Philosophy 171 - Topics in Artificial Intelligence
Prof. Hayley Clatterbuck
clatterbuck@ucla.edu

Meeting Times

Wednesday, Friday	9:30-10:45	Bunche 3153

Office Hours

Friday	11:00-1:00	Dodd 379

If you cannot meet at this time, e-mail me to set up an appointment.

Readings

All assigned readings will be posted on BruinLearn. If you need assistance accessing course materials, please let me know.

Grade Composition

Case analyses (15% each)

> You will complete three case analyses. You will be presented with a real world case and will take some time in-class to brainstorm some of the ethical issues that it raises. Then, after we discuss the topic, you will write a follow-up.

Essay (15%)

> There will be a final essay due toward the end of the quarter in which you will analyze a topic in AI ethics. More instructions will be distributed as the quarter progresses.

Final exam (25%)

> There will be a cumulative final exam, with a mixture of short-answer and longer essay questions.

Participation (15%)

> You are required to attend class meetings. You will be allowed two unexcused absences during the quarter, after which each absence will lower your participation grade. Of course, if you must miss class due to illness, religious observation, or some other valid reason, please e-mail prior to your absence so that I can excuse it.

Classroom Technology Policy

Laptops can be distracting, both for you and your classmates. There is plenty of evidence showing that students learn better when they take notes by hand. If you must use a laptop for note-taking, you must sit in one of the back two rows of the classroom. I will make my lecture notes available for review if needed.

Late Policy

Late assignments will receive a 10% penalty per day, where a day is measured from the class's start time. For example, if a paper is due Monday at the beginning of class and you were to turn your paper in on Monday night, you would incur a 10% penalty. If you were to turn it in on Tuesday after 1:00, it would be a 20% penalty.

AI Policy

AI use is not permissible during lectures. You should not use AI at all for your case studies. For the paper, you may use AI to: find sources (which you must independently check); format citations; run experiments on AIs (e.g. evaluating which LLM is most sycophantic) if doing so is a part of your paper project. If AI use is suspected, be prepared to orally explain and defend your work.

Special Accommodations

Our goal is to provide an environment that will allow every student to succeed in this course. Please contact me to discuss any special accommodations that would help you to do so.

Students needing academic accommodation based on a disability should either contact me or the Office for the Center for Accessible Education located at (310) 825-1501 or A255 Murphy Hall. When possible, students should contact the CSA within the first two weeks of the term as reasonable notice is needed to coordinate accommodations. For more information visit www.cae.ucla.edu

Academic Misconduct and Plagiarism

Be familiar with and abide by UCLA's policy on academic and intellectual integrity and their in-person COVID-19 protocols:
http://www.studentgroups.ucla.edu/dos/students/integrity/
http://www.deanofstudents.ucla.edu/Code_choice.php

Additions to this syllabus may be made at my discretion as the quarter progresses. I will notify you of all changes and will update the syllabus on BruinLearn as needed.

Schedule of Readings
(* denotes optional readings, all other readings required)

1. Alignment to what?

Fri., 9/26:	Intro to 171
Wed., 10/1:	*No class - Clatterbuck away*
Fri., 10/3:	Gabriel and Ghazavi, "The challenge of value alignment"
Wed., 10/8:	no reading

2. Algorithmic bias

Fri., 10/10:	*Case Analysis #1: Predicting student success*
Wed., 10/15:	Fazelpour and Danks, "Algorithmic bias" At least one of: Lum & Isaac, "To predict and serve" ProPublica, "Machine bias in sentencing algorithms" NYTimes: "Wrongfully accused by an algorithm" McCormick, "What happened when AI made healthcare decisions"
Fri., 10/17:	no reading

3. LLMs: training and alignment

Wed., 10/22:	*Case Analysis #1 DUE* *Case Analysis #2: Political bias of LLMs* At least one of: Wildeford, "Can we safely deploy AGI if we can't stop MechaHitler?" CNN, "Is AI woke?" Hartmann, *et al.* "The political ideology of conversational AI"*
Fri., 10/24:	Shiller, "LLMs are weirder than you think" 3Blue1Brown: LLMs explained briefly
Wed., 10/29:	Emergent misalignment - at least one of: Ornes, "AI was fed sloppy code. It became evil" Betley, *et al.* "Emergent misalignment" OpenAI, "Understanding, preventing misalignment generalization"
Fri., 10/31:	Sycophancy - at least one of one of: Cotra, "Why AI alignment could be hard with deep learning" Goedecke "Sycophancy is the first LLM dark pattern" NYTimes, "Chatbot delusional spirals"

4. Autonomous systems and responsibility gaps

Wed., 11/5:	*Case Analysis #2 DUE* *Case Analysis #3: Rogue action by an autonomous robot*
Fri., 11/7:	Nyholm, "Ethics of crashes with self-driving cars" NBC, "Tesla found partially liable in fatal autopilot crash"
Wed., 11/12:	no reading
Fri, 11/14:	Sparrow, "Killer robots"

Wed., 11/19:	no reading

5. GANs: deepfakes

Fri., 11/21:	*Case Analysis #3 DUE* Tufecki, "AOC deepfake was terrible, proposed solution is delusional"
Wed., 11/26:	Rini, "Deepfakes and the epistemic backstop"
Fri., 11/28:	*No class - Thanksgiving*
Wed., 12/3:	no reading
Fri., 12/5:	*Final essays DUE*

Final exam:

Thursday, December 11
11:30 AM - 2:30 PM