Introduction to AI Ethics - An Interdisciplinary Approach

COMSCI 88SB | Spring 2020

Instructor

Aaron Hui | aaronhui98@ucla.edu

Office Hours

Bomb Shelter | 5:00 pm - 6:00 pm | Mondays

Course Description

As technology advances at an exponential rate, it is imperative that students begin thinking about how Artificial Intelligence (AI) will interact with and impact society at different facets of daily life, with the focus on ethical implication and its implementation within AI systems. The foresight of how important it is to regulate AI through the lens of ethical discussion and implementation is of paramount significance within the setting of the rapid development of AI and how it has already permeated our daily lives.

This course will allow students to gain an interdisciplinary introduction to classical and modern ethical theory and their implications on emerging autonomous technologies. We will cover a variety of issues, but they will all address one of the six ethical principles of AI identified by Microsoft:

1. Fairness: AI systems should treat all people fairly.
2. Inclusiveness: AI systems should empower everyone and engage people.
3. Reliability and Safety: AI systems should perform reliably and safely.
4. Transparency: AI systems should be understandable by people.
5. Privacy and Security: AI systems should be secure and respect privacy.
6. Accountability: AI systems should have algorithmic accountability.

This course will include introductions to the AI Robotics Ethics Society (AIRES), the UCLA Law AI Pulse program, the AI Ethics Lab, the USC Center for Artificial Intelligence in Society (CAIS), and more.

Learning Objectives

By the end of this course, you should be able to:

1. Understand the current issues in the field of AI Ethics today.
2. Summarize arguments of modern and classical philosophy as it applies to AI ethics.
3. Construct unique rational arguments for and against ethical perspectives as they pertain to AI technology.

Course Materials

There is no textbook for this course as such a textbook does not exist yet. All materials will be provided to you on CCLE.

Grading Policy

This course is a 1-unit P/NP seminar. As with all UCLA courses, a minimum of 70% is required in order to receive a passing grade. Grading is based upon three components:

1. Class participation: 33.3%
2. Weekly forum post and response: 33.3%
3. Final Paper: 33.3%

Attendance/Participation Policy

Participation will be based on active contribution to in-class discussions or activities. Only one excused absence will be allowed, along with a make up assignment on the week's discussion. Per UCLA and USIE policy, more than one absence from the course will result in a failing grade.

Forum Posts

Readings are to be completed before class. All readings will be provided on CCLE. You must complete at least 6 of the 9 posts in order to receive this portion of the grade.

Post 1: Before the beginning of each class beginning week 1, pose a question about the week's assigned reading assignments on CCLE. This question should be open-ended and not able to be answered with a simple "yes" or "no."

Post 2: After each class and before the start of the next class, answer a classmate's question on CCLE!

Final Paper

This course will include one final paper that will be due on Week 10. The paper will be on a potential solution to one of the issues in AI ethics presented throughout the course, and should be approximately three to four pages in length. **Completion of the final paper is a necessary requirement to pass the course.** You will have the option to do it with a partner or solo if you are feeling ambitious.

Academic Integrity

Academic misconduct in any form (including plagiarism, note-selling, multiple submissions, and cheating) will be dealt with according to UCLA's policy and procedures regarding academic honesty. This includes reporting suspected violations to the Dean of Students. If you are uncertain as to what constitutes plagiarism, the library has a helpful guide: (http://guides.library.ucla.edu/citing/plagiarism/avoid).

When you submit an assignment with your name on it, you are signifying that the work contained therein is yours, unless otherwise cited or referenced. Any ideas or materials taken from another source for either written or oral use must be fully acknowledged. Penalties for academic misconduct may include a failing grade on the assignment and/or a failing grade in the course, among other possibilities. If you are unsure about the expectations for completing an assignment or taking a test exam, be sure to seek clarification beforehand.

DISCLAIMER

This syllabus is intended to give students guidance in what may be covered this quarter. The instructor reserves the right to make modifications to this information as the course progresses.

