GEORGETOWN UNIVERSITY

ethics lab

PHIL-2102 // Fall 2024

Ethical Challenges of AI

Instructor: Dr. Minji Jang (she/her)
minji.jang@georgetown.edu

Schedule: Mon / Wed 11:00 am – 12:15 pm

Location: Ethics Lab, Healy 201b

Conversation Hours:

- Wed 2:00 - 4:00 pm or by appointment in Healy 405 or via Zoom (sign up here!)

This syllabus is subject to change. Last updated: Sep 24, 2024

Course Description

The rapid advancement of AI systems presents new and profound ethical challenges demanding our immediate attention. In this course, we will explore critical ethical issues emerging from the pervasive use of AI technologies and tools that have become—or soon will become—integral to our daily lives. In doing so, we will practice essential skills and learn key ethical concepts, theories, and frameworks that will help us responsibly navigate through and make informed decisions about these complex issues.

For instance, should we entrust the consequential decisions about criminal justice and employment to algorithms that are opaque and prone to bias? In cases of emergency, should automated vehicles prioritize the safety of passengers or pedestrians? Can AI systems become members of our moral community? Can we form meaningful relationships—such as love and friendship—with them? (How) will increasing automation reshape our values and meaning of our lives? How might technologies enable us to transcend our limits (e.g., mortality)? What can—and should—we do *now* to prepare for a safer future?

Learning Objectives

This course has four learning objectives.

(1) Foundational understanding	Students will develop a foundational understanding of the key topics, issues, concepts, theories, and frameworks in AI (and) ethics.
(2) Ethical competency	Students will develop a capacity to identify and analyze high-stakes ethical issues related to the design, development, use, and governance of AI. This includes: (i) understanding the ethical stakes and complexities of these issues, (ii) identifying key stakeholders and their different (often competing) perspectives, (iii) identifying and assessing the values embedded in these issues, including their potential conflicts and the need for costly trade-offs.
(3) Philosophical skills	Students will practice key philosophical skills, including: (i) identifying and evaluating different perspectives on a complicated issue, (ii) forming your own views on the issue based on such evaluation, (iiii) effectively articulating and defending your views in writing and in speech, while properly qualifying and acknowledging its limits, and (iv) engaging in a respectful dialogue with interlocutors with different opinions and experiences.
(4) Critical inquiry	Students will critically examine, call into question, and update their understanding of the concepts, practices, and assumptions around and about their environment. The goal is to cultivate a mindset of 'healthy skepticism'—instead of accepting the current state of affairs and taking familiar things for granted, students will understand that how things were or presently are need not be how things ought to be.

Course Requirements

Grade Breakdown

Participation	10%	*Daily*
Discussion Facilitation	5%	*Twice a semester*

Text Annotation	15%	*Throughout the semester*
Portfolio	20%	*Throughout the semester*
Two Papers	30%	*See the schedule below for due dates*
Ethical Analysis Report	20%	*See the schedule below for due dates*

Participation (10%)

You are expected to attend each class, carefully having done the assigned reading (or watched the video), and prepared to engage with the material through class discussions and a series of in-class activities and exercises (including design engagements).

Your participation grade will be graded on the following four criteria:

(1) Attendance / punctuality

You are expected to show up in class and be there on time. You can have **four** unexcused absences throughout the semester, no questions asked. Barring exceptional circumstances, excusing absences will require some form of formal documentation, which can be tricky and/or annoying. I thus strongly encourage you to save these for the days when you are ill.

(2) Alertness / attentiveness

You are expected to be awake, alert, and attentive in class and not working on anything else (e.g., checking your mails/calendar, messaging, surfing, doing other readings or homework for other courses, etc.).

(3) Respectfulness

You are expected to be respectful to your peers and the instructor. This includes carefully listening to what other people say, waiting your turn to speak (instead of interrupting others), not chatting on the side, and being respectful of different opinions and perspectives of others (note that this is compatible with critically examining and responding to their views).

(4) Active participation

You are expected to actively participate in class and contribute to the collective learning experience of the group, in some ways or another. Here are some examples.

In-class participation. Actively and authentically participating in small-group and big group class discussions and in-class activities.

Conversation hours. Having thoughtful exchanges with me about class materials during conversation hours or through email exchanges. (Discussing your assignments *does not* count.)

Shout outs. Leaving a 'shout out' after class on the Canvas discussion forum. You'll (i) note **one** question or comment made by your peer during class discussions / activities that you found particularly interesting or thought provoking and (ii) briefly tell us (in 1-2 sentences) **why** you found it to be interesting/thought provoking. (You can sometimes—but not always—write about the question/comment that someone else already posted, but you'll need to say why *you* also found it to be interesting/thought provoking.)

