Yale

The Ethics of AI

EP&E TBA
Fall 2024

Meeting Information

Days of Week & Times: TBD
Location: TBD

Instructor

Max Lewis (he/him)
Office: Allwin Hall 201
Preferred contact method: max.lewis@yale.edu
I aim to respond to emails within 24 hours during the week and 48 hours on the weekend.

Office Hours

Weekly office hours are a dedicated time that I am available to answer your questions, discuss course content, and generally be of support. Please drop in or sign up for a slot on Canvas to attend office hours on Zoom or in person. If you would like help in the course but have a scheduling conflict that prevents you from attending my regular office hours, please email me to schedule an appointment. Talking with students is a highlight of my job, so feel free to come chat! I look forward to speaking with you!

Course Description

In this immersive course, we explore the ethical challenges shaping the AI revolution. From understanding AI's foundations to debating its most controversial uses, we critically engage with the pressing questions that define the future of technology and society. We explore three major sets of questions. First, we look at the moral permissibility of using and interacting with AI: Are AI algorithms biased, and if so, should we still rely on them? Do digital surveillance systems violate our right to privacy? Is it moral to use AI and robots in warfare? Could advanced AI even have moral rights? Second, we look at moral responsibility and AI: When AI causes harm—such as in war, self-driving car accidents, and medical misdiagnoses—who should bear the responsibility? Finally, we look at AI and personal relationships: Is it wrong to use AI for grief counseling, love letters, wedding vows, or eulogies? Can true friendship or romance exist between humans and machines? This course gives you the opportunity to think critically, debate passionately, and gain the conceptual, technical, and ethical tools to navigate the AI-driven future.

Format

This course is discussion-focused. In each class, we will have in-depth discussions of each of the week's readings. We will also regularly engage in group-based in-class activities. Research overwhelmingly indicates that these kinds of learning activities significantly increase individuals' critical thinking and problem-solving abilities.

Learning Objectives

This course aims to help students hone their abilities:

1. To identify, recapitulate, and critically analyze arguments.
[Through class discussions, participation in in-class active learning, Discussion posts, Leading discussion Midterm Paper, and Final Paper]

2. To grasp, express, and contrast the main concepts and arguments concerning the collection of moral and social controversies we will discuss by using concepts that structure the debates we'll read, e.g., rights, the concept of a person, killing vs letting die, consequentialism, deontology, AI, LLMs, Machine Learning, etc.
[Through class discussions, participation in in-class active learning, Discussion posts, Leading discussion Midterm Paper, and Final Paper]

3. To critically analyze the strong and weak points of the major positions such as utilitarianism, Kantianism, deontology, conservative and liberal views of AI, privacy vs surveillance, responsibility and accountability, etc.
[Through participation in in-class active learning, Discussion posts, Midterm Paper, and Final Paper]

4. To develop and be able to express their own educated opinions of these issues.
[Through class discussions, participation in in-class active learning, Discussion posts, Leading discussion, Midterm Paper, and Final Paper]

5. To apply the lessons from the debates we will examine to issues in their own lives and to those concerning public life.
[Through writing Discussion posts, Leading discussion, Midterm Paper, and Final Paper]

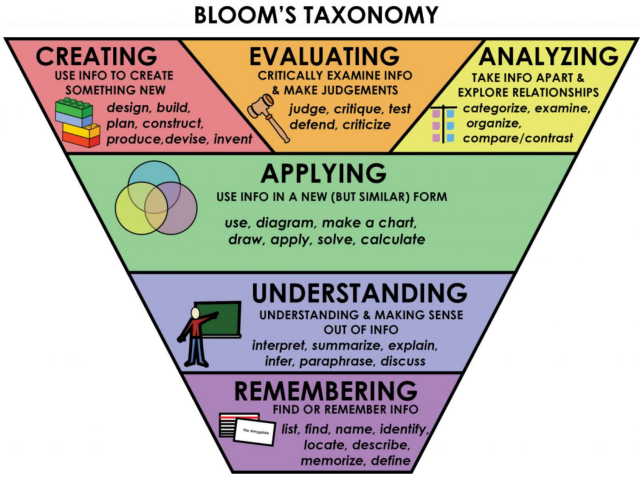
Required Materials

All required course materials can be accessed on the course's Canvas page under "Files" and "Schedule."

