Philosophy of Science and Technology: Philosophy and AI

Wednesdays at 14:05–16:30
Room 102, Teaching Building 1

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: 799268307

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 streamline 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 the extent to which this technological analogy holds true. 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 ethical issues in domains as wide ranging as education, social media, art, and warfare. Throughout the semester, we will connect these issues to their real-life manifestations in artificial intelligence technology.

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.

Letter Grade Distribution: In this course I will use the following scale to convert between numerical and letter grades:

96.00 - 100.00	A+	70.00 - 74.99	B-
90.00 - 95.99	A	67.00 - 69.99	C+
85.00 - 89.99	A-	63.00 - 66.99	C
80.00 - 84.99	B+	60.00 - 62.99	C-
75.00 - 79.99	B	00.00 - 59.99	D

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

Debriefs	Ungraded
Participation	20%
Exercises	30%
Annotation #1	25%
Annotation #2	25%

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 ungraded "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. (The exercises include standard philosophical tasks like constructing arguments and evaluating arguments. They may also include tasks specific to technology like creating a Turing machine or brainstorming principles for the assignment of responsibility related to autonomous vehicle accidents.) You will complete this exercise with a small group of students. **You should submit 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.

Annotations: Twice this semester, you will choose a topic that we covered in class. Then, you will annotate a 2000-word survey paper that covers the main ideas within that topic. This survey paper will be generated by ChatGPT. As the annotator of that paper, your responsibilities include evaluating the accuracy of the material presented, evaluating the clarity of how it is presented, and fact-checking every reference. Your grade will be based on the extent to which you identified what ChatGPT did well and what ChatGPT did poorly.

NB: to protect your privacy, I will be the one who directly interacts with ChatGPT; I will email you the paper that it generates.

Reading List and Schedule:

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

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

Week 2: Behavioralism, "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: Computationism, Turing tests
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, 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, rationality, intentionality

Optional Reading: Daniel Dennett's "True Believers: The Intentional Strategy and Why It Works"; Nick Bostrom's "The Superintelligent Will"

Week 6: Responsibility and autonomous vehicles

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

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: Originality, plagarism, and ChatGPT

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: Creativity and AI art

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

Week 10: Bias and algorithms

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: Robots, the law, 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: Artificial sentience

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

Week 13: 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 singularity and its consequences

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

Week 15: The simulation hypothesis

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

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

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