PHI 2678 – Ethics and Artificial Intelligence

Spring 2023, Wednesday and Friday 8:30 AM to 10:00 AM, AC01 LR205

Professor: Danny Weltman | danny.weltman@ashoka.edu.in

Office hours: Monday and Wednesday 12:00 to 3:00 in AC01 616 (book a slot on Canvas)

About This Course: Topic and Goals

This is a course examining the ethics of artificial intelligence. AI is having an increasingly large effect throughout the world and its use raises many ethical questions, both practical and theoretical.

Below are the **goals** for this course:

- Investigate some of the main issues in the intersection of AI and ethics
- Improve your writing skills with a focus on concise summary and argumentation

Course Content

All course material is available on the course website at https://canvas.instructure.com/courses/5671032. We will begin with some introductory topics. After that, the remaining topics will be chosen by class vote. If you have suggestions for topics, please let me know.

Assignments and Grading

There are 4 kinds of assignments in this class: **reading quizzes**, **Perusall annotations**, **500 word papers**, and **revisions**. Late assignments will lose 10% of the grade for each day they are turned in late, up to a maximum of 50% off. The late penalty is calculated per hour (0.42% lost per hour).

Reading Quizzes (10% of your grade) are to help you focus on the important parts of the reading and to get instant feedback on whether you have understood the reading. There is one quiz per reading. The lowest 6 reading quiz scores will be dropped.

Perusall Annotation Assignments (18% of your grade) allow you to collaboratively read the readings by using the Perusall website. Canvas has a document detailing examples of annotations you can make on Perusall and explaining the grading system. The lowest 8 Perusall annotation assignment scores will be dropped.

500 Word Papers (42% of your grade) are short writing assignments designed to give you practice writing concisely about philosophy and to give you more detailed feedback about how well you are understanding the material. There are 8 due and the lowest 2 will be dropped.

Revisions (30% of your grade) are revised resubmissions of 500 word papers. Using my feedback and your own thoughts about how the assignment could be improved, you rewrite one of your papers with the aim of making it better. There are 6 revisions due and the lowest 3 scores will be ignored.

Grade Breakdown:

- **10% - Reading Quizzes** (26, lowest 6 dropped)
- **27% - Perusall Annotations** (26, lowest 8 dropped)
- **48% - 500 Word Papers** (8, lowest 2 dropped)
- **15% - Revisions** (6, lowest 3 dropped)

Class Grade Rubric:

100-94% = A	<77-74% = C
<94-90% = A-	<74-70% = C-
<90-87% = B+	<70-67% = D+
<87-84% = B	<67-64% = D
<84-80% = B-	<64-60% = D-
<80-77% = C+	<60-0% = F

Disabilities

If you have disabilities which require some form of accommodation, contact me ahead of time.

Plagiarism and Academic Integrity

Any time you use **words**, **phrases**, **ideas**, or **anything else** in your writing that you did not think up on your own, you must **cite** your source the best of your ability. Words and phrases not written by you must be enclosed in quotation marks to show that you did not write them yourself. Failure to cite a source is **plagiarism** and it's not okay. You should not need to use (or cite) outside sources for this class, but if you do use them, you must cite them. It is perfectly okay to use points made by your classmates (or anyone else), *as long as you cite them to the best of your ability.* The one exception is that you do not need to cite me on your writing assignments in this class, unless you want to. Plagiarism or other academic integrity violations, like hacking into Canvas and changing your grade to an A++, may result in a zero on the assignment or in the course.

Office Hours, Email Communication, and Due Date Extensions

If you have questions or comments about the course it is best to talk during office hours. Canvas has a link for reserving office hours meeting times. If you cannot attend any office hours, email me to set up an alternative time to meet. When you contact me via email, please include "PHI 2678" in the subject line so that I know you are emailing about this course. All of the assignment due dates are available in advance, so if you anticipate not having enough time to do the assignment right before it is due, ideally you should do the assignment earlier, rather than waiting to fall behind and then asking for an extension.

Resources

My website has resources on reading, writing, and researching at dannyweltman.com/resources.html. These resources include a glossary for unfamiliar words or phrases, some of which occur in some of the readings for this course. I encourage you to examine these resources. Canvas has additional links with resources related to AI ethics in particular.

Schedule

Each day's reading has an accompanying reading quiz and Perusall annotation assignment, both of which are due 8 AM that day. You should aim to complete them much earlier than 8 AM. You have 24 hours after the due date to add replies to comments that others have written on Perusall. Each reading also has an accompanying lecture. Lectures for the readings should be watched before the readings, although sometimes the lecture will note that it can be watched after the reading instead.

Jan 25: Shevlin, Vold, Crosby, and Halina, "The limits of machine intelligence"

Jan 27: Cave, Nyrup, Vold, and Weller, "Motivations and Risks of Machine Ethics"

Feb 1: Bostrom, "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents"

Feb 3: Schulz, "Machine Grading and Moral Learning"

Feb 8: Christiano, "Algorithms, Manipulation, and Democracy"

Feb 10: Hull, "Successful failure: what Foucault can teach us about privacy self-management in a world of Facebook and big data"

Feb 15: Benn and Lazar, “What’s Wrong with Automated Influence”

Feb 17: Talbot, Jenkins, Purves, “When Robots Should Do the Wrong Thing”

The rest of the schedule will be determined via voting on topics.