Assessment and Grading

Levels of Learning

All of the assignments in this course are designed to help you achieve different levels of learning.



Forms of Assessment

Participation – 10%
Reading Engagement and Discussion – 20%
Leading a Discussion – 10%
Midterm Paper (4-6 pages) – 25%
Final Paper Outline – 10%
Final Paper (5-7 pages) – 25%

Participation

Attendance is mandatory. You must attend every class and be on time. You can participate in class in many ways: (i) asking purely clarificatory questions, (ii) answering questions that I ask, (iii) participating in in-class group exercises, (iv) commenting on the remarks or questions of your classmates, and so on. You are responsible for doing the reading and knowing the material and thus part of the participation grade will be determined by demonstrating your familiarity with the readings during class. If you are not comfortable talking in class, you can earn at least some participation points by talking to me during my office hours.

Justification: This is a seminar and as such meaningful participation is essential to the success of our class. Discussion helps us not only formulate our own ideas (by expressing them), but also understand other points of view, challenge our own presumptions, work on defending our own views, and understand something better together.

Learning Kinds/Levels: Remembering (e.g., recalling the text), Understanding (e.g., being able to summarize the text), Analyzing (e.g., being able to apply the text to novel situations), Evaluating (e.g., being able to criticize and defend the text), and (perhaps Creation, e.g., being able to think of novel versions of a theory or argument).

Reading Engagement and Discussion

Before each class, you will be required to make a post in our Canvas page's discussion board. Throughout the semester, there will be **9** class meetings on which you will have to post. **Five** of these postings must be original comments (i.e., not a reply to someone else's comment) and **Four** of them must be replies to other people's posts. All postings must engage with the reading and not just be an expression of (dis)agreement with the author/poster. Engagement will often (but not always) require quoting or citing the authors (e.g., the page number). Successful posts (both original and comments) will: (a) offer a plausible criticism of some portion of the reading (or student post in the case of commenting), (b) compare or contrast the current week's readings and previous weeks' readings, (c) illustrate how the current week's reading applies to current events, (d) suggest a more plausible route (using the author's own assumptions or commitments) that an author could have taken to get to their conclusion, (e) question the meaning of a particular passage given other passages in the same reading (e.g., if it appears as if the author has contradicted themselves or two or more of their ideas seem to be in tension). Another way of successfully commenting on an original post is to suggest how, given what we've read, the author might respond to a criticism or clarificatory question.

Requirement: Postings must be made by **8pm the day before class**. This is to ensure that other students have enough time to read and respond to original posts. Postings made after 8pm will only receive partial credit (the degree of credit will be based on the quality of the posting). Postings made during or after class

meetings will receive only 50% credit. These assignments are aimed at helping you contribute to class discussion and collaborative learning as well as to help you prepare for in-class discussion. Moreover, it is unfair to give two people equal credit when one failed to meet the deadline.

Justification: This assignment will help you use what you have learned from the readings to create your own original ideas.

Learning Kinds/Levels: Remembering (e.g., being able to recall parts of the text), Understanding (e.g., being able to summarize the text), Applying (e.g., relating ideas in the text to modern phenomena), Analyzing (e.g., understanding how different arguments in the text fit together and showing alternative paths to the same conclusion), and Evaluating (e.g., criticizing or defending part of the text).

Leading a Discussion

Once during the semester, you and another student will co-lead part of our discussion. You will choose one or two passages from the week's reading, have all of us read it in class (to ourselves or out loud), explain how they fit into the overall reading for that week, prepare discussion questions, and lead a discussion for 20 minutes.

Requirement: You and your presentation partner must meet with me at least **2 days before the presentation date** to review your passage choices and questions.

