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 Tuesday 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 have been chosen by class vote. If you have suggestions for topics for next time I teach the course, 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. Each reading also has an accompanying lecture. Lectures for the readings should be watched before the readings.

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"

Feb 22: Sparrow, “Killer Robots”

Feb 24: Robillard, “No Such Thing as Killer Robots”

March 1: Taylor, “Who Is Responsible for Killer Robots? Autonomous Weapons, Group Agency, and the Military-Industrial Complex”

March 3: Adams and Barrie, “The bureaucratization of war: moral challenges exemplified by the covert lethal drone”

March 15: Santoni de Sio and van den Hoven, “Meaningful Human Control over Autonomous Systems”

March 17: Binns, “Algorithmic Accountability and Public Reason”

March 20: 500 Word Papers on Killer Robots due at midnight

March 22: Gogol and Müller, “Autonomous Cars: In Favor of a Mandatory Ethics Setting”

March 24: Woollard, “The New Trolley Problem: Driverless Cars and Deontological Distinctions”

March 27: 500 Word Papers on Human Responsibility for AI due at midnight

March 29: van Wynsberghe and Robbins, “Critiquing the Reasons for Making Artificial Moral Agents”

March 31: Formosa and Ryan, “Making moral machines: why we need artificial moral agents”

April 3: 500 Word Papers on Self-Driving Cars due at midnight

April 5: Danaher, “The Threat of Algocracy”

April 12: McEvoy, “Political Machines: Ethical Governance in the Age of AI”

April 14: Erman and Furendal, “Artificial Intelligence and the Political Legitimacy of Global Governance”

April 15: Chomansky, “Legitimacy and automated decisions”

April 17: 500 Word Papers on Artificial Moral Agents due at midnight

April 19: Scantamburlo et al., “Machine Decisions and Human Consequences”

April 24: 500 Word Papers on AI Governance due at midnight

April 26: Jorgensen, “Algorithms and the Individual in Criminal Law”

April 28: Cen, “The Right to be an Exception in Data-Driven Decision-Making”

April 29: Free day to plan your resistance to the eventual robot uprising

May 1: 500 Word Papers on Criminal Law due at midnight if you want to revise them

May 13: 500 Word Papers on Criminal Law due at midnight if you don’t want to revise them