Philosophy of Artificial Intelligence

Note to the reader: *this syllabus is for an introductory undergraduate lecture course with no prerequisites.*

Course Description

In this course, we will explore the philosophical implications of artificial intelligence. Could we build machines with minds like ours out of computer chips? Could we cheat death by uploading our minds to a computer server? Are we living in a software-based simulation? Should a driverless car kill one pedestrian to save five? Will artificial intelligence be the end of work (and if so, what should we do about it)?

This course is also a general introduction to philosophical problems and methods. We will discuss classic problems in several areas of philosophy, including philosophy of mind, metaphysics, epistemology, and ethics. In the process, you will learn how to engage with philosophical texts, analyze and construct arguments, investigate controversial questions in collaboration with others, and express your views with greater clarity and precision.

Contact Information

Instructor: David Gray Grant
Email: dggrant@fas.harvard.edu
Office: 303 Emerson Hall
Office hours: Mondays 3:00-5:00 and by appointment

Assignments and Grading

You must complete *all* required assignments in order to pass this course. Your grade will be determined as follows:

- 10%: Participation
- 25%: 4 short written assignments (300-600 words)
- 20%: Paper 1 (1200-1500 words)
- 25%: Paper 2 (1200-1500 words)
- 20%: Final exam

Assignments should be submitted in .docx or .pdf format (.docx is preferred), Times New Roman 12 point font, single-spaced. Submissions that exceed the word limits specified above (including footnotes) will not be accepted. Papers will be graded anonymously; please do not include your name on your paper.

Attendance Policy

Attendance in lectures and sections is mandatory. Please contact me as soon as possible if you anticipate being unable to attend class. (Please note that participation constitutes 10% of your grade.)

Late Assignments Policy

Assignments are due by midnight on the day indicated on the course schedule below. Late assignments will be excused if you provide a dean's or doctor's note. In addition, you have five free late days that you can use whenever you like. You do not need to provide a reason in order to use one or more of your late days, but you do need to notify us of your plans within 24 hours after the assignment deadline has passed.

Collaboration and Academic Integrity

Philosophy is a team sport, and collaboration is essential to success. You are strongly encouraged to discuss the material from this course outside of class and section, both with each other and with your friends, relatives, neighbors, etc.

That said, this course has a zero-tolerance policy for plagiarism. You may consult external sources and discuss your ideas for your written work with others, but you must properly cite and credit the ideas and writings of others. Before our you begin work on the first written assignment, you should review Harvard's Guide to Avoiding Plagiarism. Feel free to ask if you have any questions. (When in doubt, include a citation!) Please use parenthetical citations in MLA format and include a "Works Cited" section at the end.

Accommodation

If you have a disability, then you have a right under Section 504 of the American with Disabilities Act to reasonable accommodations. To request accommodations, please contact University Disability Resources (https://accessibility.harvard.edu/contact, 617-495-1859). If you are eligible, you will receive an accommodation letter which you should bring to me as soon as possible. That way we can work together to make sure all of the course content is accessible to you.

What You Can Expect From Me

Office hours: I will be available at least two hours per week for office hours. Note that you do not need a specific reason to come to office hours or request a meeting.

Responding to emails: I will respond to all emails within two business days. If you not hear back from me in that timeframe, please do not hesitate to email me again, drop by office hours, or come talk to me after lecture.

Reading drafts: I will not be able to read whole drafts in advance, but am happy to read outlines and discuss ideas with you.

Feedback on assignments: You will receive written feedback on each assignment explaining which areas for improvement you should focus on for the next assignment. If you are confused about the feedback you receive, please do come to office hours or request a meeting!

Schedule and Readings

1. Wed 9/2 Introduction

Introduction to the course. Chiang imagines digital pets becoming conscious.

- Ted Chiang, "The Lifecycle of Software Objects" (short story)

2. Mon 9/7 No Class (Labor Day)

Part I: Machine Minds

3. Wed 9/9 The Turing Test

Turing proposes a test for intelligence. Aaronson argues that (despite appearances) no chatbot has passed it.

- Alan Turing, "Computing Machinery and Intelligence"
- Scott Aaronson, "My Conversation with Eugene Goostman"

4. Mon 9/14 The Chinese Room Argument

Searle argues that having a mind requires having a brain like ours. Bisson has some fun with the idea.

- John Searle, "Minds, Brains, and Programs"
- Terry Bisson, "They're Made of Meat"

5. Wed 9/16 Functionalism *Short written assignment 1 due*

Block explains the functionalist theory of the mind, and poses objections.

- Ned Block, "Troubles with Functionalism"

6. Mon 9/21 Emotion

Breazeal and Brooks offer a functionalist treatment of the emotions.