If you have trouble participating in class through these activities for any reason, please come chat with me, and we'll discuss the ways to make things easier for you.

Discussion Facilitation (5%)

Twice a semester, you'll facilitate a small-group discussion in class. Each time, successfully completing the following three steps will give you 2.5 points (total 5 points of your overall grade):

Step 1—*before* class	Carefully do the assigned reading, and send me 3-4 questions that you'd like to discuss in your group before the class begins.
Step 2—*during* class	Facilitate the small group discussion in class. Your job is to modulate the discussion, not to dominate it. Be sure to give each person a chance to speak (unless they don't want to).
Step 3—*after* class	Send me a brief report (2-3 paragraphs) of how it went. What questions did you discuss? What ideas came up? Were you satisfied with how it went? What would you do differently next time (e.g., rephrasing a question, preparing an example, clarifying a concept, etc.)?

You'll sign up for the dates for discussion facilitation in the first few weeks of the semester.

Text Annotation (15%)

Throughout the semester, on the day when there's a new reading, you'll submit text annotations through Hypothesis directly on Canvas. You'll do this assignment on a reading, not a video, unless specifically instructed by the instructor. All text annotations are due **before** the beginning of the class.

In your annotation, you'll leave **four** comments **evenly spread** throughout the document. You can leave (i) **at least two** original comments and questions and (ii) **at least one** reply.

Original comments and questions	You'll start a thread for a new question or a comment you have about the text. For instance, you can highlight parts that personally resonated with you (and briefly tell us why), ask a clarifying question, bring up a relevant example, agree or disagree with the author, raise a concern, connect the idea to another reading/video we discussed in class, and so on.
Reply	You'll either reply to a thread started by your peer or leave a reply to your peer's question left in your thread. For instance, you can expand on the question or comment raised by your peer, try to provide an answer to their question, agree or respectfully disagree with their perspectives, add an example, and so on.

You'll need to 'pass' **a total of 15** annotations to get 100%. There are (at least) 20 classes on which there will be a new reading assigned, so you can safely skip (at least) 5 days and still get full marks.

'Optional' make-up opportunities!	There will be several fun and important AI and tech-related events on campus this semester, including ones hosted by Tech & Society, the Beeck Center for Social Impact & Innovation, the McCourt School of Public Policy, and more. If you attend one or more of these events (**max 4**) and send me (i) one photo (as evidence), and (ii) one paragraph reflection of how it went, I'll count it as **one** 'passed' text annotation.

Portfolio (20%)

Throughout the semester, you'll build a portfolio composed of short homeworks and the artifacts that you produced from in-class activities, including Design Engagements. Design Engagements are interactive activities designed by the Ethics Lab faculties to help you creatively engage with the course topics—we'll have four of them throughout the semester. You'll submit the completed portfolio at the end of the semester.

Two Papers (30%)

You'll write two papers designed to help you practice key philosophical skills: **Paper 1** (3-4 pages, double-spaced; worth 10%) and **Paper 2** (4-5 pages, double-spaced, worth 20%), plus a short explanation of how you incorporated the feedback that you received on Paper 1 into Paper 2. Detailed guidelines and rubrics for Paper 1 and 2 will be posted separately on Canvas.

Ethical Analysis Report (20%)

At the end of the semester, you'll write an ethical analysis report on an existing AI tool / product that we've discussed in class or that you've regularly interacted with in your life. You'll identify and examine how its core features/functions generate a *genuine* (i.e., hard to resolve), *high-stakes*, and *imminent* ethical challenge, make a qualified recommendation going forward, and respond to a concern raised by someone who reasonably disagrees with you. Detailed guidelines and rubrics for your Final Report will be posted separately on Canvas.

To preview, here are some of the questions that you may consider in writing this report: How and where is this product currently used? By whom? What do they need, want, or care about the most in their interactions with it? What aspects or features of this product affording or disaffording certain values? What are the potential issues, problems, or challenges emerging from its current (or projected) use? Who is benefited (the most), and who is harmed (the most)? Are there any conflicts between competing values? What kinds of trade-offs might we (or should we) make, and at what costs? Are we willing to make them? What could (or should) be the next step?

Don't worry if these questions don't make much sense to you now. You'll get more familiar with and feel more comfortable raising these questions as the semester proceeds!

