Philosophy of Artificial Intelligence

Semester: Spring 2025	**Time:**
Room:	**Course Website:** `Canvas`

Instructor: Jamee Elder	**TA:**
Email: `jamee.elder@tufts.edu`	**Email:**
Office Hours:	**Office Hours:**
Location: 220 Miner Hall	**Location:**

This syllabus was last updated October 21, 2024.

1 Course Description

Artificial Intelligence (AI) is rapidly transforming our world, sparking both hopes for revolutionary advancements and fears of dystopian futures. This course invites students to explore the profound philosophical questions that arise from the development and application of AI technologies: Can AI ever be truly conscious, or is it simply mimicking human behavior? As AI takes on roles in our homes and workplaces, how will it reshape our social and moral landscapes? Are we prepared for the ethical challenges posed by autonomous systems like self-driving cars and AI-driven weapons? And who bears responsibility when these systems cause harm? By engaging with classic and contemporary readings, alongside thought-provoking science fiction, students will critically examine these issues and more. The course will also encourage hands-on experimentation with AI tools, challenging students to reflect on their experiences and consider the real-world implications of AI in society. This course will equip students with the knowledge and analytical skills to critically engage with ongoing debates about AI.

Note: this course description was revised in August 2024 using ChatGPT (OpenAI, 2024)

Contents

2 Learning Goals

By the end of this course, students will be able to:

- Craft and critique philosophical arguments.
- Apply philosophical theories and arguments to real-world AI applications.
- Appreciate the relationship between ethical and social values (e.g., toward the value of work, distribution of resources, accessibility) and the design and implementation of technology.

3 Text(s)

There are no required textbooks. All required course readings will be uploaded to Canvas or will be freely available online. In some cases, you may need to use your Tufts credentials to access something through the Tufts library. Please feel free to ask me if you need help accessing any of the readings.

4 Course Outline

All readings listed here are subject to change. The current version of this document will always be available on Canvas.

4.1 Topics Overview

		Date	Topic
Preliminaries	Week 1		
		16 January	Syllabus & Intros
(How) can machines think?	Week 2	21 January	Turing Tests
		23 January	
	Week 3	28 January	The Chinese Room Argument
		30 January	
	Week 4	4 February	The Case of Large Language Models
		6 February	
	Week 5	11 February	TBA: (Current issues)
		13 February	
(How) can we learn with AI?	Week 6	18 February	Black Boxes
		20 February	
	Week 7	25 February	Explainability vs. Interpretatbility
		27 February	
	Week 8	4 March	TBA: (Current issues)
		6 March	
	Week 9	11 March	AI in the Physical Sciences
		13 March	
Spring Break			
(How) should we use AI?	Week 10	25 March	Algorithmic Bias and Fairness
		27 March	
	Week 11	1 April	Robots at War and at Home
		3 April	
	Week 12	8 April	Human Autonomy and Privacy
		10 April	
	Week 13	15 April	Automation and Work
		17 April	
Wrapping Up	Week 14	22 April	The Singularity or the Apocalypse?
		24 April	

4.2 Detailed Schedule

All readings are subject to change.

Week 1: Syllabus & Introductions

Thursday:

- No required reading.

Unit 1. (How) Can Machines Think?

Week 2: Turing Tests

Tuesday:

- A.M Turing. Computing machinery and intelligence. In *Brain Physiology & Psychology*, pages 213–242. University of California Press, United States, 2023

Thursday:

- Thomas Nagel. *What is it like to be a bat?* Oxford University Press, New York, NY, 2024, Chapter 1: "What Is It Like to Be a Bat?"

Week 3: The 'Chinese Room' Argument

Tuesday:

- John R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3): 417–424, 1980. doi: 10.1017/S0140525X00005756

Thursday:

- Margaret A. Boden (1988). Escaping From The Chinese Room. In John Heil (ed.), *Computer Models of Mind*. Cambridge University Press, pp. 253-266.

Week 4: Large Language Models

Tuesday:

- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, New York, NY, USA, 2021. ACM

Thursday:

- Chalmers: https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/

Week 5: TBA (current issues)

Tuesday:

- TBA

Thursday:

- TBA

Unit 2. (How) Can We Learn With AI?

