LEH 355W: Philosophy of AI

Logistical Details

Instructor: Ryan Miller, ryan.miller77@login.cuny.edu (please include "LEH355W" in subject line). You are welcome and encouraged to email me anytime you have a question your peers cannot answer.

Course meetings: Mon 3-4:15pm, Carman 350 (plus administrative Mon on Weds 2/28, but **not** 2/22)

Blackboard:
https://bbhosted.cuny.edu/webapps/blackboard/execute/modulepage/view?course_id=_2376771_1

Office hours: Mon 2-3pm Carman 365 (no appointment expected); Weds 3-4pm
https://cuny.zoom.us/my/ryan.michael.miller (AI lab session); other times by appointment.

If you want to review paper drafts for structure and grammar, I encourage you to visit the writing center (Mon-Thurs 10a-7p; Sat 10a-2p http://tinyurl.com/LTCvirtualtutoring).

Course Description

This course discusses the nature and ethics of artificial intelligence. For both topics important philosophic articles are considered and related to concrete aspects of the way current leading Large Language Models (LLMs) function, especially ChatGPT-4/Bing Copilot Creative Mode. Students will:

(1) Gather, interpret, and assess information from a variety of sources and points of view by using Large Language Models to summarize articles and perform experiments. Students will interpret and assess the outputs in order to be responsible for their submissions and make arguments.
(2) Evaluate evidence and arguments critically and be able to appraise their usefulness by critically evaluating the outputs of LLMs for hallucinations and philosophical arguments for coherence in their daily assignments. Students will appraise how philosophical arguments and LLMs are useful for understanding each other.
(3) Produce well-reasoned written or oral arguments using evidence to support conclusions by developing their ability to provide evidence during the asynchronous modules and synthesizing this evidence into an argument for their papers.
(4) Demonstrate familiarity with methods of theoretical or abstract analysis and philosophical reasoning by making weekly summaries of philosophical and abstract theoretical computer science texts. Students will demonstrate familiarity with the modes of analysis by incorporating them into a final paper.
(5) Understand the role of theoretical and abstract reasoning in society and public policy or public concerns in which ethics or other aspects of philosophy play a role by writing summaries and term papers which explicitly connect theoretical reasoning to the concrete functioning of LLLMs, an area of acute public interest for AI policy and AI ethics.
(6) Produce an essay or written piece of research or other creative work, in "scaffolded" stages, demonstrating both an ability to express complex ideas for an educated audience as well as the ability to evaluate and utilize a variety of information of an abstract, theoretical or philosophical nature by building a final essay expressing complex interrelated philosophy and computer science concepts out of the summaries, proposal, and extended draft and prior feedback.

Course Schedule (Readings Posted to Blackboard)

#	DATE	TEXT	NOTE
1	1/29	syllabus	
		What is Artificial Intelligence?	
2	async	Mitchell, "Why AI is Harder Than We Think"	
3	2/5	Chalmers, "Could an LLM Be Conscious?"	
4	async	Turing, "Computing Machinery and Intelligence"	
	2/12	*NO CLASS – LINCOLN'S BIRTHDAY*	
5	async	V&S, "Supervised Learning"	**PyTorch Example Lab**
	2/19	*NO CLASS – PRESIDENTS' DAY*	
6	***async***	Browning & Lecun, "AI and the Limits of Language"	**No in-person 2/22**
7	2/26	Dretske, "Perception and Other Minds"	
8	**2/28**	Searle, "Minds, Brains, and Programs"	
9	3/4	Clark & Chalmers, "The Extended Mind"	
10	async	V&S, "Unsupervised Learning"	**PyTorch Example Lab**
11	3/11	Wittgenstein, from *Philosophical Investigations*	
12	async	V&S, "Deep Learning"	
13	3/18	Quine, "Translation & Meaning"	
14	async	Levenstein, "A Conceptual Guide to Transformers"	
15	3/25	Nagel, "What It's Like to Be a Bat"	
		Artificial Intelligence Ethics	
16	async	V&S, "Reinforcement Learning"	
17	4/1	Danks & London, "Algorithmic Bias in Autonomous Systems"	**Paper Proposals Due**
18	async	Pan et al, "Do the Rewards Justify the Means?"	
19	4/8	Bai et al, "Constitutional AI"	
20	async	Lowenstein, "The Obsolescence of the Horse"	
21	4/15	Yudkowsky, "AI Alignment"	
22	async	Krakovna et al, "Specification Gaming"	**Paper Drafts Due**
	4/22-9	*NO CLASS — SPRING BREAK*	
23	async	Manheim & Garrabrant, "Categorizing Variants of Goodhart…"	
24	5/6	Turner, "Reward is not the Optimization Target"	
25	async	Hadfield-Menell et al, "The Off Switch Game"	
26	5/13	Chalmers, "The Singularity"	
27	async	Hubinger, "An Overview of 11 Proposals…"	
		FINAL EXAM DAY	**Final Papers Due**

Assessments

1. Class Participation (6%)

I ask that you (a) introduce yourselves by answering the questions on the Blackboard discussion forum by 11:59 PM on Wednesday 31 January and (b) make substantial oral contributions during at least 5 of the 12 in-person class sessions and/or synchronous AI lab sessions. You all bring helpful background to this interdisciplinary course, and sharing it helps everyone learn.