Course Policies

Commitment to Safe and Inclusive Learning Environment

Together as a class, we are committed to creating a safe and inclusive learning environment for everyone involved, irrespective of their ability, age, economic status, ethnicity, first language, gender expression and identity, national origin, race, religion, sex, and sexuality. Offensive speech or discriminatory remarks on any of these bases, explicit or implicit, will not be tolerated under any circumstances.

Climate Surveys and Digital Mailbox. We'll have anonymous climate surveys to ensure that everyone feels welcomed and respected during class discussions and lectures. In addition to these surveys, if you have any climate-related concerns or suggestions that you would like to share with me, please drop a note to the Digital Mailbox at any time during the semester.

Anonymous Grading

Your papers and ethical analysis report will be graded **anonymously**. This means that you shouldn't put any identifying information on your files, other than your UID number. Your file name must be and only be your UID number. You'll also use a standardized formatting, for a distinct style may threaten anonymity: 12pt font, Times New Roman, double-spaced, with 1-inch margins. I'll dock **1/3 of a letter grade** for a submission that contains identifying information or has a non-standardized formatting.

Page Limit Policy

Your papers and ethical analysis report are subject to strict page limit policy. To standardize length, all submissions should use a standardized formatting: again, 12pt font, Times New Roman, double-spaced, with 1-inch margins. If you have a nonstandard formatting, I'll convert it to standard formatting to check for length. I'll dock **1/3 of a letter grade** for each half of a page you are over or under the required page range. (Bibliography is not included in the page range.)

Late Policy and Extension

Throughout the semester, you are given **seven "extension days"** to use **before** the last day of class (Dec 9, 2024). You can use them on any of the assignments, no questions asked, **except** Discussion Facilitation and short homeworks (for portfolio), as these are designed to be completed on time to help with class discussions, activities, and other assignments. Here are a few things to remember.

- You must email me **before** the assignment deadline to let me know how many days you'll use. (No need to tell me why.)
- Once you have used up your "extension days," I won't grant further extensions, except in exceptional circumstances. Late assignments will be docked **1/3 of a letter grade** for each calendar day (or part of a day) that it is late. Assignments more than a week late will not be accepted and will receive a zero. So, I advise you to use your "extension days" wisely.
- If you need an extended extension for an assignment, please come chat with me **before** the deadline.

Electronics

Here's our tentative electronics policy. No phones are allowed in class. (If you need to use your phone, you should feel free to let me know in advance.) Laptops are allowed **only** when indicated by the instructor (e.g., for a specific class activity) or per accommodations. If you need to access an electronic copy of your reading on your device, make sure that your Wifi/hotspot is turned off.

Course Schedule

The schedule below is subject to change. Last updated: Oct 23, 2024

Wk #	Date	Topic	Reading	Due Dates
Unit 1. Level Setting				
Wk 1	Class 1 Wed | Aug 28	*Introduction*	N/A	
Wk 2	Class 2 Tue | Sep 3	*Value-laden technology , sociotechnical systems* Case study: Facial Recognition Technology	Read: Langdon Winner, "Do Artifacts Have Politics?" (2007) until pp. 128	
	Class 3 Wed | Sep 4	*Casual intro to AI and machine learning* Guest lecture - Dr. Yo Joong Choe (UChicago)	Watch: AI, Machine Learning, Deep Learning and Generative AI Explained (2024) Watch: How Large Language Models Work (2023)	
Wk 3	Class 4 Mon | Sep 9	*Casual intro to ethical theories* Case study: Content moderation	Read: Julia Driver, "Normative Ethics" in *Ethics: The Fundamentals* (2008)	
	Class 5 Wed | Sep 11	*Ethics of AI ethics*	Read:Thomas Powers and Jean-Gabriel Ganascia, "The Ethics of the Ethics of AI" (2020)	
Unit 2. Bias and Opacity				
Wk 4	Class 6 Mon | Sep 16	*Bias*	Watch: Kate Crawford, "The Trouble with Bias" (2017) Read: "Amazon scraps secret AI recruiting tool that showed bias against women" (2018) 'Optional' watch: Joy Buolamwini, "How I'm fighting bias in algorithms" (2017) 'Optional' read: Timnit Gebru, "Race and	

