ILINA 4.0. Seminar Schedule & Syllabus

Introduction

AI is developing rapidly and could pose catastrophic risks to human safety. This course is designed to introduce participants to the field of AI safety governance and then, briefly, to foundational aspects of high-quality social science research. With regards to AI safety governance, the selected topics are meant to capture the most consequential ideas, policies, voluntary commitments, standards and laws that have been pitched, debated or adopted in the last 2 to 3 years. Apart from the topics listed, we'll consider some cross-cutting questions, including: (i) What kind of regulatory vehicles are the most promising? (ii) What does fairness require in the context of AI safety? (iii) How are geopolitical dynamics influencing the paths that are available? etc. There is no specific jurisdictional focus but we will most frequently refer to the US and the EU. Some of the topics may at first glance appear to be *too legal*. If you haven't studied law, don't let this worry you – the class discussions will be structured to ensure that the broader policy considerations and underlying logic (which are in my view the most important aspects of legal reasoning anyway) matter a lot more than legal technicalities. For the research classes, topics are selected to cover what I consider the most essential concepts that rookies ought to learn. Ultimately, I hope that participants will come out of the seminar super equipped to go on and make an important mark in AI governance.

~ *Cecil Abungu*

How to make the most of this syllabus

- Don't ask AI to summarize the readings for you from the outset – Read for yourself first.
- Where page numbers are not specified, please read in full.
- Use the content in the bullet points that appear before the list of videos or readings to guide your study.
- Make notes as you study, and frequently pause to find out more about the terms and ideas that you don't fully understand (by asking AI to explain, for example).
- Try to generate your own opinions during and after studying the material.

Overview of classes

Part I: AI safety governance classes

- Week 1 (May 23/ May 24): Understanding computers and coding
- Week 2 (May 30/ May 31): Introduction to artificial intelligence and deep learning
- Week 3 (June 6/ June 7): Could frontier AI lead to catastrophic risks? From a moral standpoint, why would that even be a problem?
- Week 4 (June 13/ June 14): How leading companies have tried to address AI safety issues at a high level
- Week 5 (June 20/ June 21): AI audits and evaluations
- Week 6 (June 27/ June 28): How much evidence is enough for stricter rules to be imposed? On whom should the onus lie?
- Week 7 (July 4/ July 5): Governing compute and training data
- Week 8 (July 11/ July 12): The information challenge
- Week 9 (July 18/ July 19): The liability challenge

Part II: Research classes

- Week 10 (July 25/ July 26): Understanding social science research at a high level + how AI is changing how it's done
- Week 11 (August 1/ August 2): Deciphering the right questions to ask
- Week 12 (August 8/ August 9): Reviewing the evidence, writing up your work, getting it out there, promoting it

Detailed breakdown of classes

Part I: AI safety governance classes

Week 1: Understanding computers and coding

- What is a computer?
- How does a computer work?
- Introduction to coding
- Practical demonstration

Mandatory:

1. Watch: Early Computing: Crash Course Computer Science 1
2. Watch: Electronic Computing: Crash Course Computer Science 2

Week 2: Introduction to artificial intelligence and deep learning

- What is AI?
- Machine learning basics
- Deep learning
- Practical demonstration

Mandatory:

1. Watch: What is Artificial Intelligence? Crash Course AI 1
2. Watch: What is Machine Learning? IBM Technology
3. Watch: But what is a neural network? 3Blue1Brown

Suggested:

1. Read: Why the Godfather fears what he's built – Joshua Rothman

Week 3: Could frontier AI lead to catastrophic risks? From a moral standpoint, why would that even be a problem?

- What are 'risks to human safety'?
- In what ways could frontier AI pose risks to human safety?
- Could the idea that frontier AI will pose risks to human safety be overstated?

