University of British Columbia

Syllabus PHIL 250

MINDS AND MACHINES

The Philosophy of Artificial Intelligence

Winter Term 1, 2024

Last updated: August 20, 2024

This syllabus unfinished and will be changed before the course begins. Once the course begins, it may be reasonably amended at any time by the instructor. This may mean dropping, adding, or changing readings or assignments as the course progresses.

PHIL 250 001: Minds & Machines (3 Credits)
Method: In class lectures
Class meetings: Tue Thurs 12:30 – 14:00 in Frederic Lasserre (LASR) - 104

Instructor

Kousaku Yui
Office hours by appointment
Email: kyui@mail.ubc.ca

Teaching Assistant

TBA
Office hours: TBA
Email: TBA

Course Description & Learning Goals

This course provides an introduction to the philosophy of mind and cognitive science, with a focus on topics in artificial intelligence (AI). As we live through another "AI Spring" (i.e. a period of optimism about AI), questions about the nature of the mind and how it exists in the world are as pressing as ever. What is the mind? Is it just a complicated physical machine—which is to say, an otherwise ordinary part of the material world—or is it something more? What is the cognitive architecture and theoretical basis that has led to such impressive results?

After covering the theoretical and conceptual background, we will also address ethical and social questions. Can computers have minds, be conscious, have free will, be morally responsible, be owed dignity? Can AI be persons, deserving moral consideration? Can AI be creative or original? How will AI change society? A more detailed list of topics can be found below.

The course will not shy away from some technical details, including in neuroscience and computer science, but no background in these fields is necessary for this course. No programming knowledge is required.

Required Materials

There is no textbook required for this course. All readings will be available on Canvas or through the UBC Library online.

Structure

	Grade %	Due Date
Reaction papers	**20%**	TBD
Course Project: Design an AI	**50%**	
Group project proposal		Week 4
Individual essay draft (1000 words)		Week 8
Individual essay final		Week 12
Group presentation		Week 13 & 14
Final Exam	**30%**	TBD

- **Reaction papers**

  Six short reaction papers on readings (less than 300 words each). The structure of this paper is as follows: Pose a question about the reading that relates the material to something we discussed in class. Answering the question is optional. Sometimes posing a really good accurate question is challenging enough.

- **Course Project - Design an AI**

  Students will work together in groups, culminating in a group presentation at the end of the term. The goal is to design an AI, and consider the various pitfalls and implications of this AI if it were to actually be made.

  1. **Group Project Proposal** - A written project proposal. Not all details need to be worked out, but the general function and design of the AI should be outlined. (about 500 words)
  2. **Individual Essay** - Each student will be tasked with writing an individual paper, on a topic that does not overlap with any other student in the group. The essay will be on an aspect of the AI project. For example, one student might choose to explore the potential for algorithmic bias and then consider ways to mitigate these issues. Another student on the same team might consider theoretical issues in cognitive science, such as whether the design of the AI is making assumptions about an aspect of psychology or sociology . Yet another student on the same team might consider the practical societal implications, such as whether and how the AI will safeguard medical privacy.

     Essays should engage with at least one or more pieces of the assigned reading or topics mentioned in the lecture. The best papers will also engage with independent research.

     Two assignments will be due: a shorter draft essay (1000 words) and a final complete essay (3000 words).
  3. **Group presentation** - At the end of the term, each group will do a presentation in front of the class. At the end of each presentation, there will be a question time where other students and I will challenge parts of your project. Presentations will be graded on detail and rigor, originality and style, responses to worries and criticisms, familiarity with the course material, etc. There will be a strict time limit as class time is limited.

- **Final Exam**

  There will be an in-class hand written final exam. Much of the material covered will be from the lectures.

Policies

Academic Misconduct & Plagiarism

Plagiarism is the act of using someone's work without giving credit and passing it off as one's own. It is a serious offense and will result in an "F" in the course. If you are unsure what constitutes plagiarism, consult the UBC guidelines. https://vancouver.calendar.ubc.ca/campus-wide-policies-and-regulations/student-conduct-and-discipline/discipline-academic-misconduct/3-academic-misconduct-ubc-students

Using AI to produce coursework

The heavy use of AI writing tools to produce written work can often be a form of plagiarism and academic misconduct. Even well meaning students might rationalize the use of AI and convince themselves that it is not really cheating. For example, producing a whole text as the basis for your first draft is cheating, even if you later do major edits to this first draft.

That said, the use of AI writing tools in more limited ways (which I will discuss in class) is permitted. Essays must include a written statement of how AI was used in the production of the essay, including what was provided to the AI.

Turnitin

Paper 1 and Paper 2 will be submitted using www.turnitin.com. This site manages assignment submissions and also checks work for signs of plagiarism.

Because Turnitin servers are located in the US, remove identifying information in order to protect your privacy.

Class ID: **TBD** Enrollment password: **TBD**

Attendance

Engaging with philosophy is best with active participation, in person with other students present. Regular attendance is expected and required. Though attendance will not be marked, it is *very* difficult to do well in this course if you do not attend regularly. If you do miss a lecture, it is your responsibility to find out what you have missed. If you plan on missing a lecture, it may be helpful to arrange for another student to share their notes with you. The essays and finals will require information presented only in the lecture.

Lecture slides will be posted on canvas, but most slides will not be understandable without attending the actual lectures. *Videos of lectures will not be provided*.

Access and Diversity

Any student in this course who has a disability that may prevent him or her from fully demonstrating his or her abilities should contact me personally as soon as possible so we can discuss accommodations necessary to ensure full participation and to facilitate your educational opportunities.

Late Assignments

Papers must be uploaded onto Turnitin before class begins on the due date. Unless excused, late papers will be penalized 5% for every additional 24 hour period they are late. If you anticipate that a paper will be late and/or you have a valid excuse, please contact me as soon as possible. It is your responsibility to contact me *before* the due date has passed to receive an extension.

Technology, Recording, & Copyright

Devices are welcome for taking typed notes. I will allow audio recordings on a case by case basis—please ask before recording. Video recordings are not permitted. Recordings may not be presented or distributed to anyone outside of the class without express written permission from the instructor.

All materials of this course (course handouts, lecture slides, assessments, course readings, etc.) are the intellectual property of the Course Instructor or licensed to be used in this course by the copyright owner. Redistribution of these materials by any means without permission of the copyright holder(s) constitutes a breach of copyright and may lead to academic discipline.

Overview of Topics

The course will be divided into 4 parts. Readings will be available on the course's Canvas site.

This list of topics is still in draft form. Some of these topics may be omitted, or may not be covered to the same degree as others.

Part 1. Historical and Philosophical background

- History of thinking about the mind
- Aristotle: symbolic logic
- Hobbes: thinking mechanisms
- Descartes: The Mind-Body Problem
- Information theory and Bayesianism
- Computationalism
- Syntax & Semantics: Formal Systems

Part 2. Cognitive Architecture

- GOFAI (Classical AI)
- The Symbol Grounding Problem
- Biological and Artificial Neural Networks (Connectionism)
- The Frame Problem
- Model training: loss function, information entropy, etc.
- Large Language Models
- Generative Adversarial Networks
- Retrieval-Augmented Generation

Part 3. Artificial Agents

- Strong AI and the Turing Test
- Kinds of intelligence and artificial general intelligence
- Consciousness, Sentience, Persons and Self
- The Extended Mind and Subsumption architecture
- Free Will, Creativity, Originality

Part 4. AI Ethics & Society

- Algorithmic Bias and Responsible AI
- The Alignment Problem
- Machine Slavery
- Mass surveillance
- AI Governance and Regulation
- Automation and the future of work
- The Singularity