Philosophy of Science and Technology: Philosophy and AI

Tuesdays at 14:05–16:30
Room 301, Teaching Building 3

Peter Finocchiaro
My office: B502
My office hours: Thursdays (14:00–16:00) and by appointment
My email: peter.w.finocchiaro@gmail.com
My QQ: 1983481653
Class QQ: 642921014

Course Description:

In 2022, OpenAI launched ChatGPT, a chatbot with a remarkable ability to mimic written language. In some ways, the launch of ChatGPT resembled the introduction of the pocket calculator in the 1970s. On the one hand, both technologies promised to automate certain tasks. On the other hand, they threatened to undermine the significance attached to those tasks: Why should I learn how to do arithmetic when a calculator can do it for me? Why should I learn how to write an essay when ChatGPT can do it for me?

In this course, we will explore how technology forces us to re-examine our lives. We will primarily (though not exclusively) focus on the technology of artificial intelligence. To that end, we will examine some core issues in the philosophy of artificial intelligence, including the difference between "strong" and "weak" AI, computationalism vs. connectionism, and the intentional stance. We will also examine a host of normative issues in domains as wide ranging as education, transportation, art, and domestic labor. Throughout the semester, we will connect these abstract issues to their real-life manifestations in artificial intelligence technology, including ChatGPT.

Required Texts: There are no required texts for this course.

Optional Texts: That being said, there are optional reading assignments. For each week, I have selected a few texts that explore that week's content in greater detail. I encourage you to read some of the optional texts when you find the content especially interesting.

Grade Distribution: The overall grade is determined by the following:

Debriefs	10%
Participation	20%
Exercises	25%
Final Assignment Proposal	15%
Final Assignment	30%

Course Goals:
As I mentioned above, our goal is to explore the extent to which technology, especially artificial intelligence technology, forces us to re-examine our lives. In service to that goal, I offer the following four smaller goals:

(i) to gain familiarity with core issues in the philosophy of artificial intelligence;

(ii) to gain familiarity with how those core issues are affected by existing technology;

(iii) to improve your ability to imagine how those core issues might be affected by technology in the future;

(iv) to improve your ability to philosophically engage with the issues underlying (i), (ii), and (iii), especially when using the English language.

Assignments

Debriefs: At the end of every class session, you will write a short "debrief" about that class. In your debrief, you will answer two questions: (1) what part of that day's class session did you find the most interesting? (2) what part of that day's class session did you find unclear or would like clarification on? You will share these debriefs with me. I will then use the debriefs to identify topics that we can review together (either because many people in the class find the topic interesting or because many people in the class would like clarification).

Participation: Philosophy is an activity that we do, and active participation in philosophy is the best way to learn to do philosophy. You are expected to interact with me and with other students inside and outside of class. It's important to note, though, that active participation is more than just being vocal; it requires carefully thinking through issues and engaging with peers, often by listening to, supporting, clarifying, or justifying their comments. Doing philosophy is not just about expressing your own ideas, but is just as much about engaging with the ideas of others. Metaphorically speaking, the ideal philosophical discussion is less like a game of ping pong and more like a soccer ("football") match. You will be graded on the extent to which you follow this model of active participation.

Exercises: Every week, I will give you an exercise intended to reinforce the lessons from that week's class. Some of these exercises will include standard philosophical tasks, such as constructing arguments and evaluating objections. Some of these exercises will include tasks specific to technology, such as developing an algorithm or using AI to create an artistic image. You will do some of these exercises on your own. But many of them you will do with a few other students and submit your results as a group. Either way, you should email the exercise results to me **by the end of the week – that**

is, by Sunday 23:59 CST. I will grade them on a "✓- / ✓ / ✓+" scale. I will also give you feedback on which parts of the exercises you did well and which parts of the exercises could be improved.

Proposal: As I mentioned above, the launch of ChatGPT has disrupted the way we think about writing, especially in the context of education. Frankly speaking, philosophy professors don't know what to do. You will help them out by writing a short proposal for a final assignment. In this proposal, you will first take a stand on the role that ChatGPT should play in philosophical education and then design a final assignment that coheres with your position. (The deadline for this assignment is TBD, but it will probably be around Week 10, after we discuss ChatGPT in class.

