Game Theory, AI, and X-Risk | Syllabus



Instructor: Aydin Mohseni
Email: amohseni@cmu.edu
Lecture Location: Baker Hall 150;
Lecture Times: TR 12:30-1:50 AM
Office Hour Location: Baker Hall 145B
Office Hour Times: by appointment

Course Description

In this seminar, we will: explore recent developments in the philosophy of artificial intelligence; examine arguments for existential risk from artificial general intelligence; and learn to use the formal tools of game theory, decision theory, and Bayesian reasoning to try to model and assess problems and puzzles in this area.

For the final project, students will select a problem in AI safety (broadly defined), conduct in-depth research, and use formal methods to address or clarify the selected problem. Students also enrolled in 15-784 Foundations of Cooperative AI have the option to use the same project for both courses.

Learning Outcomes

By the end of this course, you will be able to:

1. Describe the logical structure of central arguments on AI safety and X-risk;
2. Evaluate these arguments and evidence using probabilistic reasoning;
3. Use formal tools from game theory, decision theory, and Bayesian reasoning to model and assess problems and puzzles in this area

Course Texts

All readings will be made available in the files page of the course site. You are NOT required to purchase any textbooks or articles.

Course Structure

- Each week, we will discuss a different topic in AI, X-risk, game theory, decision theory, or Bayesian inference.
- There will be a **weekly reading** posted on course site.
- There will be a weekly **reading summary** due before the first class each week.
- You will sign up to give approximate two one-hour **presentations** over the course of the semester.
- Finally, there will be a **course paper** involving applying some of the formal tools we cover to some problem (of your choosing) in AI safety.

Course Paper

The course paper will involve applying tools from the formal frameworks we will explore, including decision theory, game theory, Bayesian inference, and social choice theory, to

address an open problem in AI safety. Examples of such problems include, but are not limited to, corrigibility, instrumental convergence, specification gaming, inner alignment, outer alignment, and acausal trade. The paper should be between 1,500 and 3,000 words in length. It must contain the following components:

- An abstract (no more than 100 words).
- An introduction and literature review (approximately 500 words).
- A description of the mathematical and/or computational model (typically less than 500 words).
- A detailed analysis of the model's results (length as necessary).
- A mathematical and/or computational appendix containing proofs and/or code.
- References and citations (in any standard format, e.g., APA, MLA, IEEE, Chicago).

Aim for clarity, conciseness, and rigor throughout your work. Ensure that the implications of your mathematical analysis are both obvious and accurate. Highlight how your analysis contributes to the existing body of literature on AI safety.

Grading Breakdown

The relative weighting of the aspects of the course are as follows.

Active participation	20%
Reading summaries	20%
Reading Presentations	20%
Course Paper, Abstract	10 %
Course Paper, Draft	10 %
Course Paper, Final	20 %

Grading Scale

A+ : 100–98 %	≥ A : 97–93 %	≥ A– : 92–90 %	≥
B+ : 89–88 %	≥ B : 87–83 %	≥ B– : 82–80 %	≥
C+ : 79–78 %	≥ C : 77–73 %	≥ C– : 72–70 %	≥
D+ : 69–68 %	≥ D : 67–63 %	≥ D– : 62–60 %	≥
F : 59–0 %			

Late Work Policy

Late work decays in grade at the rate of 10% per day. So, an assignment that is 2 days late can at best get 80%.

Exceptions to this policy are only made in cases where (1) official, school-sanctioned documentation is provided, or (2) you have contacted me at least a week ahead of the assignment due date and we have come to some sort of alternative arrangement together.

Also, your lowest grade in the reading summaries will be dropped. So, you can effectively be arbitrarily late on a reading summary without suffering any penalty to your grade.

Participation and Attendance

You get your participation grade by showing up to seminar. You can miss two sessions without any loss to your participation grade. You are not required to give a reason for these. For any more than two missed sessions, however, you must have a doctor's note or equivalent official documentation, otherwise you will lose part of your participation grade. If you are going to miss a session where you would be presenting with a team, it is your responsibility to find a replacement at least one week prior to your presentation date. If you do not find a replacement for your team, you will lose a letter grade in the course.

How to Succeed in this Course

To succeed in this course, I advise the following:

- Show up to seminars and participate in discussions.
- Do the weekly readings.
- Complete the weekly reading summaries.
- Do a great job on the presentations. (Collaborate respectfully and well with your team members.)
- Do the course paper abstract, outline, and drafts

If you do all these things, I predict you will do very well in the course!

Office Hours and Getting Help

I will hold office hours at set times during the week. You are enthusiastically encouraged to come to these hours!

Their purpose if you to ask any questions you may have about the course content, assignments; to discuss any difficulties, concerns, or questions regarding the course or your performance; and just to enjoy chatting about philosophy!

You can also contact me via email. You should allow for 48 business hours for a response via email.

Academic Integrity

I trust you to be adults. This course will follow the University policy on academic integrity.

If you are unsure about the university's policies and definitions regarding academic integrity and plagiarism see:

`www.cmu.edu/policies/student-and-student-life/academic-integrity`.

Disabilities

If you have any questions or concerns about disability accommodations, don't hesitate to speak with me—I am very happy to help out.

In general, if you have a disability for which you may be requesting an accommodation, you will want to contact the Disability Resources Center (`www.cmu.edu/disability-resources`). The Disability Resources Center will verify your disability and ensure you get appropriate accommodations for this course.

Course Readings

Below is a tentative list of readings for the course. All readings will be made available on the course website.

- **INTRODUCTION**

  - Motivation & Primer on Prediction
  - Carlsmith 2021: *Is power-seeking AI an existential risk?*
  - Kovarik, Merwijk, and Mattson 2024: *Extinction risks from AI: invisible to science?*

- **MACHINE LEARNING & INTERPRETABILITY**
  - Christiano et al 2021: *Eliciting latent knowledge: How to tell if your eyes deceive you*
  - Hutter et al. 2023: *Sleeper agents: training deceptive LLMs that persist through safety training*
  - Bricken et al 2023: *Towards monosemanticity: Decomposing language models with dictionary learning*
  - Taylor et al 2020: *Alignment for Advanced Machine Learning Systems*

- **DECISION THEORY & AGENT FOUNDATIONS**
  - MacDermott et al 2023: *Characterising Decision Theories with Mechanised Causal Graphs*
  - Soares & Levinstein 2020: *Death in Damascus*
  - Turner et al 2021: *Optimal policies tend to seek power*
  - Soares 2016: *Formalizing Convergent Goals*

- **ALIGNMENT & PRINCIPAL AGENT PROBLEMS**
  - Manheim 2018: *Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence*
  - Hadfield-Menell et al 2015: *The Off-Switch Game*
  - Zhuang & Hadfield-Menell 2021: *Consequences of Misaligned AI*

- **MULTI-AGENT COOPERATION**
  - Conitzer & Oesterheld 2023: *Foundations of Cooperative AI*
  - Dafoe et al 2020: *Open Problems in Cooperative AI*
  - Sourbut et al 2024: *Cooperation and Control in Delegation Games*

- **CORRIGIBILITY**
  - Fallenstein & Soares 2015: *Vingean Reflection*
  - Soares et al 2019: *Corrigibility*

- **SOCIAL CHOICE & ETHICAL DILEMMAS**
  - Manheim & Garrabrant 2019: *Categorizing Variants of Goodheart's Law*
  - Eckersley 2018: *Impossibility and Uncertainty Theorems in AI Value Alignment*