Mandatory:

1. Read: The Alignment Problem from a Deep Learning Perspective – Richard Ngo et al
2. Read: AI Safety Summit Discussion Paper – DSIT – page 15 to 29
3. Read: 10 arguments that AI poses an x risk – Katja Grace and Nathan Young

Suggested:

1. Read: Measuring AI long-task capability – METR
2. Read: Alignment faking study and responses
    - Anthropic study
    - Summary of the study – Sourav Hun
    - Nuanced response – Jan Leike
3. Read: Why Global South Countries Need to Care about Advanced AI – Cecil Abungu et al
4. Read: Six thoughts on AI safety – Boaz Barak

5. Read: Arvind Narayanan and Melanie Mitchell Discuss Artificial and Human Intelligence – CITP Blog

Week 4: How leading companies have tried to address AI safety issues at a high level

- Corporate structuring to ensure AI is developed responsibly
- Development and implementation of responsible scaling policies
- Can we trust these companies?

Mandatory:

1. Read: The governance of AI companies: Reconciling Purpose with Profits – Paul Oudin and Teodora Groza – page 1 to 22
2. Read: On preparedness frameworks generally – Federation of American Scientists
3. Read: FLI study – IEEE summary
4. Read: Google breaks its promises – Shakeel Hashim

Suggested:

1. Read: Safety and security committee – Jenna Baron
2. Read: AI is testing the limits of corporate governance – Roberto Tallarita
3. Read: On responsible scaling policies – Zvi Mowshovitz
4. Read: What AI labs can learn about self-regulation – Nicholas Caputo
5. Read: OpenAI ditches plan to convert to for-profit business – George Hammond and Cristina Criddle
6. Read: Activating AI safety level 3 protections – Anthropic

Week 5: AI Audits and evaluations

- How AI audits and evaluations work
- The technical and governance gaps in AI audits and evaluations

Mandatory:

1. Read: Under the Radar Report – Ada Lovelace Institute – page 19 to 96
2. Read: Challenges in red teaming AI systems – Anthropic
3. Read: Short preview of key questions in new Berkeley course on AI evals – Ben Recht

Suggested:

1. Read: An Institutional View of Algorithmic Impact Assessments – Andrew Selbst – page 119 to 127

2. Read: Reasons to doubt the impact of AI risk evaluations – Gabriel Mukobi
3. Read: The way we evaluate AI model safety might be about to break – Lynette Bye
4. Read: Big Four firms race to develop audits for AI products – Ellesheva Kissin

Week 6: How much evidence is enough for stricter rules to be imposed? On whom should the onus lie?

- The current state of affairs generally (ie. the common rules and institutions in existence)
- The problem
- Safety cases

Mandatory:

1. Read: Pitfalls of evidence-based AI policy – Stephen Casper et al
2. Read: The International Obligation to Regulate Artificial Intelligence – Bryan Druzin et al – Page 3 to 9; page 21 (where Part II starts) to 35
3. Read: Safety cases for frontier AI – Marie Davidsen Buhl et al

Week 7: Governing compute and training data

- What does it mean to govern compute?
  - Controlling access to high-end GPUs
  - Monitoring and reporting requirements
- Can training data be governed in any consequential way?
- Why might these be useful from an AI safety perspective?
- What are the downsides?

Mandatory:

1. Read: Compute governance literature review – Sophia Jarvis
2. Read: Nonproliferation is the wrong approach to AI misuse – Helen Toner
3. Read: Article 10 (1)-(5), EU AI Act; Article 53 (1) (a)-(d) of the EU AI Act

Suggested:

1. Read: Computing Power and the Governance of Artificial Intelligence – Girish Sastry et al
2. Read: With its latest rule, the U.S. tries to govern AI's global spread – Sam Winter-Levy

3. Read: Trump administration modifies direction of regulating AI chips – O'Melveny
4. Read: Entity-Based regulation in frontier AI governance – Dean W. Ball and Ketan Ramakrishnan
5. Read: Biden administration issues more restrictions on advanced chips and AI models – O'Melveny
6. Read: The global AI divide – Alex Satariano and Paul Mozur
7. Read: At Amazon's biggest data center, everything is supersized for AI – Karen Weise and Cade Metz

