CS395T

AI Alignment Theory

UT Austin, Fall 2026

https://utexas.instructure.com/courses/1462155

Time and Place	Mondays and Wednesdays 2-3:15PM ECJ 1.308
Instructor	Scott Aaronson aaronson@cs.utexas.edu www.scottaaronson.com Office: GDC 4.422 Office hours: After lectures or by appointment
TA	Lin-Lin Lee llee3@utexas.edu Office hours: TBD

The astounding progress of AI over the past decade has been accompanied by a rising fear: do we really understand how to align and control powerful AI systems—how to get them reliably to do what we wanted, or would want them to do on reflection, rather than merely what we said? If we succeed at building general-purpose superhuman intelligences along the current paradigm, should we expect that development to go well for humanity? Can we modify the design, training, monitoring, or scaffolding of those intelligences to help ensure that it goes well? While there's been a great deal of recent empirical work touching on these questions, this course will concentrate mainly on theoretical and mathematical foundations. As a warning, the theoretical foundations of AI alignment have not yet gelled into any one coherent body of results accepted as canonical by the field. Nevertheless, in this course, we'll read and debate many of the conceptual and mathematical works that have been most influential in the AI alignment field, from both before and during the current LLM revolution. Student presentations, reports, and projects will play a central role.

Prerequisites. To get the most out of this course, it is highly recommended that students have taken at least one previous course on theoretical computer science (and ideally on computability and complexity), as well as at least one previous course on AI.

Blog. We plan to run a blog for the class, to which rapporteur reports and class project reports will be posted. Students who don't want to disclose their real names will be allowed to use pseudonyms.

Reading. All students will be expected to do the reading for every class (yes, there's a lot of it). Serious engagement with the papers will be factored into the class participation grade.

Rapporteur Requirement. The first class requirement is to serve as "rapporteur" for one class meeting. This means that you speak for roughly 20 minutes about a paper (or, in some cases, multiple related papers), answer questions about the paper and help lead the discussion of the paper, and then turn in a written report about not only the paper itself, but also the main points from the in-class discussion. To assist with this, you can make a recording of the class for personal use only, unless your fellow students object. We will plan to post the rapporteur reports on the class blog.

A first draft of the rapporteur report is due one week after the class in question. We will then suggest revisions and give you another week to submit a revised report.

Final Project. The final project can be done either individually or in pairs, and can concern any aspect of AI alignment so long as it has a substantial theoretical component. You can, for example, run

an experiment, build a system, address an open problem from a paper, survey a subfield (reading multiple papers about it, not just one), or develop your own AI alignment idea in conversation with the existing literature. In the middle of the semester, we'll solicit project proposals, and will discuss the proposals with each student or team individually. At the end of the semester, we'll schedule project presentations, and also require project reports to be turned in.

Grading. Grading will be done as follows:

20% rapporteur presentation
20% rapporteur report
40% final project
20% class participation

Lateness Policies. Late permission must be obtained for a reason such as illness, family emergency, religious observance, or UT-related travel. General busyness or other coursework are not valid reasons. For the rapporteur reports, 1 point will be removed for every day of unexcused lateness. For final project reports, lateness will prevent the reports from being graded at all before the final grades are due.

Academic Integrity Policies. Plagiarism is taken very seriously at UT. Your rapporteur report and final project report must represent your own work, with any contributions from others appropriately credited. You are responsible for understanding UT's Academic Honesty and the University Honor Code which can be found at `https://deanofstudents.utexas.edu/conduct/standardsofconduct.php`

AI Policies. Students are allowed and encouraged to use AI to help with their rapporteur report and final project, under two conditions. First, they must fully disclose how they used AI (meaning, which models were used for which purposes). Second, they must take personal responsibility for every sentence of every work that they turn in, and be able to answer questions about it. Violation of these policies will be considered an academic integrity issue. After all, the whole motivation for this course is that AI is *not* yet at a point where we can trust it to be aligned with our interests reliably.

Resources. While we'll read many papers and articles, there's no official course textbook. Nevertheless, there are some canonical books on AI alignment that might be helpful:

- *Superintelligence: Paths, Dangers, Strategies* by Nick Bostrom
- *If Anyone Builds It, Everyone Dies* by Eliezer Yudkowsky and Nate Soares

There are also previous AI alignment courses that have websites packed with resources:

- Boaz Barak's course at Harvard - https://boazbk.github.io/mltheoryseminar/
- Elad Hazan's course at Princeton - https://sites.google.com/view/cos598aisafety/
- Lionel Levine's course at Cornell - https://lionellevine.github.io/MAIS.html
- Roger Grosse's course at Toronto - https://alignment-w2024.notion.site/
- Iliad Intensive course - https://www.lesswrong.com/posts/dWQnLi7AoKo3paBXF/the-iliad-intensive-course-materials

Disabilities notice. The University of Texas at Austin provides upon request appropriate academic accommodations for qualified students with disabilities. For more information, contact the Office of the Dean of Students at 471-6259, 471-6641 TTY.