Final: After everyone has submitted their proposal, we will deliberate as a class and decide on which proposal to adopt. I will then settle all of the final details, including the schedule. You will then complete the assignment as designed in the adopted proposal (though I reserve the right to modify the assignment slightly to better fit the course goals).

Reading List and Schedule:

Below is a tentative schedule of the material that we will cover throughout the semester.

Week 1: Introductions, automation, education, and value
Optional Reading: Eric Schliesser's "What ChatGPT Reveals About the Collapse of Political/Corporate Support for Humanities/Higher Education"

Week 2: Technology and "strong" AI vs. "weak" AI
Optional Reading: John Searle's "Minds, Brains, and Programs"; Margaret Boden's "Escaping from the Chinese Room"; Koji Tanaka's "A Chinese Perspective on the Chinese Room"

Week 3: Computationalism, intelligence, and algorithms
Optional Reading: Alan Turing's "Computing Machinery and Intelligence"; Fintan Mallory's "In Defense of a Reciprocal Turing Test"; Matthew Crosby's "Building Thinking Machines by Solving Animal Cognition Tests"

Week 4: Connectionism, neural networks, and machine learning
Optional Reading: Frank Gabels "Some Studies in Machine Learning Using the Game of Checkers"; Ron Sun's "Connectionism and Neural Networks"; Paul Smolensky's "Connectionist Modeling: Neutral Computation / Mental Connections"

Week 5: Beliefs, desires, and the intentional stance
Optional Reading: Daniel Dennett's "True Believers: The Intentional Strategy and Why It Works"; Nick Bostrom's "The Superintelligent Will"

Week 6: Autonomy, responsibility, and moral agency

Optional Reading: Christian List's "Group Agency and Artificial Intelligence"; Amitai Etzioni and Oren Etzioni's "Incorporating Ethics into Artificial Intelligence"

Week 7: Artificial morality and "top-down" vs. "bottom-up" approaches

Optional Reading: Stephen Omohundro's "The Basic AI Drives"; Colin Allen et al.'s "Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches"

Week 8: Generative AI, LLMs, and semantic pollution

Optional Reading: Luciano Floridi and Massimo Chiriatti's "GPT-3: Its Nature, Scope, and Consequences"; Joni Salminen et al.'s "Creating and Detecting Fake Reviews of Online Products"

Week 9: Aesthetic value, creativity, and AI art

Optional Reading: Margaret Boden's "Creativity in a Nutshell"; Andy Clark and David Chalmers's "The Extended Mind"

Week 10: Design, accessibility, and algorithmic bias

Optional Reading: Linus Huang et al.'s "Ameliorating Algorithmic Bias, or Why Explainable AI Needs Feminist Philosophy"; Joy Buolamwini and Timnit Gebru's "Gender Shades"

Week 11: Existential risks, AI regulation, and the value alignment problem

Optional Reading: Neil Richards and William Smart's "How Should the Law Think About Robots?"; Nick Bostrom's "Existential Risks"

Week 12: Sentience and the responsibilities of creators

Optional Reading: Luke Roelof's "Sentientism, Motivation, and Philosophical Vulcans"; Thomas Metzinger's "Two Principles for Robot Ethics"

Week 13: AI containtment and AI rights

Optional Reading: Eric Schwitzgebel and Mara Garza's "Designing AI with Rights, Consciousness, Self-Respect, and Freedom"; S. Matthew Liao's "The Moral Status and Rights of Artificial Intelligence"

Week 14: The simulation hypothesis

Optional Reading: Nick Bostrom's "Are We Living in a Computer Simulation?"; David Chalmers's "The Virtual and the Real"

Week 15: The singularity and the aim(s) of philosophy

Optional Reading: David Chalmers's "The Singularity: A Philosophical Analysis"; Drew McDermott's "Response to David Chalmers"

Week 16: Reserved for holiday cancellations, delays in class, or other unexpected events

(**NB:** if you take a picture of yourself and East Lake during a sunrise or a sunset and send it to me before the end of Week 5, I will give you 1 extra credit point.)