AI, Society, and Humanity

Fall 2023 T/Th 12:30-1:45pm | **Room** Porter Hall 100 | **Course** 80249

Carnegie Mellon University

Instructor

Dr. Derek Leben (he/him)

I say it: "Lee-ben."
I can be addressed as "Professor Leben," "Dr. Leben," or "Derek"

Email

dleben@andrew.cmu.edu

Office Location & Hours

Tepper 4120; MW 12:30pm-2:30pm and by appointment.

My working hours are 9am-6pm on weekdays.

Course Description

AI and robotic technologies are developing rapidly and are increasingly incorporated into decisions, practices, and activities that impact individual and social interests. To ensure that these technologies advance meritorious goals without undermining important values or relationships, stakeholders must be able to understand the diverse ways in which new technologies can impact the lives of individuals and communities, the diverse dimensions on which such impacts can be evaluated and measured and where in the lifecycle of product development these various impacts might be anticipated and addressed. Through a series of case studies of current or near-future AI and robotics technologies students in this course will explore frameworks for assessing, evaluating and regulating novel technologies with the goal of ensuring that they support and advance human interests and social values.

Learning Objectives

The goals of this course are: (1) Understand and "consume" analyses of technology from multiple disciplinary perspectives (a knowledge component), and (2) explain many of the different ways that AI & robotic technologies can impact people (a skills component). Performance in the knowledge component will be assessed by quizzes, and performance in the skills component will be assessed by papers. Regular engagement will be assessed by mandatory attendance.

Textbook

There is no textbook, all readings are posted on Canvas under "modules" for each week.

Requirements

- Attendance (7%)
- Quizzes (8.25% x 4)
- Group Presentation (10%)
- Paper 1 (25%)
- Paper 2 (25%)

Detailed Requirements

Attendance

During each class, you will mark your attendance on a physical piece of paper with a physical writing device. Each student has exactly two "free" absences. This means that there should be no need to contact the instructor with excuses for absences, unless it is with a record of a *legitimate reason* (below) which requires more than two absences.

The slides are minimalist by design, and are intended to be used as visual "section headings" for the material in lecture and class discussion, which means you will need to take notes during lecture.

Audio recordings of lectures will be shared individually with students (via a Box link) whenever a student cannot attend class for a *legitimate reason*, which is shared with me before the time of the lecture.

Quizzes

There will be four in-class quizzes. You will be asked to label key terms and answer multiple-choice questions about definitions, positions, and arguments.

If you cannot attend class on a quiz day for a legitimate reason, you may take a make-up quiz during my office hours.

Group Presentations

By the end of week 1, all students must sign up for a presentation team on Canvas. The presentation list is located under the module for Week 1. The groups will all have a maximum of 3 students; if a group has 3 students, you must sign up for another group. If all groups are filled, please email me. Presentations must be recorded and uploaded to Canvas by 11:59pm on 11/14. There will be a discussion in-class on 11/21, where students will view at least three presentations and come prepared with objections.

Groups will present on a specific product/service which incorporates AI, and the ethical challenges presented by that product/service. They will select either the "support" or "oppose" side, where the oppose side will argue that this product/service violates an ethical principle, and the "support" side argues that it is ethically permissible.

Unlike the papers, I don't expect you to make use of course readings and materials to make your points (although this is welcome). The rubric for presentations is on Canvas.

Papers

There are two papers for the class, which are both argumentative papers. This means the goal is to propose a thesis, assume the reader disagrees with your thesis, and try to convince the reader. The first paper will be about the more theoretical material from the first half of the course, and the second paper will be about the more practical/ethical material from the second half of the course. The due dates are:

You are welcome to meet with me to go over drafts. I do not schedule meetings on evenings, weekends, or the paper due date. The rubric for the papers is available on Canvas.

Late grades will be penalized with a deduction of 3 points per day late.

