PHIL 225
Minds and Machines: Philosophical Issues in Generative AI

Fall 2024

Instructor:	Megan Hyska	**Time:**	12:30-1:50
Email:	megan.hyska@northwestern.edu	**Place:**	Kresge 2-415

Instructor Office Hours: Wednesdays 3-5 and by appointment in Kresge 3-433

Course Description:

This course is a survey of philosophical questions about generative artificial intelligence: Are contemporary AI models conscious, and how could we tell if they were? How are AI models' linguistic abilities different from, and how similar to, those of human beings? Can we understand why models generate the results that they do, and does this matter for their use in health care, law enforcement, and hiring? Can they be genuinely creative? What do we mean when we say that AI has automated human labor? How does the ubiquity of synthetic images affect political communication? And should we take seriously the worry that AI poses an existential risk to humans?

Learning Objectives:

1. Reflect on how theories and research from philosophy and the social and behavioral sciences help elucidate factors underlying contemporary socio-technical problems, as well as inform potential solutions.
2. Demonstrate knowledge and understanding of social scientific theories about the influence of technology on the behavior of individuals, interpersonal relationships, and/or group dynamics
3. Develop the ability to critique theories, claims, and policies in the social and behavioral sciences through careful evaluation of an argument's major assertions, assumptions, evidential basis, and explanatory utility.

Course Materials: All readings are either available on Canvas under "Files" or else are linked to in this document. The instructor reserves the right to alter any readings and coursework, though always with at least a week's notice.

Course Outline:

DAY	**READINGS AND COURSEWORK**
Sept 24	Introduction
Sept 26	Class Cancelled, instructor away
	Introductions assignment due at 12:30pm
	Consciousness
Oct 1	Van Cleave, *The Mind-Body Problem*

	Comment on classmates' introduction posts by 12:30
Oct 3	Chalmers, *Could a Large Language Model be Conscious?*
	Understanding and Representation
Oct 8	Bender and Koller, *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data*
Oct 10	Mandelkern and Linzen, *Do Language Models' Words Refer?*
	Explanation
Oct 15	Fleisher, *Understanding, Idealization, and Explainable AI*
Oct 17	Messeri and Crockett, *Artificial intelligence and illusions of understanding in scientific research*
	Creativity and Labor
Oct 22	Young and Terrone, *Growing the image: Generative AI and the medium of gardening*
	Epstein et al., *Who Gets Credit for AI-Generated Art?*
Oct 24	Taylor, *The Automation Charade*
	Data Workers Inquiry short doc
	Presentation sign-up goes live at 4pm
	Images and Political Epistemology
Oct 29	Rini, *Deepfakes and the Epistemic Backstop*
Oct 31	Hyska, *Deepfakes, Public Announcements and Political Mobilization*
Nov 5	Class Cancelled
	AI and Elections Assignment due
	Catastrophe, Fairness, and Alignment
Nov 7	Bostrom, excerpt from *Superintelligence*
Nov 12	Bostrom continued
Nov 14	Gabriel, *Artificial Intelligence, Values, and Alignment*
Nov 19	Tubert and Tiehen, *Existentialist risk and value misalignment*

	Final Assessment
Nov 21	Student Presentations
Nov 26	Student Presentations

Grade breakdown:

- Final Presentation 40%
- Quizzes (every Tuesday) 35%
- Participation and Attendance 15%
- Other assignments 10%

Assignments:

Final presentation

All of us as individuals currently have to make our own decisions about how to relate to AI: whether and how to use it to help with our work or creative endeavours; whether to relate to it like a therapist, friend or even romantic partner, or else as a mere tool; whether to employ it on behalf of our preferred political agendas; whether to replace human employees with it; to some degree, whether to share our data with the companies who will use it to train the next generation of models. Across the world, people are making wide-ranging decisions about what they want their personal relationships to AI to look like.

For your final assessment in this class, you will present a way that some people out there are currently choosing to use AI and make an argument that either criticizes or defends this use. Your presentation must:

- include a set of slides. The final slide in your deck should be a bibliography. Slides must be submitted to the instructor by 5pm on the date 2 days prior to your presentation.
- show some evidence that the use in question is one that people are really engaging in. This evidence needn't be a peer-reviewed academic article; it might be a news story, or a collection of social media posts. Just show good judgment while assessing the credibility of this evidence.
- draw on philosophical ideas touched on in class, as well as at least 1 additional academic source, in your argument for the appropriateness/ innappropriateness of this use.
- be controversial in a good way: don't argue for something with which it would be hard to imagine many reasonable people disagreeing. This doesn't however mean you have to choose to defend something "edgy" or with which you yourself don't agree.
- demonstrate thoughtfulness, creativity, and rigor.
- be given either individually, or in a group of 2 or 3.
- abide by the following time constraints:

  Single presenter: 3 minutes of presentation, 2 minutes of answering audience questions.

  2-person group: 6 minutes of presentation, 4 minutes of answering audience questions.

  3-person group: 9 minutes of presentation, 6 minutes of answering audience questions.

Presentations will be given on either Nov 21st or Nov 26th, the last two days of classes. While attendance is always mandatory, it is particularly important that you be present and participatory for both these days, not merely to give your own presentation but to ask questions after your colleagues' presentations too.

Presentation signup will go live on October 24th at 4pm, first-come first-served. Because your sign-up requires you to commit to whether your are presenting by yourself or in a group, you will need to figure out in advance of this date whether there are classmates who you share a common interest with such that you would like to work together.

Quizzes

Rather than having a traditional midterm exam, this class will examine your grasp of the last week's material every Tuesday starting on Oct 8, and running till Nov 19th (except for Nov 5th). Quizzes are administered at the beginning of Tuesday's class and are designed to take approximately 10-15 minutes. Quizzes are closed-book and questions will concern comprehension of the readings covered the previous week, rather than an evaluation of their arguments.

Other assignments:

- Introductions (due Sept 26th)
- Connections event (due before end of quarter)
- AI and Elections assignment (due Nov 5th)

Participation and attendance: This is an in-person class, so by default in-person attendance is required. Email your instructor in advance regarding any anticipated absences. Except in extraordinary circumstance, in order for an absence to be excused it must be communicated *in advance* of the lecture you will be missing. Note that, per WCAS regulation, a student can not pass this course unless they are present for at least 50% of course meetings.

You earn your participation grade largely by being active (asking questions, participating in conversation) in lecture. This means being ready to discuss the readings every day.

Use of Generative Artifical Intelligence: As determined together on the first day of class, in this course, the use of generative AI tools (e.g. Copilot) is permitted for the following activities:

- Brainstorming research questions
- Drafting an outline to organize your thoughts
- Editing grammar

The use of generative AI tools is not permitted in this course for the following activities:

- Writing a full draft of your presentation or any other assignment
- Writing entire sentences, paragraphs, presentations, or papers to complete class assignments.

Any use of generative AI should be accompanied by a disclosure at the end of an assignment explaining (1) what you used it for; (2) the specific tool(s) you used; and (3) what prompts you used to get the results. Any use of generative AI beyond where permitted will be viewed as a potential academic integrity violation.

Other Policies: This course follows the Northwestern University Syllabus Standards. Students are responsible for familiarizing themselves with this information.