CAMPUS RESOURCES FOR STUDENTS

Academic Achievement Program (AAP):

AAP advocates and facilitates the access, academic success, and graduation of students who have been historically underrepresented in higher education; informs and prepares students for graduate and professional schools; and develops the academic, scientific, political, economic, and community leadership necessary to transform society. Learn more at http:// www.aap.ucla.edu/

Undergraduate Writing Center:

The Undergraduate Writing Center offers UCLA undergraduates one-on-one sessions on their writing. The Center is staffed by peer learning facilitators (PLFs), undergraduates who are trained to help at any stage in the writing process and with writing assignments from across the curriculum. PLFs tailor appointments to the concerns of each writer. Multiple locations and hours available. For more information or to schedule an appointment, visit http://wp.ucla.edu/wc/

Center for Accessible Education (CAE):

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

Counseling and Psychological Services (CAPS):

CAPS supports student mental health needs as they pursue their academic goals. Their services are designed to foster the development of healthy well-being necessary for success in a complex global environment. CAPS offers a variety of services to meet student needs including: crisis counseling available by phone 24 hours a day/7 days a week 310-825-0768, emergency Intervention, individual counseling and psychotherapy,

group therapy, psychiatric evaluation and treatment, psychoeducational programs and workshops, and campus mental health and wellness promotion. Please visit http://cousneling.ucla.edu for more information.

Full Course Reading Schedule

Week 1: Introduction to Ethics in AI and Machine Learning

- This unit will cover the structure of the seminar, expectations, and goes over the syllabus. We will Introduce the main topic of the course and review the current debates in AI ethics and the fundamentals of machine learning. Discussion topics will focus on why AI ethics is necessary in developing AI, and whether ethics can be hard coded into AI systems.

- Optional Readings:
    1. "The Ethics of Artificial Intelligence: Mapping the Debate" by Brent Daniel Mittelstadt et al. (Big Data & Society, 2016).
    2. Toward an ethics of algorithms: Convening, observation, probability, and timeliness by Ananny, M. (Science, Technology, and Human Values, 2016).
    3. "Why Zuckerberg and Musk are Fighting About the Robot Future" by Ian Bogost (The Atlantic, July 27, 2017).
    4. "Ethics needs to keep up with economics when it comes to AI, experts warn" by James McLeod (Financial Post, April 12, 2019).
    5. "The Problem With AI Ethics" by James Vincent (The Verge, April 3, 2019).
    6. "Thinking Machines: The Search for Artificial Intelligence" by Jacob Roberts (Chemical Heritage Foundation, 2016)
    7. "A Few Useful Things to Know about Machine Learning" by Pedro Domingos (University of Washington)

Week 2: How to Analyze and Create Arguments

- This unit will cover the basics of common argument forms involving inductive and deductive arguments. Discussion topics will include debates in various topics raised by students but utilizing proper argument forms instead of resorting to violence and rhetoric or emotional appeals.

- Readings:

    1. "Four Argument Strategies" by JW Gray (10 min)
    2. "Argue with strangers online? A philosophy professor offers some tips for arguing well" by ABC Radio (ABC Radio, February 18, 2018). (5 min)

- Watch:

    1. Deductive and Inductive Arguments:
        a. https://www.youtube.com/watch?v=BwtCScUoL_w (12 min)
        b. https://www.youtube.com/watch?v=3PLUgFYhGvM (5 min)
    2. How to Argue - Philosophical Reasoning:
        a. https://www.youtube.com/watch?v=NKEhdsnKKHs (10 min)

- Optional Readings:
    1. "Five Steps To Arguing On The Internet, According To Philosophy" by Mitch Alexander (JUNKEE, November 14, 2016)

Week 3: The Different Ethics

- This unit will cover the basics of Utilitarianism, Deontological, and Virtue Ethics and how they pertain to AI ethics. Discussion topics will compare the three ethical standpoints and see how each one can contribute to AI ethics.

