AISST Policy Fellowship Syllabus

Here, you can find the syllabus we'll use throughout the fellowship. It contains all the papers we'll read together, as well as several others, marked "additional readings," which we've provided for participants to (optionally) read independently if they're especially interested in a particular topic.

Week 1: What is AI? When might transformative AI arrive?

1. Large Language Models Explained (Part 1, Part 2)
2. Why do people disagree about when powerful AI will arrive?
3. Situational Awareness: From AGI to Superintelligence

Further readings

How AI works

- Large Language Models explained briefly
- But what is a neural network? (3Blue1Brown, 2017)
- How Does AI Learn? A Beginner's Guide with Examples (Adam Jones)

AI timelines

- Why we might expect weird things to happen sometime soon: This Can't Go On
- Open questions: Unresolved debates about the future of AI by Helen Toner (until "#3: Will future AIs still basically be tools, or something else?")
- The Case for Short Timelines
    - The Case for AGI By 2030
    - "Timelines Forecast" by Eli Lifeland et al., AI 2027 (2025)
        - Response:
            - When Decades Become Days: Dissecting AI 2027
            - A deep critique of AI 2027's bad timeline models
    - "Situational Awareness: From GPT4 to AGI: Counting The OOMs" by Leopold Aschenbremmer, situational-awareness.ai (2024)
    - Long Timelines Have Gotten Crazy Short by Helen Toner
- The Case for long(er) timelines
    - The Case for Multi-Decades AGI Timelines
    - Why Do Large Language Models Hallucinate?
    - Where's my Ten-minute AGI?
    - How Fast Can Algorithms Advance Capabilities?
- Pathways to AGI

    - Scaling compute
        - The Scaling Hypothesis
        - The Bitter Lesson
        - Trends in AI: Training Costs and Diffusion Speed
        - "Can AI Scaling Continue Through 2030?" by Jaime Sevilla et al., EpochAI (2024)
    - Recursive self-improvement
        - Will AI R&D Automation Cause a Software Intelligence Explosion?
        - The implausibility of an intelligence explosion
    - Scaling inference-during-training: Inference Scaling Reshapes AI Governance (last section only)
- AI as (Ab) Normal Technology (AI diffusion)
    - AI As Normal Technology by Arvind Narayanan & Sayash Kapoor, Knight Institute (2025) [Part 1]
    - AI As Profoundly Abnormal Technology by Scott Alexander, AI Futures Project (2025) [Parts 1-2]
- Benchmarks / capability evals
    - AGI's Last Bottlenecks
    - Evaluating frontier AI R&D capabilities of language model agents against human experts
    - "Measuring AI Ability to Complete Long Tasks" by Thomas Kwa et al., METR (2025)
    - Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity
    - Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad

Week 2: Misuse

1. Statement on AI Risk [1 minute]
2. The Vulnerable World Hypothesis (455-459, stop at "Type-2a") [15 mins]
3. An Overview of Catastrophic Risks from AI (pp. 6-8) [5 minutes]
4. Optional: Superintelligence Strategy (Section 5, stop at "loss of control")
5. Nonproliferation is the wrong approach to AI misuse [15 minutes]

Further readings

Misuse

- Cyberattacks on critical infrastructure:
    - Superintelligence Strategy: Cyberattacks on Critical Infrastructure (3.2.2)
    - The Future of Phishing

    - ‘3cb’: The Catastrophic Cyber Capabilities Benchmark
    - Cyber attacks on critical infrastructure
- Geoengineering: On the Extinction Risk from Artificial Intelligence: Malicious Geoengineering (25-32)

Structural / gradual risks

- Gradual Disempowerment
- Beyond Accidents and Misuse: Decoding the Structural Risk Dynamics of Artificial Intelligence (30 mins)

Week 3: Loss of control

1. The Problem
2. Misaligned AI is no longer just theory
3. Optional: AI-Enabled Coups: How a Small Group Could Use AI to Seize Power (Abstract, Summary, Introduction, Conclusion)

Further reading:

What loss of control might look like:
We're Not Ready for Superintelligence: AI 2027

Arguments for loss of control:
The alignment problem from a deep learning perspective (Ngo et al., 2022).
Is Power-Seeking AI an Existential Risk?
“Risks from power-seeking AI systems”, by Cody Fenwick (2024)
Empirical examples:
“Alignment Faking in LLMs” by Anthropic (2024)

Week 4: Theories of Victory

1. AI governance needs a theory of victory (just the Overview)
2. AI Pause Proposal (Pause AI)
3. Situational Awareness: The Free World Must Prevail
4. Analysis of Global AI Governance Strategies (Cooperative Development section)
5. An International Agreement to Prevent the Premature Creation of Artificial Superintelligence (presumably just a subsection of it)

Week 5: What should the US do today?

1. Radical Optionality by Christop Winter and Charlie Bullock (1-18) [Ask for authors' permission]
2. Alex Bores AI Framework
3. SB53 (legislative digest, optional to skim the wording of the law itself)
4. ~~The State of State AI Law~~
5. A National Center for Advanced AI Reliability and Security (up to recommendation 2)

Week 6: Liability and Private Governance

1. A simple solution to regulate AI
2. We're not Ready for AI Liability
3. The case for AI Liability
4. If-then Commitments for AI risk reduction
5. Private regulation: Regulatory Markets (Abstract, Section 4, Section 5.1)

Week 7: International governance

1. Computing power and the governance of AI
2. How US export controls have (and haven't) curbed Chinese AI
3. Securing Model Weights (OR Situational Awareness: Lock Down the Labs)
4. Is China serious about AI safety?
5. Verifying International Agreements on AI (Summary)

Week 8: Career Planning and Next Steps

1. Feedback form
2. AI governance and policy (80,000 Hours)
3. Priority list of opportunities
4. Emergingtechpolicy.org/ - Choose 1-2 articles to read (and bookmark other articles of interest for later)
5. Career Planning Template

Addendum: Concentration of Power

1. Extreme power concentration: 80k Profile
2. AI-Enabled Coups: How a Small Group Could Use AI to Seize Power
    a. Podcast discussion
3. "The Intelligence Curse" by Luke Drago and Rudolf Laine (2025)
4. How AI Threatens Democracy
5. In search of a dynamist vision for safe superhuman AI

6. Training AGI in Secret would be Unsafe and Unethical
7. AI Safety and the Age of Dislightenment