- Cynthia Breazeal and Rodney Brooks, "Robot Emotion: A Functional Perspective"

7. Wed 9/23 The Knowledge Argument

Jackson argues against physicalism about conscious experience.

- Frank Jackson, "Epiphenomenal Qualia"

8. Mon 9/28 The Hornswoggle Problem

Churchland defends physicalism about conscious experience.

- Patricia Churchland, "The Hornswoggle Problem"

9. Wed 9/30 Unit Review *Short written assignment 2 due*

- No readings

Part II: Machine Persons

10. Mon 10/5 Obligations to Machines

Basl argues that we do not have obligations to machines—at least not yet.

- John Basl, "Machines as Moral Patients We Shouldn't Care About (Yet)"

11. Wed 10/7 Uploading I

Locke introduces the problem of personal identity over time, and offers an account.

- John Locke, *An Essay Concerning Human Understanding*, Book II, Chapter XXVII ("On Identity and Diversity")
- *Black Mirror* Season 3 Episode 4 ("San Junipero")

12. Mon 10/12 No Class (Indigenous People's Day)

13. Wed 10/14 Uploading II

Parfit develops Locke's account of personal identity and canvasses alternatives.

- Derek Parfit, *Reasons and Persons*, Chapter X ("What We Believe Ourselves To Be")

14. Mon 10/19 Uploading III *Short written assignment 3 due*

Is uploading metaphysically possible? Chalmers says "maybe."

- David Chalmers, "Mind Uploading: A Philosophical Analysis"

15. Wed 10/21 The Simulation Argument I

Descartes argues that you shouldn't trust your senses. Nagel explains the problem of skepticism.

- Rene Descartes, *Meditations on First Philosophy*, Meditations I and II
- Jennifer Nagel, "The Problem of Skepticism" (video)
- *The Matrix* (film)

16. Mon 10/26 The Simulation Argument II

Bostrom argues that you should believe you are living in a computer simulation.

- Nick Bostrom, "The Simulation Argument: Why the Probability that You are Living in a Matrix is Quite High"

17. Wed 10/28 The Simulation Argument III

Bostrom, continued. Moore refutes the skeptic with his hands.

- G.E. Moore, "Proof of an External World"

18. Mon 11/2 The Simulation Argument IV

Chalmers argues that the hypothesis that we are living in a simulation isn't a skeptical hypothesis.

- David Chalmers, "The Matrix as Metaphysics"

19. Wed 11/4 Unit Review *Paper 1 due*

- No readings

Part III: Machine Ethics

20. Mon 11/9 Autonomous Vehicles I

Hao reports on the MIT Media Lab's Moral Machines project. Midgley argues that what you ought to do doesn't depend on where you are from.

- Karen Hao, "Should a Self-Driving Car Kill the Baby or the Grandma? Depends on Where Youre From" (*MIT Technology Review*)
- Mary Midgley, "Trying Out One's New Sword"

21. Wed 11/11 Autonomous Vehicles II

Jacques criticizes the Moral Machines project. Nyholm and Smids consider the relevance of trolley problems for the design of autonomous vehicles.

- Abby Jacques, "The Moral Machine is a Moral Monster"
- Sven Nyholm and Jilles Smids, "The Ethics of Accident-Algorithms for Self-Driving Cars"

22. Mon 11/16 Algorithmic Discrimination I *Short written assignment 4 due*

ProPublica argues that an algorithm used widely in the criminal justice system is biased against African Americans. Hellman considers the nature of discrimination.

- Julia Angwin et al., "Machine Bias" (*ProPublica*)
- Deborah Hellman, *What Makes Discrimination Wrong?*, Chapter I ("The Basic Idea")

23. Wed 11/18 Algorithmic Discrimination II

Long responds to ProPublica's investigation. Forman explores the causes of mass incarceration.

- Robert Long, "Fairness in Machine Learning: Against False Positive Rate Equality as a Measure of Fairness," sections 1-3
- James Forman, Jr., *Locking Up Our Own*, pages 3–8.5

24. Mon 11/23 Algorithmic Discrimination III

Long and Forman, continued.

- Robert Long, "Fairness in Machine Learning: Against False Positive Rate Equality as a Measure of Fairness," section 4 and 5
- James Forman, Jr., *Locking Up Our Own*, pages 8.5–14

25. Wed 11/25 The End of Work

Lepore considers whether robots are coming for our jobs. Gheaus and Herzog examine what makes work valuable to us.

- Jill Lepore, "Are Robots Coming for Your Job?" (*The New Yorker*)
- Anca Gheaus and Lisa Herzog, "The Good of Work (Other Than Money!)"

26. Mon 11/30 No Class (Thanksgiving Recess)

27. Wed 12/2 Conclusion and Exam Review *Paper 2 due*

- No readings

FINAL EXAM