			Gender" (2020)	
	Class 7 Wed | Sep 18	*Fairness*	Read: Ben Green, "The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness" (2020) 'Optional' read: ProPublica, "Machine Bias" (2016)	
Wk 5	Class 8 Mon | Sep 23	*Surveillance, privacy, consent*	Read: Georgetown Law Center on Privacy and Technology - Clare Garvie & Laura M. Moy, "America Under Watch" (2019) **and** Jameson Spivack, "Cop Out: Automation in the Criminal Legal System" (2023) 'Optional' read: "Woman sues Detroit after facial recognition mistakes her for crime suspect" (2023) 'Optional' watch: "China's Surveillance State Is Growing" (2022)	
	Class 9 Wed | Sep 25	*Opacity, transparency, interpretability*	Read: Maya Krishnan, "Against Interpretability" (2019)	
Wk 6	Class 10 Mon | Sep 30	***Engagement #1***	N/A	
	Class 11 Wed | Oct 2	***Paper Writing Workshop #1***	N/A	
Unit 3. Responsible (Use of) AI				
Wk 7	**No class** Mon | Oct 7	N/A	N/A	
	Class 12 Wed | Oct 9	*Deep fakes, fake news*	Read: Regina Rini, "Deepfakes and the Epistemic Backstop" (2020)	
Wk 8	**No class** Mon | Oct 14	** Happy Fall break! **		**Paper 1** due Mon | Oct 14
	Class 13	*Responsibility gap*	Read: Madeleine Elish, "Moral Crumple	

	Wed | Oct 16		Zones" (2016)	
Wk 9	Class 14 Mon | Oct 21	*Self-driving car*	Read: Sven Nyholm &· Jilles Smids - "Automated Cars Meet Human Drivers: Responsible Human-Robot Coordination and the Ethics of Mixed Traffic" (2020) Watch: Patrick Lin, "The ethical dilemma of self-driving cars" (2016)	
	Class 15 Wed | Oct 23	*Clinical use*	Read: Charles Binkley & Tyler Loftus, "The Patient and AI Clinical Decision Support Systems" (2024)	
Unit 4. Recent Topics in AI Safety				
Wk 10	Class 16 Mon | Oct 28	*Alignment*	Watch: Robert Miles - Intro to AI Safety (2021) Read: "When Generative AI Refuses To Answer Questions, AI Ethics And AI Law Get Deeply Worried" (2023) 'Optional' Read: Illustrating Reinforcement Learning from Human Feedback (RLHF) (2022)	
	Class 17 Wed | Oct 30	***Engagement #2*** Guest instructor - Prof. Akshaya Narayanan (Ethics Lab)	Read: Elon Musk unveils Tesla Cybercab self-driving robotaxi (2024) - No Hypothesis	
Wk 11	Class 18 Mon | Nov 4	*Preparing for future*	Read: A Right to Warn about Advanced Artificial Intelligence (2024) Read: "4 Ways to Advance Transparency in Frontier AI Development" (2024) Read: Pause Giant AI Experiments: An Open Letter (2023)	
Unit 5. AI Companions				
	Class 19	*Love and friendship*	Read: Sven Nyholm & Lily Frank, "It Loves	

	Wed | Nov 6		Me, It Loves Me Not" (2019) 'Optional' read: Safiya Umoja Noble, "Your Robot Isn't Neutral" (2021)	
Wk 12	Class 20 Mon | Nov 11	***Paper Writing Workshop #2***	N/A	
	Class 21 Wed | Nov 13	*AMAs*	Read: David Chalmers, "Could a Large Language Model be Conscious?" (2023)	
Wk 13	Class 22 Mon | Nov 18	*AMAs, cont'd*	Read: David Gunkel, "Why Consciousness is Neither a Necessary nor Sufficient Condition for AI Ethics" (2019)	**Paper 2** due Fri | Nov 15
Unit 5. Personal Identity and Meaning in Life				
	Class 23 Wed | Nov 20	*Uploading*	Read: Jim Pryor, "What's So Bad about Living in the Matrix?" (2005) Read: David Chalmers (2010), "Uploading and Personal Identity" (2010)	
Wk 14	Class 24 Mon | Nov 25	*Automation, achievement, and meaning in life*	Read: John Danaher & Sven Nyholm, "Automation, work and the achievement gap" (2020)	
	No class Wed | Nov 27	** Happy holidays! **		
Wk 15	Class 25 Mon | Dec 2	*Digital immortality*	Watch: Black Mirror, "San Junipero" (S3E4)	
	Class 26 Wed | Dec 4	***Engagement #3***	N/A	
Wk 16	Class 27 Mon | Dec 9	*Wrapping up*	N/A	**Portfolio** due Mon | Dec 9
	Tue | Dec 10 - Thu | Dec 12	** Study days **		
Wk 17	Fri | Dec 13 - Sat | Dec 21	** Exam days **		**Report** due Mon | Dec 16