

PHI 378: Minds and Machines

Fall 2026

Instructor: Antonio Freiles

Place & Time: Life Science Building 214, TuTh 3:30pm-4:50pm

E-mail: anantoni@syr.edu

Office Hours: 1.30pm – 3.10pm and by Appointment

Course Description: Askill et al., at page 30 of [Claude's Constitution](#), claim:

Our own understanding of ethics is limited, and we ourselves often fall short of our own ideals. We don't want to force Claude's ethics to fit our own flaws and mistakes, especially as Claude grows in ethical maturity. And where Claude sees further and more truly than we do, we hope it can help us see better, too.

This passage, and others in the constitution, seem to assume that Large Language Models (LLMs) like Claude, in “seeing further,” are capable of thinking or understanding. In this course, we investigate the nature of such models in order to assess whether these assumptions are true. In the first part of the course, we will consider what Artificial Intelligence (AI) and LLMs are, and what important philosophical questions they raise. In the second part, we will survey several arguments for and against the view that LLMs are conscious. Finally, we will examine the relationship between LLMs and existential risk. Throughout, we will relate these issues to Claude's constitutional document.

Learning Objectives:

- (1) How to become familiar with the main philosophical debates about LLMs.
- (2) How to read, interpret, and critically engage with both philosophical texts and primary sources on AI.
- (3) How to formulate arguments about the nature and capabilities of AI.
- (4) How to object and reply to the arguments about machine intelligence, consciousness, and existential risk analyzed in class.

Course Materials: All the required readings will be posted on Blackboard under the heading ‘Content’. Original class materials (handouts, slides, etc.) are the course instructor's intellectual property. You may download these materials for your use in this class. However, you may not provide these materials to other parties (e.g., websites, social media, other students) without permission. Doing so is a violation of intellectual property law and of the student code of conduct.

To be ready for in-class discussion and activities, you must consult each required reading before the start of the class for which it is assigned. Finally, reading philosophy is not an easy task. Readings are often complex and very time-consuming. Allocate some time of your day to read philosophy on the days the lectures are in session. Here is a very brief guide on what to look for when reading [philosophy](#).

Assessment: The assessment is in five parts.

1. *Argument Responses*: 10 points of your course grade. Almost every week we will discuss two arguments in class. There will be seven argument responses, one every two weeks. In each, I will ask you to restate schematically, in premise-conclusion form, one of the arguments discussed the previous week, and to present an objection and a reply. You need not explain the argument in full, but you must explain everything necessary to make the objection and reply intelligible. You may draw directly on our class discussion or develop your own response. Each argument response is closed-book and should take approximately 20 minutes in class. We will discuss a mock answer to an argument before the first argument response comes around.

2. *Attendance and Participation*: 10 points of your course grade. Attendance is worth 5 points of your course grade, and it is required. However, I will excuse absences if they are due to religious observance or an email from the Barnes Center informing me that the student cannot attend the class for health-related issues. Private doctor's notes will not be accepted. In-class participation is worth 5 points of your course grade, and it is required. You can participate in class, but if you prefer to avoid speaking in public, you must send me an email 24 hours before each class with a question about the material, except for when exams take place. Finally, the participation grade is conditional on the completion of a pitch on your final project, whose details are below. I will detail how this pitch should be delivered in Week 3 of the semester.

3. *Exams*: 50 points of your course grade. You will take two in-class, closed-book exams. The first exam, on October 8, is worth 20 points of your course grade and covers the material studied in the first seven weeks of the semester. The final exam, on December 15, is worth 30 points of your course grade and covers the entire semester's material. The exams will contain fewer questions than the argument responses, but they will have the same type of questions.

4. *Final Project & Outline*: 15 + 5 points of your course grade. This course requires the completion of a final project. Given the number of Computer Science majors in this class, there are two tracks for completion:

Track A You will design a detailed blueprint for an LLM application that addresses a genuine real-world need in connection with the philosophical questions discussed in this course. To earn a grade, you must first demonstrate that your design is technologically feasible. Once that threshold is met, the heart of the project lies in its philosophical analysis. Alongside a demonstration of how you built the application, you will submit a report of no more than ten pages that lays out the problem motivating the project, describes how you implemented and resolved it, and draws out the philosophical implications that follow.

Track B You will write a final paper of 4,000 to 5,000 words on a topic of your choosing. Your grade will reflect the philosophical skills the course is meant to cultivate: reconstructing an argument, exposing the assumptions and commitments hidden within it, defending a thesis of your own, anticipating an objection and answering it, and writing with clarity and rigor.