Justification: (1) People gain a deeper grasp of material when they focus on a small section of it, construct their own questions, and are responsible for answering questions about it. (2) Gives you agency in the class as you get to choose the reading and passage(s) as well as which aspects of the passages to discuss, and (3) facilitates collaborative learning.

Learning Kinds/Levels: Remembering (e.g., being able to recall parts of the text), Understanding (e.g., being able to summarize the text), Analyzing (e.g., being able to understand and explain how selected passage(s) fit into the text as a whole), and Evaluating (e.g., being able to criticize or defend part of the text, asking questions that show your sense of where their might problems).

Midterm Paper

You will write a 4-6-page paper either in response to a prompt provided or on a topic that you and I agree upon beforehand. You will not be required to do any outside reading for the papers. You are allowed to rely only on class readings and in-class discussions. However, you are permitted to use outside resources, if you would like.

Learning Kinds/Levels: Understanding (e.g., being able to summarize an argument in your own words), Analyzing (e.g., understanding which parts of the text are the conclusion/thesis, which parts constitute reasons offered in favor of the conclusion, which claims function as stipulations or assumptions and being able to identify similarities and differences between different authors), Evaluating (e.g., being able to criticize a premise of the argument), Creating (e.g., providing a novel defense of some idea or connection between authors).

NOTE: In order to write on a prompt that you come up with, you will need to have a meeting with me to discuss it

at least ***one week before to the due date****.*

NOTE: If you did not receive an extension, you will receive a 3-point grade deduction for every day you are late. In addition, in cases where extensions are granted, your graded paper will generally be returned later than those who hand in the paper on time.

Final Paper Outline

You must summit a 1–2-page outline of your planned final paper telling me your thesis, your main argument in support of that thesis, and what objection you will consider against your view.

Learning Kinds/Levels: Understanding (e.g., being able to summarize an argument in your own words), Analyzing (e.g., understanding which parts of the text are the conclusion/thesis, which parts constitute reasons offered in favor of the conclusion, which claims function as stipulations or assumptions and being able to identify similarities and differences between different authors), Evaluating (e.g., being able to criticize a premise of the argument), Creating (e.g., providing a novel defense of some idea or connection between authors).

Final Paper

You will write a 5-7-page paper either in response to a prompt provided or on a topic that you and I agree upon beforehand. You will not be required to do any outside reading for the papers. You are allowed to rely only on class readings and in-class discussions. However, you are permitted to use outside resources, if you would like.

Learning Kinds/Levels: Understanding (e.g., being able to summarize an argument in your own words), Analyzing (e.g., understanding which parts of the text are the conclusion/thesis, which parts constitute reasons offered in favor of the conclusion, which claims function as stipulations or assumptions and being able to identify similarities and differences between different authors), Evaluating (e.g., being able to criticize a premise of the argument), Creating (e.g., providing a novel defense of some idea or connection between authors).

NOTE: In order to write on a prompt that you come up with, you will need to have a meeting with me to discuss it at least ***one week before to the due date****.*

NOTE: If you did not receive an extension, you will receive a 3 point grade deduction for every day you are late. In addition, in cases where extensions are granted, your graded paper will generally be returned later than those who hand in the paper on time.

Extensions, Late Submissions, and Absences

Extensions will be granted for assignments only in two cases: (1) you request one within 72 hours of the assignment due time or (2) you provide me with evidence of extenuating circumstances (e.g., family illness or death, religious holiday, personal injury or sickness, or other kinds of emergencies). Unexcused late submissions will be subject to the above-mentioned penalties.

Absences will be excused in extenuating circumstances (family illness/death, religious observance, family emergency, etc.). The same holds for showing up late or leaving early. An absence, tardiness, or early departure that is not excused will negatively affect your participation grade. With regard to any of these

situations, please contact me as soon as possible.

