at explaining the arguments for catastrophic risk intuitively, and it's engaging.
2. An Overview of Catastrophic Risks from AI - Summary (15 mins)
    a. Catastrophic AI risks can be grouped under four key categories which are summarized in this paper.
3. Video: Why would AI want to do bad things? (10 mins)
    a. How can we predict that AGI with unknown goals would be dangerous by default?
4. Why Might Misaligned, Advanced AI Cause Catastrophe? (15 mins)
    a. Explore philosophical arguments for why very advanced AIs put into modern society's structures may lead to catastrophic outcomes.

Highly Recommended Readings:

1. What is AI Alignment? (10 mins)
    a. Briefly learn about the technical challenges of creating AIs that do what we want them to. Try to focus more on the high level arguments than the technical details.
2. How Likely is Deceptive Alignment by Evan Hubinger

3. Learning from Human Preferences by Paul Christiano, Alex Ray and Dario Amodei (2017)
4. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback by Casper et al. (2023)
5. Alignment Faking in Large Language Models (2024)

In-Session Activities:

- How Complex Systems Fail (10 mins)
    - Understand how all systems inherently are imperfect and yet our governmental systems, universities, your favorite big corporation, etc. all still overall keep running.
    - Build intuitions for not necessarily needing to eliminate risk but instead tolerating them within a reasonably decided upon margin.
- Activity: Risk Debate
    - Divide into groups of two or three
    - Each group will pick their own risk or harm with no overlap.
    - We will give you 30 minutes to do research to argue for why your risk is the biggest deal. This is very broad, and we are leaving it so intentionally. It is up to you to define what this means and argue why this definition is best.
    - Then, for another 30 minutes, you will debate with each other. We aren't being very serious about what precisely is the worst risk. You can, for example, say things that you don't actually agree with because it supports your side. This doesn't have to be a professional debate. Try to have a good time here!
    - Finally, we'll take 10 minutes to reflect on the activity.
- Take 5 minutes to reflect on this week's content overall.

Week 3: Economic and social impacts

Transformative AI will radically reshape the global economy. This week, we examine the future of labor and consider how society may adapt to change.

Core readings:

1. Predicting AI's impact on Work (20 mins)
    a. Provides an introduction to automation evaluations—what they are, their various limitations, and their potential uses for predicting AI's effects on human occupations.
2. The Economic Impacts and the Regulation of AI (15 mins)
    a. Read through sections 2 and 3 of this paper. These two sections will give insight into the potential impacts that development of AI will have on the labor market and economic growth.
3. How much economic growth from AI should we expect, how soon? (5 mins)
    a. Read section "Getting automation to impact growth is harder than it seems"
    b. This short section provides some basic intuition on how automation can realistically change the economy. If time permits, read through some other sections to get a greater understanding of the economic forces behind AI advancement.
4. Anthropic's Economic Index (10 mins)
    a. See how a frontier lab tracks its impact on the labor market

In-Session Activities

1. **Activity: (15 minutes)** List 10 specific economic milestones on cards:

    - "50% of current trucking jobs automated"
    - "AI replaces majority of junior lawyers"
    - "Fully autonomous fast food restaurants common"
    - "AI therapists covered by insurance"
    - "50% of code written without human input"
    - "AI runs a Fortune 500 company"
    - "Universal Basic Income in 3+ countries"
    - "AI professors teaching university courses"
    - "Construction work 50% automated"
    - "AI-generated movies win major awards"

    Each person gets 100 "confidence points" to bid on when these will happen (2025-2050 or "never").
    The highest bidder on each item must defend their timeline to the group.
    Others can challenge with specific bottlenecks.
2. Activity: Debate! (25 min)
    a. Split into two teams (regardless of actual beliefs):

- Team A argues everything happens in 5 years
- Team B argues everything takes 20+ years

b. You'll have 15 minutes to research and discuss amongst your team.

c. Each team will then spend 3 minutes presenting their position, alternating speaking time.