2. Philosophical Summaries/Responses (10%)

By 10:00 AM before each in-person course meeting, post either a one-paragraph summary of the text with one cited exact quotation of the main point, or a response disagreeing with how another student has summarized the text--is the main point different than what was expressed (perhaps stronger or weaker?) or is the argument for the main point different than what was given?

You may use an LLM to generate these responses (indeed, you are encouraged to do so) but you must note which LLM you use and what prompt you gave it, and you are responsible for the correctness of the post, including quotations and citations.

Your two lowest grades will be automatically dropped, so you may skip this assignment without penalty for two course meetings if you wish.

3. Asynchronous "Lab Reports" (13%)

By 11:59 PM on the Thursday of an asynchronous meeting week, perform the experiment requested and post your results to the Blackboard forum for that week. You must do these every week, as they are the only record of your asynchronous participation.

You may either do the reading, perform the experiment, and make the post completely asynchronously on your own time or you may drop into the Wednesday afternoon synchronous AI lab office hours, where we will discuss and perform the experiments together (though each student is responsible for submitting their own unique response and for the quality of that response). The AI lab office hours will be recorded and posted to Blackboard for the benefit of those working asynchronously on Wednesday evenings or Thursdays. Synchronous attendance is therefore entirely optional, but working on the material together may prove easier. Substantial contributions also count towards your class participation. You are also welcome to work on these collaboratively at other times, as long as you each submit your own answers.

For one week of your choice (after class 5 / 19 February), skip this assignment and instead do the assignment below.

4. Google Colab Notebook / PyTorch Experiment (1%)

For one week of your choice, submit a Google Colab Notebook holding a PyTorch experiment instead of the simpler no-code experiments discussed in (3). Submit the post by 11:59 PM on the Thursday of an asynchronous meeting week, link to a Google Colab Notebook holding the experiment you have performed, and discuss how the experiment you performed supports or undermines the argument of the author for the week.

In the synchronous AI lab office hours for class 5 we will go over how to do these experiments, and you may ask for help in any other AI lab office hours or in private office hours in-person or by appointment.

This assignment is only 1% of your grade—thus if you are terrified of writing code and skip it the penalty will be minor, and if you are a computer science student well-accustomed to PyTorch you will gain no great advantage. I nonetheless suggest that you take the assignment seriously, because (a) working with LLMs at a technical level is a wonderful item for your resume or a letter of recommendation and (b) performing such an experiment is crucial for the argument of your term paper (about which more below).

5. Paper Proposal (5%) – Monday 1 April

By 11:59 PM Monday 1 April, submit a three-paragraph paper proposal arguing that either (a) there is some concrete detail about the working of Large Language Models which lends support to a philosophical conclusion or (b) there is a philosophical argument which should change the way that software engineers concretely build Large Language Models. In either case, summarize the philosophical view, describe the relevant feature of Large Language Models, and explain how the one should influence the other. The proposal may (indeed, should) overlap with already submitted or future submissions of (1), (2) and (3). The instructor will provide samples of papers of both (a) and (b) type. You are warmly invited to discuss proposals during office hours or by email before Blackboard submission.

No late work will be accepted, as brief comments will be provided by 4/5 so that you can incorporate the feedback into your papers. If you want more time to work on the paper, you may submit the proposal early and you will receive feedback sooner.

6. Paper Draft (25%) – Wednesday 17 April

By 11:59 PM Wednesday 17 April, expand the proposal into an eight-page paper with correct citations which accurately summarizes the philosophical view, demonstrates that Large Language Models actually work as described, and clearly explains how the one should influence our understanding of the other. Instructor feedback on the proposal should be reflected in the draft. Please address any grammar, style, or structure issues with the writing center before Blackboard submission.

You are responsible for keeping a digital file in case Blackboard crashes. Essays submitted between 4/18 and 4/21 (inclusive) incur a 10% deduction. Essays will not be accepted after 4/21 without a prior negotiated extension or a medical excuse (enjoy your spring break!).

All direct quotations must be enclosed in quotation marks and all direct quotations and paraphrases must be cited in-text (MLA or APA style) or with a footnote. All cited works, including primary sources, must appear in a bibliography (which does not count towards the length of the essay). Do not submit a separate document for your bibliography; it must appear at the bottom of your essay. Turning in work without proper citations of all words and ideas borrowed from another is plagiarism. All essays are checked for plagiarism, which is a violation of the academic integrity policy and will result in (at minimum) a zero on the assignment.

7. Final Paper (40%) – Assigned Final Exam Day

By the date of the assigned final exam, revise the paper into a twelve-page finished product, incorporating instructor feedback. Please address any grammar, style, or structure issues with the writing center before Blackboard submission. The same plagiarism standards apply as in the draft.

Essays must be submitted on Blackboard by the listed final exam time. No late work will be accepted as grading must be completed.