Ethics and Artificial Intelligence

Carnegie Mellon University

Course 46898
M3; Spring 2024

T/Th 4pm - 5:45pm
Tepper 2111

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 10am-12pm and by appointment.

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

Description

This course will explore the ethical challenges that businesses face when making use of AI, map out policies which have been proposed as solutions to these challenges, and analyze the normative arguments behind these policies. The goal of the course is to acquire knowledge of the ethical challenges which emerge from AI, and the skills to develop responsible corporate practices around AI. The course is organized around six core principles for the responsible use of AI, and applications to illustrate each principle (separated into domains of products):

Topic	Product Domains
Autonomy	Media and Marketing
Explainability	Credit and Finance
Discrimination	Hiring
Fairness	Criminal Justice
Benefit	Transportation, Security
Responsibility	Healthcare

This course will NOT address the long-term risks/impacts of AI, such as: machine personhood and rights, human unemployment, the singularity, existential risk, etc.

Goals

The goals of the course are: (1) to gain an understanding of the ethical challenges raised by new technologies in business, and (2) to develop skills for criticizing and defending corporate policy about the use of these technologies. We are defining a "policy" as a statement about the conditions under which your company believes it is permissible to deploy AI in some domain for some purpose/task (and the conditions under which it is *not* permissible), in as much detail as possible.

Requirements

- Attendance (5%)
- Quizzes (15%)
- Presentation (20%)
- Paper 1 (30%)
- Paper 2 (30%)

Scoring

The scoring system in this class is:

89.5-100% = A

79.5-89.49% = B

…

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* (see below) which requires more than two absences.

Presentation

Every Tuesday in class (starting Week 2), teams will present a 10-15 minute evaluation of a case-study. The teams will research information about the company and its practices and products, and describe: (1) how does this company use AI in its practices or products, (2) how does this involve the ethical topic from THAT WEEK, (3) what are some policies that the company might consider to mitigate this specific ethical risk? Please focus ONLY on the particular ethical principle being discussed that week.

> *Policies should have the general format*: "we believe that it is a violation of (principle) to deploy AI for (domain, task) under the following conditions… and according to this policy, (case study) is/is not a violation of (principle)."

Teams should use the lectures and readings from that section to support their case. The rubric for the presentation is available on Canvas.

In addition, I ask that teams post a 1-page summary of their policy and argument on the Canvas discussion board by Monday evening before their presentation, and that each student come prepared with one objection to raise. Part of the presentation grade is how these objections are addressed in real time.

Quizzes

There will be four quizzes. You will be asked to define key terms and answer multiple-choice questions about positions and arguments.

Papers

The papers are arguments for a specific policy about ONE case selected from the assignment description on Canvas, using ONE normative theory and focusing on ONE principle. Any standard formatting and citation styles are acceptable.

I expect that the papers will make use of the readings and lectures to argue for a detailed policy about how you believe the company should address the case-study.

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

Evaluation

Late Penalties and Make-Ups

Papers will be penalized with a deduction of 3 percentage points per day late. There is a maximum lateness deduction of 50 percentage points. I will accept work turned in by the end of the semester for the maximum lateness penalty.

Missed presentations and quizzes cannot be made up. If you are unable to present or take the quiz on the required day, please see me ahead of time and we can almost certainly find a solution.

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, nor is adding classes for next semester. Job or medical school interviews are legitimate. As a CMU student, you have agreed to be present for all classes in the regular term, including those before vacation periods, and you must make personal or family plans around your class schedule.

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 'R' in the course

Cheating

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

The Use of Generative AI

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.

Accommodations

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

Diversity, Equity, 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.

Course Schedule

Articles highlighted in pink are optional

Week	Topic	Readings	Class Dates
Week 1	AI and Ethics: Definitions and Frameworks	"Principled Artificial Intelligence" Fjeld, et al. (2020) *Berkman Klein Center*	1/16
		Optional "The Global Landscape of AI Ethics Guidelines" Jobin et al. (2020) *Nature Machine Intelligence* "Constitutional AI: Harmlessness from AI Feedback" Bai et al. (2022) "Principles Alone Cannot Guarantee Ethical AI" Mittelstadt (2019) *Nature Machine Intelligence* Blueprint for an AI Bill of Rights (2022) *The White House* The AI Act (2023, revised) *European Union Parliament*	