Week 6: Black Boxes

Tuesday:

- Buckner, C. (2019). Deep learning: A philosophical introduction. Philosophy Compass, 14(10), 1-19

Thursday:

- Emily Sullivan. Understanding from Machine Learning Models. *The British journal for the philosophy of science*, 73(1):109–133, 2022

Week 7: Explainability & Interpretability Tuesday:

- Creel, Kathleen A. 2020. "Transparency in Complex Computational Systems." *Philosophy of Science* 87 (4): 568–589. https://doi.org/10.1086/709729

Thursday:

- Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019. doi: 10.1038/s42256-019-0048-x

Week 8: TBA (current issues)

Tuesday:

- TBA

Thursday:

- No class (JE away at conference)

Week 9: AI in the Physical Sciences

Tuesday:

- Guest, Dan, Kyle Cranmer, and Daniel Whiteson. 2018. "Deep learning and its application to LHC physics." *Annual Review of Nuclear and Particle Science* 68:161–181.
- King, Martin. (draft: access via Canvas)

Thursday:

- Doboszewski, Juliusz and Jamee Elder (draft: access via Canvas)

Unit 3. (How) Should We Use AI?

Week 10: Algorithmic Bias & Fairness

Tuesday:

- John W. Patty and Elizabeth Maggie Penn. Algorithmic fairness and statistical discrimination. *Philosophy Compass*, 18(1), 2023

- (OPTIONAL BACKGROUND) Watch "Algorithmic Bias and Fairness: Crash Course AI #18" for a more accessible introduction to key ideas.

Thursday:

- Binns, R. (2018). What can political philosophy teach us about algorithmic fairness? IEEE Security & Privacy, 3, 16, 73-80.

Week 11: Robot Companions and Carers

Tuesday:

- Shannon Vallor. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press, New York, 2016, Chapter 9: "Robots at War and at Home: Preserving the Technomoral Virtues of Care and Courage"
- (OPTIONAL BACKGROUND) Shannon Vallor. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press, New York, 2016, Chapter 6: "Technomoral Wisdom for an Uncertain Future: 21st Century Virtues"

Thursday:

- "Simulacrum" in Ken Liu. *The Paper Menagerie and Other Stories*. Saga Press, London, 2016

Week 12: Human Autonomy & Privacy

Tuesday:

- Prunkl, C. "Human Autonomy and Artificial Intelligence: A Philosophical Perspective." Minds and Machines. (LINK)

Thursday:

- Sax, M. (2018) "Privacy from an Ethical Perspective", The Handbook of Privacy Studies. An Interdisciplinary Introduction, Bart van der Sloot & Aviva de Groot (ed.), Amsterdam University Press. (LINK)

Week 13: Automation and Work

Tuesday:

- James, A. (2020). Planning for Mass Unemployment. Chapter 6 of Ethics of Artificial Intelligence, Oxford University Press, 183-211

Thursday:

- Angela Watercutter. AI, the WGA Strike, and What Luddites Got Right. *Wired*, 2023. URL https://www.wired.com/story/wga-strike-artificial-intelligence-luddites/
- "The Evolution of Human Science" Ted Chiang. *Stories of Your Life and Others*. Small Beer Press, Easthampton, MA, 1st ed. edition, 2010

Week 14: The Singularity or the Apocalypse?

Tuesday:

- Nick Bostrom (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, 22(2), pp. 71-85.

Thursday:

- Chapter 1: "The Six Epochs" in Kurzweil, R. (2005). *The singularity is near: When humans transcend biology.* New York: Viking.
- Allen, P. (2011). "The Singularity Isn't Near". *MIT Technology Review*, LINK

---

5 Course Requirements

5.1 Meaningful Participation (15%):

Philosophy is a group activity; together we are more than the sum of our parts. With this in mind, students are expected to carefully study all required readings for each week and come prepared to discuss them, raise questions about them, and draw attention to their strengths and weaknesses.

In addition, students will complete weekly **reading journals** and occasional **one-minute papers**, which will contribute to their 'meaningful participation' grade. These are required, but will not themselves be graded.