Week 8: The information challenge

- Scrutinizing training data
- Scrutinizing model operation
- On the other flank: privacy, trade secrets and information security

Mandatory:

1. Read: Deconstructing design decisions – Andrew Selbst et al – page 416 to 433
2. Read: Article 11 (as read with Annex IV), Article 12, Article 21, Article 78, Article 91, Article 92 of the EU AI Act

Suggested:

1. Read: The future of court-ordered scrutiny of AI training data – Hollywood Reporter

Week 9: The liability challenge

- What challenges does advanced AI create for traditional ways of determining legal responsibility for harms caused?
- What are the potential ways forward?

Mandatory:

1. Read: Addressing the Liability Gap in AI Accidents – Amrita Vasudevan
2. Read: European Commission Report from the Expert Group on Liability and New Technologies – page 19 to 30.
3. Read: Moffat v Air Canada – Barry Sookman

Suggested:

1. Read: AI and aggregate litigation – Daniel Wilf-Townsend

2. Read: The Law of AI is the Law of Risky Agents without Intentions – Ian Ayres and Jack Balkin
3. Read: On the utility of network theory in determining AI liability – Anat Lior
4. Read: Regulating downstream developers – Jonas Schuett et al
5. Read: Megan Garcia v Character Technologies

Part II: Research classes

Week 10: Understanding social science research at a high level + how AI is changing how it's done

- The point of a good research project
- How AI is generally changing research
- Critical skills to hone for good research
- The architecture of a well-designed research project
- The research process that you should follow

Mandatory:

1. Read: What will AI do to (p)research? – Joshua Gans
2. Read: Which research process should you follow? – Cecil Abungu

Suggested:

1. Read: Tips for Writing (Policy) Research Papers – Cecil Abungu

Week 11: Deciphering the right questions to ask and the methodology to use when answering them

- How to decide the main question that you should study
- How to figure out which sub-questions to study
- How to know the follow-up questions that you ought to be interested in
- How to ensure you're following the right logic of inquiry and methodology
- How AI can help with these tasks

Mandatory:

1. Do: Pre-class exercises (to be shared one week prior)

Suggested:

1. Read: My research process: Understanding and cultivating research taste – Neel Nanda

Week 12: Reviewing your sources, writing up your work, getting it out there, promoting it

- What makes an impressive analysis of sources?
- When do you know that you have enough evidence to back a crucial claim?
- Some tips to improve your writing
- Some tips for getting your work out there and promoting it
- How AI can help with these tasks

Mandatory:

1. Do: Pre-class exercises (to be shared one week prior)

ILINA SEMINAR EXTRA SESSIONS

Introduction

From weeks 2 to 11, we'll have optional 30 to 40 minute extra sessions immediately after the main seminar session. In each session, one ILINA team member will share reflections on a topic they have been thinking about deeply. We believe that these sessions will enrich the professional paths of our seminar participants.

Breakdown

Week 2: Grace Chege, *Navigating Sacrifice and Tradeoffs in Your Career.*
Week 3: Mark Gitau, *The Trouble With Optionality.*
Week 4: Tomilayo Olatunde, *Moving to Foreign Places for Work.*
Week 5: Cecil Abungu, *Thinking like Your Assessor.*
Week 6: Jean Cheptumo, *Contending with Gray Areas.*
Week 7: Gathoni Ireri, *On Productive Transience.*
Week 8: Raqda Sayidali, *Loneliness, Connection and 'Adult' Friendships.*
Week 9: *Q & A with George Gor.*
Week 10: Marie Iradukunda, *Navigating the discomfort that comes with growth: My personal experience in research/academia.*
Week 11: Michelle Malonza, *Traits I've noticed that make it easier or harder to be a researcher.*