3. **Activity:** Identify the "cruxes" of the opposing argument. What observable evidence would help update your views towards or against either side?

Week 4: National Security

This week, we'll be studying the national security implications of AI from the perspective of the US and its allies, although the lessons learned can be applied to other nations as well. Experts expect the "AI race" to accelerate and promote geopolitical instability, followed by a centralization of power once superhuman AI arrives. We discuss the arguments for this projection and their implications on things like compute governance, the Global South, and international cooperation. More details to come.

Core readings:

1. Superintelligence Strategy (20 mins)
   a. This is the main reading for this week. Focus especially on Section 3 and Section 5, which will be important grounding for future weeks and work on export controls.
2. The H20 Problem (10 mins)
   a. Understand weaknesses and strengths of current export controls as they apply to the H20 chip, a powerful inference GPU from NVIDIA.
3. Securing AI Model Weights (20 mins)
   a. Read the introduction, security levels, and conclusion, at minimum. A foundational paper in AI security, it introduced terms like "SL-5" and has

been referenced many times by both private companies and the intelligence community in protecting against AI sabotage.

4. U.S. AI Statecraft (15 mins)
    a. Read “What Remains..” until the conclusion. An analysis of current bilateral deals with Gulf State countries, including what the US and its allies may give up with these deals.
5. Situational Awaropeeness: The Free World Must Prevail (10 mins)
    a. An opinionated piece from Leopold Aschenbrenner motivating a democratic race towards superintelligence. Contrast with the previous.

Activity

1. (15 mins) Begin with a discussion of US-Gulf State deals. Split into 2 groups:
    a. Group 1: Argue for US-Gulf State deals, and explain why they benefit American competitive and national security interests despite the costs.
    b. Group 2: Argue against the continuation of US-Gulf State deals, and explain why they harm American interests in the short and long term.
    c. Do this in two rounds, coming back with counterarguments.
2. (25 mins) We’ll be analyzing US-China competitive dynamics and how they may impact other countries. Split into two different groups from before.
    a. Group 1: Propose actions US allies (so not the US itself) could do to strengthen the impact of export controls and preserve a US lead in AI.
    b. Group 2: Propose actions Chinese allies (so not the PRC itself) could do to weaken the impact of export controls and erode a US lead in AI.
    c. After discussing with your group, come up with 3 bullet point actions your team does (e.g., “France uses CeSIA to accelerate studies into compute location monitoring techniques”). Then, have the facilitator/Claude/GPT/LLM of choice adjudicate the outcome this would have. Build on this outcome for the next round.
    d. Repeat the above twice. At each step, after adjudication, discuss as a full group what impacts this might have on the economy of (1) the US, (2) China, and (3) other nations.

Week 5: Private governance

This week, we tackle two problems: understanding the inputs required for scaling AI and how policy controls that process. We'll contrast different governmental and corporate approaches to responsible scaling and critique their merits.

Core readings:

1. Scaling Law Introduction (10 mins)
    a. Learn about power laws, and see why they are useful for predicting the capability of future, more advanced AI systems.
2. How Predictable is LLM Benchmark Performance? (10 mins)
    a. Explore research around how well we are able to predict how future models will behave when they pop out of training.
3. OpenAI o1 Models (5 mins)
    a. Understand a new approach to scaling for performance jumps where rather than scaling your training compute, you scale your inference compute and allow an already fully trained model to produce outputs in silence, meta-process them, and then create a final response.
4. Model Evaluation for Extreme Risks (10 mins, skim through paper)
    a. In this paper, the authors argue that continuous testing of frontier AI models' capabilities and alignment should inform decisions about AI.
    b. It promotes building capacity for model evaluations that can reliably detect extreme risks.
5. Frontier AI Regulation: Managing Emerging Risks to Public Safety (25 mins, read the Executive Summary and Section 3: "Building Blocks for Regulation")
    a. Understand why Frontier AI is particularly the most in need of regulation.
    b. Learn about proposals for mechanisms to ensure compliance from AI organizations and the standards to uphold them to.
6. OpenAI Charter and OpenAI LP Announcement (5 mins)
    a. These short documents enumerate one set of existing organizational commitments to AI safety and the broad distribution of AI benefits.