Legitimate Reasons for Absence

Medical, legal, or other serious obligations which override your obligations as a student. Travel for personal or family plans is not a legitimate reason. As a CMU student, you have agreed to be present for all lectures in the regular term, including lectures before vacation periods, and you must make personal or family plans around your class schedule.

Accommodations for Students with Disabilities

If you have a disability and have an accommodations letter from the Disability Resources office, I encourage you to discuss your accommodations and needs with me as early in the semester as possible. I will work with you to ensure that accommodations are provided as appropriate. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Resources, I encourage you to contact them at access@andrew.cmu.edu.

Student Wellness Resources

As a student, you may experience a range of challenges that can interfere with learning, such as strained relationships, increased anxiety, substance use, feeling down, difficulty concentrating and/or lack of motivation. These mental health concerns or stressful events may diminish your academic performance and/or reduce your ability to participate in daily activities. CMU services are available, and treatment does work. You can learn more about confidential mental health services available on campus at: http://www.cmu.edu/counseling/. Support is always available (24/7) from Counseling and Psychological Services (CaPS): 412-268-2922. If you or someone you know is feeling suicidal or in danger of self-harm, call someone immediately, day or night:

- CaPS: 412-268-2922
- Re:solve Crisis Network: 888-796-8226

If the situation is life threatening, call the police

- On campus: CMU Police: 412-268-2323
- Off campus: 911

Diversity, Equality, Inclusion

The university encourages anyone who experiences or observes unfair or hostile treatment on the basis of identity to speak out for justice and support, within the moment of the incident or after the incident has passed. Anyone can share these experiences using the following resources:

Center for Student Diversity and Inclusion: csdi@andrew.cmu.edu, (412) 268-2150

Report-It (Links to an external site.) online anonymous reporting platform: net (Links to an external site.) username: tartans password: plaid

All reports will be documented and deliberated to determine if there should be any following actions. Regardless of incident type, the university will use all shared experiences to transform our campus climate to be more equitable and just.

Academic Integrity

You are responsible for reading and understanding the Academic Integrity Guidelines in the Student Handbook. The optional readings and SEP entries should contain all of the materials you will need, and you should also consult the CMU libraries as well as your instructor if any further materials are needed. Below are some specific interpretations of these guidelines which are relevant for this course.

Plagiarism and Improper Use

In this course, plagiarism is a deliberate attempt to submit another person's words or ideas as your own. As such, it constitutes deception. We do employ the "Turn it in" function on Canvas to screen every paper for plagiarism. If there is material from another source presented without proper acknowledgment, we will evaluate the case based on factors like:

Was the material cited, but just improperly acknowledged?

Was this material from an official text in the course, or from outside material?

How much material was copied/used?

Depending on the severity of the offense, the following penalties may be applicable, along with an official Academic Integrity Violation report:

Small grade penalty (e.g., half a letter grade), Moderate grade penalty (e.g., up to a letter grade), Regrade with elimination of the copied material, Grade penalty and regrade with elimination of the copied material, Grade of 0 on the paper, Grade of 'F' in the course

Cheating

In this course, cheating is the use of outside material (including the work of other students) on a quiz/exam. No exams in this course may make use of any material during the exam itself. Cheating on exams will depend on the severity of the case. Taking attendance for another student is also an instance of cheating.

The Use of Generative AI for Papers

The use of generative AI as a research tool for finding references and information is an improper use of materials. Models trained on data from the Internet are not the same as search engines, and will often "hallucinate," or fabricate, information that is presented as true.

The use of generative AI for producing *the content or words* of a paper, even a single sentence, is plagiarism. Turning in work which has been written by another agent, either human or machine, without acknowledging proper authorship, is always a violation of academic integrity.

It is permissible to use generative AI as inspiration, the way you might look at the papers of another student as inspiration. But you should treat any content generated by AI the same way you would treat the content generated by other students in the class. You may also engage with AI models in mock debates the way you would with other students, challenging it to present objections and criticisms of the arguments which you have written.