If you cannot physically make it to class (e.g., because you have COVID or had to leave campus for an emergency), but you want to participate anyway, please email me as soon as you know you cannot make attend. I can arrange for AV to provide us with a Meeting OwlLinks, which will allow you to hear and see us clearly and to speak (if you want).

General Advice for Doing Well

Not required:

1. Being super ready to defend your point to the death every time.
2. Making impassioned speeches.
3. Being able to recite the text from memory.

Some tips for doing well:

- Pay attention in class (obvious, but true).
- Become invested in the texts and the questions we're considering. Re-read the parts you find the most interesting.
- Carefully and patiently read the texts.
- Think deeply and carefully about what they mean and how they relate to each other.
- Come to office hours.
- Start assignments early

My Job

My fundamental job is to help you: (1) understand the questions we will ask in this course, (2) understand the texts we will read, (3) make connections between the readings, and so on.

In a nutshell: My job is to help you do well in this class.

One way I can help is during office hours. Weekly office hours are a dedicated time that I am available to answer your questions, discuss course content, and generally be of support. Please drop in or sign up for a slot on Canvas to attend office hours on Zoom or in person. If you would like help in the course but have a scheduling conflict that prevents you from attending my regular office hours, please email me to schedule an appointment. Talking with students is a highlight of my job, so feel free to come chat! I look forward to speaking with you!

In Class Use of Technology: Out of respect for your fellow classmates and me, I ask that you please not use phones during class unless you have a medical condition that warrants it. You are allowed to use laptops and tablets in class as long as they are being used for class purposes (e.g., taking notes or reviewing course readings).

Academic Integrity

Yale punishes academic dishonesty severely. The most common penalty is suspension from the university, but students caught plagiarizing are also subject to lowered or failing grades as well as the possibility of expulsion. You are responsible for understanding the university's rules regarding academic integrity. Please be sure to review Yale's Academic Integrity Policy at: http://yalecollege.yale.edu/content/cheating-plagiarism-and-documentationLinks to an external site..

Collaboration with ChatGPT or other AI composition software is not permitted in this course except for the following tasks:

- Correcting grammar or spelling
- As a thesaurus
- To check historical facts (e.g., the year or location that someone was born)
- To translate from one language to another

If there are other tasks that you think might be appropriate to use AI composition software for, please do not hesitate to contact me about them.

Diversity, Equity, Inclusion, & Belonging

I am committed to ensuring that students from all backgrounds and perspectives are equally served by this course. This requires, amongst other things, that discussions be respectful of differences in gender, sexuality, ability, age, socioeconomic status, ethnicity, race, culture, and so on. That does not mean that we cannot disagree or criticize each other's arguments. However, we must ensure to do so respectfully and humbly. Respecting diverse backgrounds and perspectives is not only the right thing to do, it is beneficial for each of our educations. Being presented with a diverse set of views is one of the best ways of learning about a topic and getting to the truth. I would encourage you to let know how I can improve the effectiveness of the course for you personally or for others. To do so anonymously, please use this form: [INSERT HYPERLINK]

Accessibility

I am committed to creating a course that is inclusive in its design. If you encounter barriers, please let me know immediately so that we can determine if there is a design adjustment that can be made or if an accommodation might be needed to overcome the limitations of the design. I am always happy to consider creative solutions as long as they do not compromise the intent of the assessment or learning activity. You are also welcome to contact Student Accessibility Services to begin this conversation or to establish accommodations for this or other courses. I welcome feedback that will assist me in improving the usability and experience for all students.

Academic & Wellness Supports

At Yale, you have important resources available for both academic and wellness support. Academic support is available in the form of both writing and language tutors. Wellness support is available in the form of acute care, social needs, mental health care and counselling and Yale Well, which is committed to your emotional, physical, social, intellectual, professional, and spiritual wellbeing. For more information and resources, please visit: https://yalecollege.yale.edu/getting-help

Readings & Assignments

Meeting 1: What is Morality?