In-Session Activities

1. Anthropic's Updated Responsible Scaling Policy (35 mins)
    a. Read through Anthropic's extensive plan for ensuring safe outcomes as they responsibly scale up their model capabilities.
2. Activity:
    a. For the following activity, you all together are going to practice thinking about the potential capabilities of AI and timelines that they might arrive on. You likely don't have too much background in this currently, but that is okay. This is a fun exercise. In working on it, you'll uncover the big uncertainties in your mind and work towards better modeling the big cruxes that would shift your mental models of the future.
    b. 1) Map out on the white board what different capabilities apply to the different ASL thresholds: ASL-2, ASL-3, ASL-4, and ASL-5. These aren't always entirely specified, particularly for the later thresholds so try to brainstorm about what they might be with your group. Think about what capabilities would be dangerous
    c. 2) Once you have the capabilities all outlined, try to think about what timelines you expect these to arrive on. Create an estimate for the average year you all agree on. Create a confidence range/interval for the years you are 50% confident that it will occur in. Then, create a confidence range/interval for the years you are 90% confident that it will occur in.
    d. As you write out these estimations, also write below what somewhat realistic key events or factors if they were to happen would most shift your estimations.
        i. For example, if OpenAI's inference compute scaling approach as done with their o1 model continues to improve noticeably over the next year (don't worry if you don't know what this is!), then this would make me think X capability would arrive around Y to Z years earlier.
    e. We understand that this will be challenging. The end goal is not to create the most perfect and calibrated estimates. No one knows when things will happen. It's mostly just to practice this sort of thinking. If you do this, then you'll have done a good job!

Week 6: Regulatory Approaches

AI regulation is still a burgeoning field. This week, we'll read about current legislative paradigms and proposed frameworks to regulate AI around the world. This is meant to give you a birds-eye view of the space, and there is a vast amount of optional content you can read.

Core Readings:

1. The Overton Window Explained (10 mins)
    a. Understand how the Overton Window influences what policy ideas are viable.
2. The EU AI Act, Briefly (15 mins)
    a. Read about the most important current standard for AI regulation.
3. SB 1047, Briefly (10 minutes)
    a. Explore the failed California bill to regulate frontier AI.
4. Preparing for Launch (15 minutes)
    a. A vision of the future of AI progress and how to actualize it.
5. 12 Tentative Ideas for US AI Policy (6 mins)
    a. This article proposes some relatively simple AI policy ideas to spark thought on what the possibilities are.
6. Find or create an interesting AI policy idea to share with the group. Think about its potential benefits and drawbacks, what coalitions might support or oppose it, and any failure modes. (15 mins)

Additional Readings:

1. Exploring International Institutions for AI Governance (20 mins, read the executive summary and section 3)
    a. Explore several potential international institutions that could support the benevolent development of AI and proliferation of its benefits
2. Secure Infrastructure for Advanced AI (15 mins)
    a. OpenAI's call for enhancing infrastructure security, including trusted computing, network and tenant isolation guarantees, operational and physical security for datacenters, and more.

In-Session Activities:

- Let's read: American AI Plan (15 mins)
    - Understand the new administration's view on AI safety and what it means for policy rhetoric
- How to Write A Policy Brief - UNC Writing Center (10 mins)

    - A brief guide which offers some tips on brief-writing.
- Let's write a policy brief! (45 mins)
    - Take the remaining time to ideate, research, and write a quick draft of your policy pitch. Consider your audience and theory of change. It's okay if you don't finish!

Week 7: Further Involvement

This week, we'll be writing an op-ed and discussing further involvement in AI governance, including career support. More details to come.