If the graders suspect that any portion of a paper has been generated by AI, and an AI detection tool suggests that this is also the case, there may be an interview scheduled with that student, where we will ask the student to review past writing and compare styles. We will use evidence from toolkits, our own judgments, and comparison with past writing styles to evaluate if plagiarism has occurred.

Course Schedule

READINGS HIGHLIGHTED IN BLUE ARE OPTIONAL

Week	Topic	Reading	Dates
Week 1	**Introduction**	Artificial Intelligence: A modern approach, ch.1 Stuart Russell and Peter Norvig	8/29, 8/31
		A New Kind of Science, ch.2-3 Stephan Wolfram SEP: "Artificial Intelligence", "Turing Machines," "The Computational Theory of Mind"	
Week 2	**Philosophy Lens: Can Machines Think?**	"Computing Machinery and Intelligence" (1950) Alan Turing "Is LaMDA Sentient?" (2022) Blake LaMoine and <collaborator> **Movie of the Week:** *Ex Machina*. (2014) Director: Alex Garland	9/5, 9/7
		"Minds, Brains, and Programs" (1980) John Searle SEP: "Alan Turing," "The Chinese Room Argument," "Functionalism",	
Week 3	**Philosophy Lens: Consciousness and Identity**	"Could a Large Language Model be Conscious?" (2022) David Chalmers "Could You Merge With AI?" (2020) Cody Turner and Susan Schneider	9/12, 9/14

Week	Topic	Reading	Dates
		Movie of the Week: *Black Mirror: Be Right Back*. S2e1 (2013) Director: Owen Harris	
		"Consciousness in Artificial Intelligence" (2023) Patrick Bultin et al. "Personal Identity" (1971) Derek Parfit "What is it like to be a bat?" (1974) Thomas Nagel "Integrated Information Theory: from consciousness to its physical substrate" (2016) Giulio Tononi et al SEP: "Zombies", "Consciousness", "Personal Identity"	
Week 4	**Psychology Lens: Trust and Relationships**	"Trust in Automation" (2014) Hoff and Bashir **Movie of the Week:** *Her*. (2013) Director: Spike Jonze	9/19, 9/21
		"Symposium" (c.385 BC) Plato "Tying the knot with a robot: legal and philosophical foundations for human-artificial intelligence matrimony" (2021) Greg Yanke	

Week	Topic	Reading	Dates
		"Sex Robots: Are we ready for them? An exploration of the psychological mechanisms underlying people's receptiveness of sex robots" (2022) Junzhao Ma, Dewi Tojib, and Yelena Tsarenko "From Sex Robots to Love Robots: Is mutual love with a robot possible?" (2017) Sven Nyholm and Lily Frank SEP: "Trust," "Game Theory," "Love"	
Week 5	**Economics Lens: Augmentation and Transformation of Work**	"Artificial intelligence, automation, and work" (2018) Daron Acemoglu & Pascual Restrepo **Movie of the Week:** Modern Times (1936) Director: Charlie Chaplin	9/26, 9/28
		"Why are there still so many jobs?" (2015) David Autor "Will life be worth living in a world without work? Technological unemployment and the meaning of life" (2017) John Danaher SEP: "Philosophical Approaches to Work and Labor", "The Meaning of Life"	
Week 6	**Policy and Governance Lens: Rights and Liabilities**	"Machines without principles: liability rules and AI" (2019) David Vladeck	10/3, 10/5

Week	Topic	Reading	Dates
		"Robots Should be Slaves" (2009) Joanna Bryson **Movie of the Week:** Metropolis (1927) Director: Fritz Lang	
		"Is it time for robot rights?" (2021) Vincent Muller "The other question: can and should robots have rights?" (2018) David Gunkel "When Something Goes Wrong: Who is responsible for errors in ML decision-making?" (2023) Andrea Berber and Sanja Sreckovic SEP: "Computing and Moral Responsibility", "Rights", "Personhood and Ethics"	
Week 7	Ethics Lens:	"Principled Artificial Intelligence" Fjeld, et al. (2020) **Movie of the Week:** Blade Runner (1982) Director: Ridley Scott	10/10, 10/12
		"Perspectives and Approaches in AI Ethics: East Asia" (2021) Danit Gal "Race and Gender" (2020) Timnit Gebru SEP: "Ethics of AI and Robotics"	