Required Material

- Mark Timmons, "A Moral Theory Primer" from *Disputed Moral Issues* (~34)
- David McNaughton and Piers Rawling, "Deontology" (~11 pages, two column layout)

- John Stuart Mill, *Utilitarianism* (selection from *Disputed Moral Issues*) (~6 pages, two column layout)
- Immanuel Kant, *Groundwork for the Metaphysics of Morals* (selection from *Disputed Moral Issues*) (~7 pages, two column layout)

Meeting 2: What is Artificial Intelligence?

Required Material

- Vincent C. Müller, "Philosophy of AI: A Structured Overview" (~20 pages)
- S. Matthew Liao, "A Short Introduction to the Ethics of Artificial Intelligence" (~23)
- Google, "Introduction to Large Language Models": https://developers.google.com/machine-learning/resources/intro-llms (~4 pages)
- Google, "What is Machine Learning?": https://developers.google.com/machine-learning/intro-to-ml/what-is-ml (~4 pages)
- Google, "Supervised Learning": https://developers.google.com/machine-learning/intro-to-ml/supervised (~4 pages)

Optional Material

- https://1000wordphilosophy.com/2024/02/13/artificial-intelligence/
- MIT Technology Review Narrated: What is AI?

Meeting 3: Bias

Required Material

- Gabbrielle M. Johnson, "Algorithmic bias: on the implicit biases of social technology" *Synthese* (2021) 198: 9941–9961 https://doi.org/10.1007/s11229-020-02696-y (~16 pages)
- Duncan Purves and Jeremy Davis Should Algorithms that Predict Recidivism Have Access to Race? *American Philosophical Quarterly* Volume 60, Number 2, April 2023 (~13 pages 2-column layout)
- Huang, Linus Ta-Lun, Hsiang-Yun Chen, Ying-Tung Lin, Tsung-Ren Huang, and Tzu-Wei Hung. 2022. "Ameliorating Algorithmic Bias, or Why Explainable AI Needs Feminist Philosophy." *Feminist Philosophy Quarterly* 8 (3/4). (~24 pages)

Optional Material

- Lauren Leffer, "Humans Absorb Bias from AI--And Keep It after They Stop Using the Algorithm." *Scientific American* (~5 pages)

Meeting 4: Privacy

Required Material

- Andrei Marmor, "What Is the Right to Privacy?" *Philosophy & Public Affairs* 43: 3–26. doi: https://doi.org/10.1111/papa.12040 (~23 pages)
- James Stacey Taylor, "In Praise of Big Brother: Why We Should Learn to Stop Worrying and Love Government Surveillance," *Public Affairs Quarterly*, Jul., 2005, Vol. 19, No. 3 (Jul., 2005), (~15 pages).
- Carissa Véliz, "The Surveillance Delusion" in *The Oxford Handbook of Digital Ethics* (~15 pages)

Optional Reading

Mark Coeckelbergh, "Ch. 7: Privacy and the Other Usual Suspects" from *AI Ethics* (~11 pages)

Meeting 5: AI in War

Required Material

- N. Sharkey,'Killer Robots in War and Civil Society', video talk,10 August 2015.
- Sparrow,R. (2007).'Killer robots',*Journal ofAppliedPhilosophy*,24,62–77 (~13 pages)
- Simpson,T. W . and Muller, V . C. (2016) Just war androbot's killings,*Philosophical Quarterly*, 66 (263) (~20 pages)

Optional Material

- N. Sharkey (2010) Saying 'No!' to Lethal Autonomous Targeting, *Journal of Military Ethics*, 9:4,369–383

Meeting 6: Self-Driving Cars

Required Material

- Kauppinen, Antti (2023) "Who Should Bear the Risk When Self-Driving Vehicles Crash?" *Journal of Applied Philosophy*. (~14 pages)
- Frances Kamm, "The Use and Abuse of the Trolley Problem," in *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020) (~23 pages)
- Sven Nyholm. 'The ethics of crashes with self-driving cars: A roadmap, I' *Philosophy Compass* 13,e12507 (~7 pages)
- Sven Nyholm. 'The ethics of crashes with self-driving cars: A roadmap, II' *Philosophy Compass* 13,e12506 (~8 pages)