- **Reading journals**: By 5:00pm each Wednesday, you will submit a journal entry (via Canvas) in which you write a short paragraph about the relationship between the Tuesday and Thursday readings for that week. The goal is raise questions or thoughts about why these readings were paired for the week and explore common threads.
- **One-minute papers**: At the end of each Thursday class, we will devote a few minutes to writing "one-minute papers" in which you briefly respond to a prompt (given in class), reflecting on what we have discussed that week.

Your participation grade will be based on the consistency and quality of your contributions in class discussions, reading journals, and one-minute papers.

Attendance: You should attend every class. However, I understand that extenuating circumstances arise that can make this difficult. If you cannot attend a class or will be more than 15 minutes late, please let me know.

Any more than two unexcused absences or missed journal entries will negatively impact your participation grade.

5.2 Short Papers (40%)

You will complete two short papers, each 1000-1500 words in length, responding to a specific question or reacting to a short reading. Each of these papers will be worth 20% of your final grade, so the short papers will be worth 40% of your grade in total.

- Short Paper 1 (20%): Due **Friday, 23 February** by **5:00pm**.
- Short Paper 2 (20%): Due **Friday, 29 March** by **5:00pm**.

5.3 Final Paper (45%):

The final paper (approx. 3000 words) will be due on **Friday, 3 May** by **5:00pm**. For this paper, you will choose a recent debate or controversy about a particular kind of digital technology and subject it to careful analysis, based on independent research.

This project will be broken up into several steps:

1. A research question and annotated bibliography (15%)
    - due **Friday, 8 March** by **5:00pm**

2. A first draft that argues for a particular answer to your research question (this should be submitted via canvas by the due date but will not be graded. Instead, you will get feedback via in-class peer-review)
    - due **Friday, 12 April** by **5:00pm**
3. A peer review assignment, where you provide constructive feedback on peer drafts and submit a brief report reflecting on the peer-review process (5%)
    - due **Friday, 19 April** by **5:00pm**
4. Final paper (25%)
    - due **Friday, 3 May** by **5:00pm**

Further instructions for writing this paper will be distributed separately.

Note that all major assignment deadlines appear in the appropriate place in the following course outline, indicated by a "⋆⋆⋆"

5.4 Letter Grade Conversion

>= 93.00	A	73.00 - 76.99	C
90.00 - 92.99	A-	70.00 - 72.99	C-
87.00 - 89.99	B+	67.00 - 69.99	D+
83.00 - 86.99	B	63.00 - 66.99	D
80.00 - 82.99	B-	60.00 - 62.99	D-
77.00 - 79.99	C+	<= 59.99	F

What do these grades mean?

According to Tufts, letter grades such as As, Bs, and Cs, translate as follows:

A. Superior work.

B. Meritorious work.

C. Work without marked merit or defect.

D. Unsatisfactory work but allowable for credit, subject to the restrictions specified under the requirements for graduation. Some departments disallow credit toward the concentration requirement.

In this class, work in the A range must be clear, accurate and insightful, both in the interpretation of other texts and in their arguments for original claims. (Note: A final grade of A+ will be given rarely, and only for consistently exceptional work). Work in the B range is good work that is mostly clear, accurate, and may offer good analysis that lacks the original insight of an A paper. Work in the C range is satisfactory work that may be unclear in places, contain inaccuracies, or lack original analysis. Assignments in the D range may be incomplete, or deficient in clarity and accuracy.

6 Course Policies

6.1 Classroom Learning Agreement

6.2 Late or Incomplete Work

We all live busy lives outside of the classroom and we each face our own unique challenges. I understand that these challenges will sometimes make it difficult to complete class assignments or to show up for class ready to make our best contributions.

Please let me know as soon as possible (and ideally before the deadline) if you know that you will struggle to meet a deadline. When you do, we can determine a reasonable timeline for you to complete the assignment or, under some circumstances, an alternative way for you to demonstrate your learning

All students are entitled to three no-questions-asked 24-hour extensions, which can be used for any major assignment (short papers, annotated bibliography, final paper, peer review report). You still need to let me know if you want to use one of these extensions! Further extensions are at the instructor's discretion, but I promise to be as flexible as possible in offering reasonable extensions.

Late work without such an extension will be penalized by 1/3 of a letter grade per day (so, e.g., an A- handed in one day late would become a B+).