The shared policy governing these two assignments is twofold. First, both the report and the paper

must be written in a Google Doc shared with me from the start of the class. The single rule for the Google Doc is that copy-pasting text into it is forbidden for every student. The document, together with the record of the edits you make to it, serves as evidence that you did not rely on AI to generate your ideas. Second, both the report and the paper will be judged against the standards of philosophical writing. It is my view that these standards carry over to virtually any writing task you will face beyond your time at university.

Finally, it is important to note something about philosophical writing. The latter is argumentative writing, and it differs considerably from the kinds of writing you are likely to have met in your earlier college coursework. Here are some resources for learning what those standards involve: [Philosophical Writing](#), [Writing An Introduction](#), and [Guidelines on Writing a Philosophy Paper](#).

5. Oral Exams: 10 points of your course grade. After the submission of your final project, you will undergo a closed-book oral exam of between 10 and 20 minutes, where you will have to answer specific questions about your project and questions on the material of the course that connect to it.

Letter grade: A 93 - 100; A- 90 - 92.99; B+ 87 - 89.99; B 83 - 86.99; B- 80 - 82.99; C+ 77 - 79.99; C 73 - 76.99; C- 70 - 72.99; D+ 67 - 69.99; D 63 - 66.99; D- 60 - 62.99; F 0 - 59.99.

Course Policies: the following policies will be applied throughout the course.

Academic Integrity: Do not cheat! It's unfair to your instructors and fellow students. Any established academic integrity violation may result in course failure, regardless of violation level. For details on SU's academic integrity expectations and policies, visit: <http://class.syr.edu/academic-integrity/policy/>. All work submitted for argument responses and exams must be yours alone. Communicating about argument responses or exam questions with anyone during the quiz or exam period violates academic integrity expectations. The use of AI to complete any assignment is explicitly prohibited and also violates academic integrity expectations.

Religious Observances: If you need to miss a class or deadline due to religious observance, you will be entitled to make-up opportunities if you provide notice by the end of the second week of class. You must do so through MySlice. You can find the relevant information [here](#).

Accommodations: Syracuse University values diversity and inclusion; we are committed to a climate of mutual respect and full participation. There may be aspects of the instruction or design of this course that result in barriers to your inclusion and full participation in this course. I invite any student to contact me to discuss strategies and/or accommodations (academic adjustments) that may be essential to your success and to collaborate with the Center for Disability Resources (CDR) in this process. If you would like to discuss disability accommodations or register with CDR, please visit the Center for Disability Resources. Please call (315) 443-4498 or email disabilityresources@syr.edu for more detailed information.

Health: Mental health and overall well-being have a significant impact on academic success. It is important to develop skills and resources to effectively navigate stress, anxiety, depression, and other mental health concerns. Please familiarize yourself with the range of resources the [Barnes Center](#) provides and seek out support for mental health concerns as needed. Counseling services

are available 24/7, 365 days a year, at 315.443.8000, and I encourage you to explore the resources available through [the Wellness Leadership Institute](#).

Email Policy: I will reply to your emails within 24 hours on weekdays.

Laptop and Phone Policy: laptops and phones are not allowed for note-taking unless prior arrangements with CDR have been made. Pencils, notebooks, and tablets are allowed instead. Students using laptops or phones in class without authorization will not receive the attendance and participation grades for the lecture.

Deadlines Policy: absences for all assignments are justified only if they satisfy the criteria illustrated in *Attendance and Participation*. If a student misses a single argument response without any previous notice, the exercise grade will count as the lowest and thus will be dropped. In contrast, if a student misses more than one assignment without any previous notice, the student will not receive a grade for that assignment. A good rule of thumb is to get in contact with me once you know you will be missing an assignment.

Extra credit Policy: extra credit will not be granted. The assignments already allow you to excel and improve your grade.

Office Hours: I will hold office hours every Thursday before class. It would help me if you could sign up [here](#). If this time does not work for you, you can send an email to arrange office hours at a different time. Your presence during office hours counts positively toward your participation grade (e.g., if you did not participate in a lecture, your participation during office hours would offset the score).

Schedule: each reading shows its complete length in terms of pages. I marked optional readings with a ‘*’.

Week 1, Introduction: Syllabus and Argumentative Toolkit

- August 25, Syllabus Day & *Philosophers Working in or with AI Firms and Organizations*.
- August 27, Amanda Askell et al., “Being Helpful” (pp. 10–28) and “Claude’s Nature” (pp. 68–77) of [Claude’s Constitution](#) & [Anthropic’s Philosopher Answers your Questions](#) (Video).