Optional Material

- Shariff, Rahwan, and Bonnefon "Whose Life Should Your Car Save?" *The New York Times* https://www.nytimes.com/2016/11/06/opinion/sunday/whose-life-should-your-car-save.html
- Derek Leben(2017). "A Rawlsian algorithm for autonomous vehicles." *Ethics and Information Technology* 19 (2):107-115. https://philpapers.org/rec/LEBARA
- Patrick Lin , "The ethical dilemma of self-driving cars," TEDtalk, 8 December 2015
- Iyad Rahwan. 'What moral decisions should driverless cars make?', TEDtalk, 8 September 2017

Meeting 7: Using AI in Human Relationships

Required Material

- John Danaher, "Toward an Ethics of AI Assistants: an Initial Framework" *Philosophy and Technology* ***31****, 629–653 (2018). https://doi.org/10.1007/s13347-018-0317-3* (~23 pages)
- Evan Selinger. (2014). Today's Apps are Turning us Into Sociopaths. *WIRED* 26 February 2014 - available at https://archive.is/464Jp (~ 4 pages)
- John Danaher, "The Case for Outsourcing Morality to AI" *Wired* (~7 pages)
- Ethical Issues with Artificial Ethics Assistants (~15 pages)
- Max Lewis, Relational Actions and AI Interpersonal Assistants (manuscript). (~22 pages)

Meeting 8: Griefbots

Required Material

- Joel Krueger and Lucy Osler, Communing with the Dead Online (~22 pages)

- Nora Freya Lindemann, "The Ethics of 'Deathbots" *Sci Eng Ethics* 28, 60 (2022). https://doi.org/10.1007/s11948-022-00417-x (~13 pages)
- Regina E. Fabry and Mark Alfano, "The Affective Scaffolding of Grief in the Digital Age: The Case of Deathbots" Topoi (2024) 43:757–769 (~11 pages two column layout)

Optional Material

- Jack Holmes, "Are We Ready for AI to Raise the Dead?" *Esquire* (~5 pages)
- *Morning Edition* , "**Ethical implications of making a chatbot using the voice or likeness of someone"** https://www.npr.org/2024/10/22/nx-s1-5154869/ethical-implications-of-making-a-chatbot-using-the-voice-or-likeness-of-someone
- *In Machines We Trust*
    - https://podcasts.apple.com/us/podcast/technology-that-lets-us-speak-to-our-dead-relatives/id1523584878?i=1000674111360

Meeting 9: Friendship and Romance with AI

Required Material

- Dean Cocking and Steve Matthews, "Unreal Friends" *Ethics and Information Technology* **2:** 223–231, 2000. (~9 pages, two column layout)
- Sven Nyholm and Lily Frank, 2017, "From Sex Robots to Love Robots: Is Mutual Love with a Robot Possible?", in Danaher and McArthur 2017: 219–243. In *Robot Sex: Social and Ethical Implications* (~21 pages)
- John Danaher, "The Philosophical Case for Robot Friendship", Journal of Posthuman Studies, 3(1): 5–24. doi:10.5325/jpoststud.3.1.0005 (~16 pages)
- Danaher, "Embracing the Robot," https://aeon.co/essays/programmed-to-love-is-a-human-robot-relationship-wrong (~12)

Optional Material

- *Her* by Spike Jonze

Meeting 10: The Moral Status of AI

Required Material

- Heller, Nathan. (2016) "If animals have rights, should robots?" The New Yorker https://archive.is/X3fJq (~9)
- S. Matthew Liao, "The Moral Status and Rights of Artificial Intelligence," in *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020) (~17 pages)
- Parisa Moosavi. (2024). Will intelligent machines become moral patients? *Philosophy and Phenomenological Research*, 109, 95–116. https://doi.org/10.1111/phpr.13019 (~19 pages)