Possible Topics

The readings listed below for each topic may vary to some degree once we pick our final topics, so as to accommodate enthusiasm for topics and a sensible reading and assignment schedule.

AI Agency and Responsibility: Popa, “Human Goals Are Constitutive of Agency in Artificial Intelligence (AI)”; Matthias, “The responsibility gap: Ascribing responsibility for the actions of learning automata”; Formsa and Ryan, “Making moral machines: why we need artificial moral agents”

AI Moral Patients: Gordon and Gunkel, “Moral Status and Intelligent Robots”; Shevlin, “How Could We Know When a Robot was a Moral Patient?”; Müller, “Is it time for robot rights? Moral status in artificial entities”; Bryson, “Patiency is not a virtue”; Neely, “Machines and the Moral Community”; Agar, “How to Treat Machines that Might Have Minds”

AI Governance: McEvoy, “Political Machines: Ethical Governance in the Age of AI”; Erman and Furendal, “Artificial Intelligence and the Political Legitimacy of Global Governance”; Butcher and Beridze, “What Is the State of Artificial Intelligence Governance Globally?”; Cihon, Maas, and Kemp, “Fragmentation and the Future: Investigating Architectures for International AI Governance”; Danaher, “The Threat of Algocracy: Reality, Resistance and Accommodation”; an overview for me: Mapping global AI governance: a nascent regime in a fragmented landscape

Algorithm Bias: Johnson, “Algorithmic Bias: On the Implicit Biases of Social Technology”: Caliskan, Bryson, and Narayanan, “Semantics derived automatically from language corpora contain human-like biases”; Waller and Waller, “Assembled Bias: Beyond Transparent Algorithmic Bias”

Algorithm Fairness: Creel and Hellman, “The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems”; Zimmermann and Lee-Stronach, “Proceed with Caution”; Hedden, “On Statistical Criteria of Algorithmic Fairness”; Lipper-Rasmussen, “Using (Un)Fair Algorithms in an Unjust World”; Holm, “The Fairness in Algorithmic Fairness”; Castro and Loi, “The Fair Chances in Algorithmic Fairness: A Response to Holm”

Algorithm Values: Johnson, “Are Algorithms Value-free? Feminist Theoretical Virtues in Machine Learning”

Autonomous Vehicles: Gogoll and Muller, “Autonomous Cars: In Favor of a Mandatory Ethics Setting”; Bhargava and Kim, “Autonomous Vehicles and Moral Uncertainty”

Black Boxes: Burrell, “How the machine ‘thinks’: Understanding opacity in machine learning algorithms”; Vredenburgh, “Freedom at Work: Understanding, Alienation, and the AI-Driven Workplace”

Criminal Law: Jorgensen, “Algorithms and the Individual in Criminal Law”; Scantamburlo et al., “Machine Decisions and Human Consequences”

Human Responsibility for AI: Verdicchio and Perin, “When Doctors and AI Interact: on Human Responsibility for Artificial Risks”; Binns, “Algorithmic Accountability and Public Reason”

Killer Robots: Sparrow, “Killer Robots”; Robillard, “No Such Thing as Killer Robots”

> **Killer Robots + Trust:** Roff and Danks, ““Trust but Verify”: The Difficulty of Trusting Autonomous Weapons Systems”

Medical AI: Alvarado, “Should we replace radiologists with deep learning? Pigeons, error and trust in medical AI”; Palmer and Schwan, “Beneficent dehumanization: Employing artificial intelligence and carebots to mitigate shame-induced barriers to medical car”; Weissglass, “Contextual bias, the democratization of healthcare, and medical artificial intelligence in low- and middle-income countries”

> **Medical AI + Black Boxes:** Pierce et al., “A riddle, wrapped in a mystery, inside an enigma: How semantic black boxes and opaque artificial intelligence confuse medical decision-making
>
> **Human Responsibility for AI + Medical AI:** Sand et al, “Responsibility beyond design: Physicians’ requirements for ethical medical AI”
>
> **Algorithm Fairness + Medical AI:** Holm, “Handle with care: Assessing performance measures of medical AI for shared clinical decision-making”

Moral Robots: Allen et al., “Prolegomena to any future artificial moral agent”; Bonnemains et al., “Embedded ethics: some technical and ethical challenges”; van Wynsberghe and Robbins, “Critiquing the Reasons for Making Artificial Moral Agents”

Trust: Al, “(E)-Trust and Its Function: Why We Shouldn’t Apply Trust and Trustworthiness to Human – AI Relations”; Alvorado, “What kind of trust does AI deserve, if any?”

> **Trust + Medical AI:** Duran and Jongsma, “Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI”; Ferrario, “Design publicity of black box algorithms: a support to the epistemic and ethical justifications of medical AI systems”; Hatherly, “Limits of Trust in Medical AI”; Ferrario, Loi, and Vigano, “Trust Does Not Need to Be Human: It Is Possible to Trust Medical AI”; Starke and Ienca, “Misplaced Trust and Distrust: How Not to Engage with Medical Artificial Intelligence”; Nickel, “Trust in medical artificial intelligence: a discretionary account”

Value Alignment: Gabriel and Ghazavi, “The Challenge of Value Alignment: From Fairer Algorithms to AI Safety”