Week	Topic	Readings	Class Dates
Week 2	Autonomy	The GDPR in the Age of Surveillance Capitalism” Andrew and Baker (2019) *Journal of Business Ethics* Generative AI Has an Intellectual Property Problem (2023) Appel, Neelbauer, and Schweidel *Harvard Business Review* **Optional** “A Right to Reasonable Inferences” Wachter and Mittelstadt (2019) *Columbia Business Law Review* Foundation Models and Fair Use (2023) Henderson, Li, Jurafsky, Hashimoto, Lemley, and Liang *Stanford Law and Economics Online Working Paper* “The Ethical Application of Biometric Facial Recognition” Smith and Miller (2022) *AI and Society*	1/18, 1/23

Week	Topic	Readings	Class Dates
Week 3	Explainability	"Is Explainable AI Intrinsically Valuable?" Colaner (2022) *AI and Society* "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Rudin (2019) *Nature Machine Intelligence* **Optional** "Beyond Explainability: Justifiability and Contestability of Algorithmic Decision Systems" Henin and Lemetayer "Why a Right to Explanation of Automated Decision Making Does Not Exist in the GDPR" Wachter et al. (2017) *International Data Privacy Law* "Explainable AI as Evidence of Fair Decisions" Leben (2023) *Frontiers* "Explainable Artificial Intelligence (XAI)" Arrieta et al. (2020) [Technical Survey] *Information Fusion* "A Survey of Methods for Explaining Black Box Models" [Technical Survey] Guidotti et al. (2018) *ACM Computing Surveys*	1/25, 1/30

Week	Topic	Readings	Class Dates
Week 4	Discrimination	"Discrimination in the Age of Algorithms" Kleinberg, et al. (2018) *Journal of Legal Analysis*	2/1, 2/6
		Optional "Big Data's Disparate Impact" Barocas and Selbst (2016) *California Law Review* "Choosing How to Discriminate: navigating ethical trade-offs in fair algorithmic design for the insurance sector" Loi and Christen (2021). *Philosophy and Technology* "Challenging Biased Hiring Algorithms" Kelly-Lyth (2021) *Oxford Journal of Legal Studies* Chapter 4: Equal Treatment and Discrimination Chapter 5: Relevance *Fairness for AI* (2024) Leben	

Week	Topic	Readings	Class Dates
Week 5	Fairness	Chapter 1: Measuring Fairness *Fairness for AI* (2024) Leben	2/8, 2/13
		Optional *Fairness in Machine Learning* (ch.3: "Classification") Barocas, Hardt, and Narayanan (2021) "Do the Ends Justify the Means? Variation in the Distributive and Procedural Fairness of Machine Learning Algorithms" Morse et al. (2021) Journal of Business Ethics Fairness in ML: lessons from political philosophy Binns (2019). *Journal of Machine Learning Research* "Beyond Bias: Re-Imagining the Terms of 'Ethical AI' in Criminal Law" Barabas (2020) Georgetown Journal of Law Chapter 3: Demo- AI for Mortgages *Fairness for AI* (2024) Leben	

Week	Topic	Readings	Class Dates
Week 6	Benefit	"Ethical and Social Risk of Harm from Large Language Models" (2021) Laura Weidinger et al. **Popular Article:** "The Robot Car of Tomorrow May Be Programmed to Hit You" Lin (2017) *Machine Ethics and Robot Ethics* **Optional** "AI and Machine Learning in Financial Services" Financial Stability Board "Measuring Automated Vehicle Safety" Blanar et al. (2018) RAND Corporation "Autonomous Driving Ethics" Geisslinger et al. (2021). Philosophy and Technology Chapter 6: Avoiding Collisions *Ethics for Robots* (2018) Leben Chapter 8: Fairness vs. Accuracy *Fairness for AI* (2024) Leben	2/15, 2/20

Week	Topic	Readings	Class Dates
Week 7	Responsibility & Control	"Are Current Tort Liability Doctrines Adequate for Addressing Injury Caused by AI?" Sullivan and Schweikart (2019) *AMA Journal of Ethics* **Popular Article:** "When Artificial Intelligence Botches Your Medical Diagnosis, Who's to Blame?" Hart (2017) *Quartz* **Optional** "Ethical Implications and Accountability of Algorithms" Martin (2018) *Journal of Business Ethics* "On the Ethics of Algorithmic Decision-Making in Healthcare" Grote and Berens (2020) *The Journal of Medical Ethics* "Machines Without Principles: liability rules and AI." Vladeck (2019). *Washington Law Review* When AI is Marketed As a Doctor, Companies Take on New Responsibilities" Leben and Schweikart (in prep)	2/22, 2/27