7 Other Policies and Resources

7.1 Tufts Academic Resources

The StAAR Center offers a variety of FREE resources, available to *all students*. Students may make an appointment to work on any writing-related project or assignment, attend subject tutoring in a variety of disciplines, or meet with an academic coach to hone fundamental academic skills like time management or overcoming procrastination. Students can make an appointment for any of these services by visiting go.tufts.edu/TutorFinder, or by visiting their website: https://students.tufts.edu/staar-center

7.2 Accommodations

Tufts is committed to providing equal access and support to all qualified students through the provision of reasonable accommodations. If you have a disability that requires reasonable accommodations, you are encouraged to contact the StAAR Center at:

- StaarCenter@tufts.edu or
- 617-627-4539.

Please be aware that accommodations cannot be enacted retroactively, making timeliness a critical aspect for their provision.

In addition, students with or without a formally documented disability are warmly encouraged to contact me about accommodations. I am committed to collaborating with students to ensure that my course does not present unreasonable or inequitable barriers to their success.

7.3 Religious Accommodations

Tufts University faculty, staff, and administration highly value and acknowledge the religious diversity of its student body. Students seeking religious accommodations related to their holy days are encouraged to collaborate with faculty to make arrangements during the first week of each semester. The religious holy days calendar, including the holy days policy from the Faculty Handbook, is available here for your reference. Students seeking additional support may refer to the University Religious Accommodations Policy, available here. The University Chaplaincy is also available to respond to questions on religious observances; their contact information is available here.

7.4 Academic Integrity

All members of the Tufts community are responsible for integrity in their own behavior and for contributing to an overall environment of integrity at the university. You can find resources relating to academic integrity in the Tufts Academic Integrity handbook (click here). It is your responsibility to familiarize yourself with the requirements of ethical behavior and academic work as described in Tufts' Academic Integrity handbook.

If you ever have a question about the expectations concerning a particular assignment or project in this course, please ask me for clarification.

The Faculty of the School of Arts and Sciences and the School of Engineering are required to report suspected cases of academic integrity violations to the Dean of Student Affairs Office. If I suspect that you have cheated or plagiarized in this class, I must report the situation to the dean.

7.5 Guidelines for the use of AI — Revise! (in collaboration with students)

In this course, you may use AI tools for your learning, just as you can collaborate with your peers for things such as brainstorming, getting feedback, revising, or editing of your own work. However, you may not submit any work generated by an AI program as your own. This is a violation of Tufts Academic Integrity policies.

To help guide you in the use of AI in this course, consider the following guidelines:

- Familiarize yourself with AI tools, including that: Bias is embedded in the creation of these systems and in their output and you may encounter harmful language and ideas; AI platforms can produce inaccurate or false information with confidence (so called hallucinations, e.g, it frequently invent false references); Text from AI may closely mimic human knowledge, understanding and even human emotions; Many of these tools retain the rights to use your information and the content shared with them in a variety of ways.
- Cite all AI tools when used or referred to in assigned work. See How to Cite ChatGPT from the APA & How to Cite Generative AI from the MLA. Identify the way it contributed to your work. For example, you can include a statement that you asked an AI to "identify any grammatical or spelling errors" in your writing, or you used it to get started in thinking about topics for your paper. Any statement directly generated by an AI system should be in quotes.

- If you have questions please ask via email, in office hours or during class.

7.6 Student Support, Including Mental Health

As a student, there may be times when personal stressors or difficulties interfere with your academic performance or well-being. The Dean of Student Affairs Office offers support and care to undergraduates and graduate students who are experiencing difficulties, and can also aid faculty in their work with students. In addition, through Tufts' Counseling and Mental Health Service (CMHS) students can access mental health support 24/7, and they can provide information on additional resources. CMHS also provides confidential consultation, brief counseling, and urgent care at no cost for all Tufts undergraduates as well as for graduate students who have paid the student health fee. To make an appointment, call 617-627-3360. Please visit the CMHS website: http://go.tufts.edu/Counseling to learn more about their services and resources.

8 Other Resources

8.1 The Stanford Encyclopedia of Philosophy

Find it online here: https://plato.stanford.edu/

The SEP is a good place to start when learning about a new philosophical topic. The SEP entries are written by experts in that area and generally provide a good overview of the issues.

A good next step is often to follow up by reading the sources that the SEP entry sites on a particular issue.