Philosophy 385: The Philosophy of AI M 2:30–3:50 PM/W 1:00–2:20 PM Conklin Hall 352

Instructor: Cameron Domenico Kirk-Giannini
Office: Conklin Hall 417
Office Hours: Mondays 1:00-2:00 PM (email for appointment)
Email: camerondomenico.kirkgiannini@rutgers.edu

COURSE DESCRIPTION

Can AI systems think? Can they see, hear, and feel? Are they conscious? Do they matter morally? Do we owe anything to them? Do they perpetuate existing social inequalities? Do they pose a threat to our jobs? Do they pose a threat to our lives? Are they our best chance at immortality?

In this course, we will think through these difficult and timely questions. In addition to taking a close look at some of the most transformative AI technologies developed in the past few years (e.g. generative text and image models like GPT-4 and DALL-E 3), we will touch on foundational topics related to AI in the philosophy of mind, ethics, metaphysics, and decision theory.

LEARNING OUTCOMES

- By critically engaging with the course material, students will gain an understanding of some of the most important philosophical issues raised by AI.
- The topics covered in the course will also serve as convenient introductions to major concepts in philosophy including: consciousness, personal identity, bias, meaning, propositional attitudes, and decision theory.
- Through class discussion and structured writing exercises, students will develop crucial philosophical abilities like reconstructing and evaluating arguments, articulating ideas in conversation, and writing clearly and cogently.

TEXTBOOK

There is no textbook for this course. All readings will be made available online.

LEARNING MATERIALS

I will upload course handouts to the course's Canvas site.

ASSIGNMENTS AND GRADING

There will be three significant course requirements. First, you will be required to attend class and participate in discussion of the course material. Second, you will be required to lead one class meeting of your choice during the semester. This will involve preparing a handout, discussion questions, and so on. Third, you will be required to write a paper and then present it to the class as a talk followed by a question-and-answer period in the format of a typical philosophy conference.

Grades will be determined as follows:

- Attendance and class participation: 50%

- Leading discussion for one class: 25%
- Final paper, presentation, and responses to questions: 25%

Grading Scale:

- A = 89.5-100
- B+ = 84.5-89.49
- B = 79.5-84.49
- C+ = 74.5-79.49
- C = 69.5-74.49
- D = 59.5-69.49
- F = 0-59.49

SEMESTER OVERVIEW

Week 1: 1/17

Reading: None (Course Introduction)

Introduction to AI and Machine Learning

Week 2: 1/22 and 1/24

Readings:

1. Ben Levinstein, “A Conceptual Guide to Transformers, Part 1."
2. Andrew Ng, “AI for Everyone," <https://www.youtube.com/watch?v=gtokFuP4Bfs>

AI Ethics

Week 3: 1/29 and 1/31

Readings:

1. Katie Crawford, *Atlas of AI*, Chapter 1 (“Earth") and Chapter 2 (“Labor”).

Week 4: 2/5 and 2/7

Readings:

1. Katie Crawford, *Atlas of AI*, Chapter 3 (“Data") and Chapter 4 (“Classification”).

Can AI Systems Think?

Week 5: 2/12 and 2/14

Readings:

1. John Searle, “Minds, Brains, and Programs,” with replies by Block, Dennett, and Fodor.

Week 6: 2/19 and 2/21

Readings:

1. Emily Bender and Alexander Koller, “Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data.”
2. David Chalmers, “Does Thought Require Sensory Grounding? From Pure Thinkers to Large Language Models.”

Week 7: 2/26 and 2/28

Readings:

1. Eric Schwitzgebel, "Belief."
2. Simon Goldstein and Cameron Domenico Kirk-Giannini, "AI Wellbeing."

Could AI Systems Be Conscious?

Week 8: 3/4

NOTE: CLASS ON 3/4 BY ZOOM; NO CLASS ON 3/6

Readings:

1. David Chalmers, "Could a Large Language Model Be Conscious?"

Week 9: 3/18 and 3/20

Readings:

1. Jeff Sebo and Robert Long, "Moral Consideration for AI Systems by 2030."
2. Simon Goldstein and Cameron Domenico Kirk-Giannini, "A Case for AI Consciousness: Language Agents and Global Workspace Theory"

Digital Minds and Virtual Reality

Week 10: 3/25 and 3/27

Readings:

1. David Chalmers, "Uploading: A Philosophical Analysis."
2. Joseph Corabi and Susan Schneider, "Metaphysics of Uploading."

Week 11: 4/1 and 4/3

Readings:

1. Nick Bostrom, "Are We Living in a Computer Simulation?"
2. David Chalmers, *Reality+*, Chapter 10 ("Do Virtual Reality Headsets Create Reality?")

Week 12: 4/10

NOTE: NO CLASS ON 4/8

Readings:

1. David Chalmers, *Reality+*, Chapter 17 ("Can You Lead a Good Life in a Virtual World?")

AI Safety

Week 13: 4/15 and 4/17

Readings:

1. Stephen Omohundro, "The Basic AI Drives."
2. Joe Carlsmith, "Is Power-Seeking AI an Existential Risk?"

Week 14: 4/22 and 4/24

Readings:

1. Nick Bostrom, *Superintelligence*, Chapter 9 ("The Control Problem").
2. Simon Goldstein and Cameron Domenico Kirk-Giannini, "Language Agents Reduce the Risk of Existential Catastrophe."

Week 15: 4/29

Readings: None

4/29: Final Paper Presentations

COURTESY

It is important that all discussion be conducted calmly and respectfully. Professional courtesy and consideration for our classroom community are especially important with respect to topics dealing with differences such as race, color, gender and gender identity/expression, sexual orientation, national origin, religion, disability, age, and veteran status.

Meaningful and constructive dialogue requires mutual respect, a willingness to listen, and tolerance of opposing points of view. Respect for individual differences and alternative viewpoints will be maintained at all times in this class. Our choices of words and use of language are critical components of respectful discourse as we work together to achieve the full benefits of creating a classroom in which all people can feel comfortable expressing themselves.

ACADEMIC INTEGRITY

As an academic community dedicated to the creation, dissemination, and application of knowledge, Rutgers University is committed to fostering an intellectual and ethical environment based on the principles of academic integrity. Academic integrity is essential to the success of the University's educational and research missions, and violations of academic integrity constitute serious offenses against the entire academic community.

Academic Integrity Policy: http://academicintegrity.rutgers.edu/academicintegrity-policy/

ACCOMMODATIONS FOR STUDENTS WITH DISABILITIES

Every effort will be made to accommodate students who present a valid Letter of Accommodations. For more information, see: https://ods.rutgers.edu/my-accommodations/letter-of-accommodations

RELIGIOUS OBSERVANCE

I am happy to accommodate special needs related to students' religious practices. However, I require that you notify me in writing within the first two weeks of class if you will need such accommodation at any point during the semester.

COUNSELING SERVICES

Counseling services are available at the Counseling Center, Room 101, Blumenthal Hall. For more information, call

(973) 353-5805 or visit http://counseling.newark.rutgers.edu/. Please note that I am required to report certain sensitive information you might relate to me to the University.