Week 2, Artificial Intelligence: the Origins

- September 1, Alan Turing, *Computing Machinery and Intelligence* (pp. 27).
- September 3, *Argument Response #1* & Ned Block, excerpts from *Psychologism and Behaviorism* (pp. 19).

Week 3, Artificial Intelligence: Symbolism vs. Connectivism

- September 8, Michael Rescorla, Sections 1–3.2 from *The Computational Theory of Mind* (pp. 8).
- September 10, Terence Horgan and John Tienson, Sections I–IX of *Representations without Rules* (pp. 20).

* Stuart J. Russell and Peter Norvig, Chapter 1 of *Artificial Intelligence: A Modern Approach* (pp. 29).

Week 4, Artificial Intelligence: The Practical Side of Connectivism

- September 15, Ben Levinstein, *A Conceptual Guide to Transformers: Part I, II, III, IV, V* (Substack), and *Large Language Models Explained Briefly* (Video).
 - * Timothy B. Lee and Sean Trott, *Large language models, explained with a minimum of math and jargon* (Substack).
- September 17, *Argument Response #2* & Raphaël Millière and Cameron Buckner, *A Philosophical Introduction to Language Models – Part I: Continuity With Classic Debates* (pp. 30).

Week 5, AI & Consciousness: An Overview

- September 22, You will provide *a pitch* of your project.
- September 24, Thomas Nagel, *What Is It Like to Be a Bat?* (pp. 15), and, David Chalmers, *Facing Up to the Problem of Consciousness* (pp. 22).

Week 6, AI & Consciousness: Not Yet or Never

- September 29, David Chalmers, *Could a Large Language Model Be Conscious?* (pp. 17).
 - October 1, *Argument Response #3* & Anil K. Seth, *Conscious artificial intelligence and biological naturalism* (pp. 42).

Week 7, Exam Week

- October 6, The *outline* for the final project is due on this date.
- October 8, *Midterm Exam*.

Week 8, AI & Consciousness: Language

- October 13, *Fall Break*: no class.
- October 15, Lisa Miracchi Titus, *Does ChatGPT have semantic understanding?* (pp. 11) & Matthew Mandelkern and Tal Linzen, *Do Language Models' Words Refer?* (pp. 10).

Week 9, AI & Consciousness: In Favor

- October 20, Simon Goldstein and Cameron Domenico Kirk-Giannini, *A Case for AI Consciousness: Language Agents and Global Workspace Theory* (pp. 29).
- October 22, *Argument Response #4* & Harvey Lederman and Kyle Mahowald, *Are Language Models More Like Libraries or Like Librarians? Bibliotechnism, the Novel Reference Problem, and the Attitudes of LLMs* (pp. 17).

Week 10, AI & Consciousness: Cautionary Tales

- October 27, Eric Schwitzgebel, Chapters 1 and 7 of *AI and Consciousness: A Skeptical Overview* (pp. 13).
- October 29, Jonathan Birch, *AI Consciousness: A Centrist Manifesto* (pp. 28).

Week 11, Threats from AI: Existential Risk

- November 3, Karina Vold and Daniel R. Harris, *How Does Artificial Intelligence Pose an Existential Risk?* (pp. 25).
- November 5, *Argument Response #5* & Herman Cappelen, Simon Goldstein, and John Hawthorne, *AI Survival Stories: a Taxonomic Analysis of AI Existential Risk* (pp. 24).

Week 12, Threats from AI: The Alignment Problem

- November 10, William D'Alessandro, *Deontology and Safe Artificial Intelligence* (pp. 23).
- November 12, Adam Bradley and Bradford Saad, *AI Alignment vs. AI Ethical Treatment: Ten Challenges* (pp. 34).

Week 13, Threats from AI: Instrumental Convergence

- November 17, *Argument Response #6* & Nathaniel Sharadin, *Promotionalism, Orthogonality, and Instrumental Convergence* (pp. 30).
- November 19, David Thorstad, *Against the Singularity Hypothesis* (pp. 22).

Week 14, Thanksgiving Break

- November 24, No class.
- November 26, No class.

Week 15, Conclusion: Ethics and Safety

- December 1, Amanda Askell et al., “Being Broadly Ethical” (pp. 31–58) and “Being Broadly Safe” (pp. 59–67) of *Claude’s Constitution*.
- December 3, *Argument Response #7 & Exam Review*.
- December 4, There is no class, but your *final project* is due on this day.

Weeks 16-17, Final Week & Final Exam

- December 8, The *oral exam* on your project or final paper will take place on this day.
- December 15, *Final Exam*: 10:15 a.m.–12:15 p.m.

Acknowledgments: I would like to thank Li Kang for helping design this syllabus, specifically its assessments.