Week	Topic	Reading	Dates
Week 8	**FALL BREAK**		
Week 9	Predictive AI: Criminal Justice	"Do the Ends Justify the Means? Variation in the Distributive and Procedural Fairness of Machine Learning Algorithms" (2022) Lily Morse, et al "Heat Listed" (2018) Matt Stroud "Machine Bias" (2016) Julia Angwin et al. **Movie of the Week:** Minority Report (2002) Director: Steven Spielberg	10/24, 10/26
		"Discrimination in the Age of Algorithms" (2019) Jon Kleinberg, et al. "Using AI to Prevent Crime: Implications for due process and criminal justice" (2022) Kelly Blount "A Review of Predictive Policing from the Perspective of Fairness" Kiana Alikhademi et al. (2022)	
Week 10	Embodied AI and Robots: Military Applications	"Autonomous Weapons Systems and the Ethics of AI" (2020) Peter Asaro **Movies of the Week:** The Terminator (1984) Director: James Cameron Robocop (1987) Director: Paul Veerhoven	10/31, 11/2

Week	Topic	Reading	Dates
		"Lethal Autonomous Weapon Systems and Respect for Human Dignity" (2022) Leonard Kahn "The Humanitarian Imperative for Minimally-Just AI in Weapons" (2021) Jason Scholz and Jai Galliott	
Week 11	Embodied AI and Robots: Autonomous Vehicles	"The Robot Car of Tomorrow May Just Be Programmed to Hit You" (2017) Patrick Lin "Autonomous Driving Ethics" (2021) Maximillian Geisslinger, et al. **Movie of the Week:** Wall-E (2008) Director: Andrew Stanton	11/9
		"Measuring Automated Vehicle Safety" (2018) Blanar et al. "Implementable Ethics for Autonomous Vehicles" (2016) Chris Gerdes and Sarah Thornton "The Ethical Knob" (2017) Giuseppe Contissa et al.	
Week 12	GenAI: Large Language Models	"Bing's AI Chat: I Want to Be Alive" (2022) Kevin Roose "Ethical and Social Risk of Harm from Large Language Models" (2021) Laura Weidinger et al.	11/14, 11/16

Week	Topic	Reading	Dates
		Movie of the Week: Upgrade (2018)	
		"Constitutional AI: Harmlessness from AI Feedback" (2023) Yuntao Bai et al. "Sparks of AGI" (2023) Sebastien Bubeck et al.	
Week 13	Student Presentation discussions	View at least 3 presentations	11/21
Week 14	GenAI: Image and Music Generators	"AI is Blurring the Definition of an Artist" (2019) Ahmed Elgammal "AI Agents are Not Artists" (2022) S. Will Chambers	11/28, 11/30
		"We're Witnessing the Birth of a New Artistic Medium" (2022) Stephan Marche SEP: "The Definition of Art", "Philosophy of Digital Art"	
Week 15	The Singularity and X-Risk	"The Singularity" (2020) Ray Kurtzweil "Ghost in the Cloud" (2017) Meghan O'Gieblyn **Movie of the Week:** *2001: A Space Odyssey*. (1968) Director: Stanley Kubrik	12/5, 12/7

Week	Topic	Reading	Dates
		Pensees [excerpt] (1669) Blaise Pascal “The Superintelligent Will-motivation and instrumental rationality in advanced artificial agents” (2005) Nick Bostrom “Silicon Valley’s Quest to Build God and Control Humanity” (2023) Edward Ongweso, Jr.	