- Readings:

    1. "The Virtuous Machine - Old Ethics for New Technology?" by Nicolas Berberich et al. (Univ. of Munich, 2018) (20 min; Read pages 10-23)
    2. "Deontological Ethics" by the Stanford Encyclopedia of Philosophy (10 min; skim through but focus on intro)
    3. "Act and Rule Utilitarianism" by Stephen Nathanson (IEP, 2014) (10 min; skim through but focus on intro)

- Optional Readings:

    1. "Briefing: Kant's Ethics" (Moral Robots, January 6, 2017).
    2. "Should your autonomous car protect you? And at what cost? (AI + Kantian Ethics)" by Truman Halladay (Medium, January 24, 2018).
    3. "Virtue Ethics" by McCombs School of Business.

Week 4: DEBATE WEEK

- This until will be spent on reviewing all the covered ethical standpoints and the class will be split into groups advocating for different ethical perspectives. They will debate on which one of the ethical standpoints they find best suited for implementing into AI systems. Students will be randomly assigned to one of the three ethical perspectives.

- Readings: NONE

- Optional Readings: NONE

Week 5: Issues in Fairness

- This unit will cover the issues of fairness in AI in terms of bias towards certain groups of people. Discussion topics will include a case study on the problems of racial bias for algorithms designed to determine prison sentences for criminals.

- Readings:

    1. “This is how AI bias really happens—and why it’s so hard to fix” by Karen Hao (MIT Technology Review, February 4, 2019). (10 min)
    2. “Artificial Intelligence Has A Problem With Bias, Here's How To Tackle It” by Bernard Marr (Forbes, January 29, 2019). (10 min)
    3. "A computer program used for bail and sentencing decisions was labeled biased against blacks. It’s actually not that clear" by Avi Feller, Emma Pierson, Sam Corbett-Davies and Sharad Goel (Washington Post, 2016) (10 min)

- Optional Watch:

    1. The Trouble with Bias - NIPS 2017 Keynote - Kate Crawford #NIPS2017 https://www.youtube.com/watch?v=fMym_BKWQzk (50 min)

- Optional readings:
    1. “Should Prison Sentences Be Based On Crimes That Haven’t Been Committed Yet?” by Anna Maria Barry-Jester, Ben Casselman and Dana Goldstein (FiveThirtyEight, 2015)
    2. “A.I. Bias Isn’t the Problem. Our Society Is” by By Alex Salkever and Vivek Salkever (Fortune, April 14th, 2019)

3. “Technology is Biased too. How do we Fix it?” By Laura Hudson (Five Thirty Eight, July 20, 2017).
4. “Machines Taught by Photos to Learn a Sexist View of Women” by Tom Simonite (Wired Magazine, August 21, 2017)
5. “Why Stanford Researchers Tried to Create a ‘Gaydar’ Machine” by Heather Murphy (New York Times, 2017)

Week 6: Issues in Inclusiveness

- This unit will cover the issue of how humans remain useful in a world of autonomous machines. Discussion topics will include a case study of the relationship between the American trucker and AI.

- Readings:

1. “Will AI And Robots Force You Into Retirement?” by Stephen Chen (Forbes, April 4, 2019) (5 min)
2. “Artificial Intelligence Paradox: As Robots Take Over, People Skills Become More Critical” by Joe McKendrick (Forbes, November 29th, 2018). (5 min)
3. “End of the Road: Will automation put an end to the American trucker?” by Dominic Rushe (The Guardian, 2017) (10 min)

- Optional readings:

1. “How Frightened Should We Be of A.I.?” by Tad Friend (The New Yorker, May 7, 2018).
2. Cathy O’Neil, Weapons Of Math Destruction : How Big Data Increases Inequality And Threatens Democracy. New York: Crown; 2016.
3. “Will AI And Robots Force You Into Retirement?” by Stephen Chen (Forbes, April 4, 2019).
4. “The Relentless Pace of Automation” by David Rotman (MIT Technology Review, 2017)
5. “Where machines could replace humans, and where they can’t (yet)” by Michael Chui, James Manyika, and Mehdi Miremadi (McKinsey Quarterly, 2016)
6. "Everything You Need To Know About Sophia, The World's First Robot Citizen" by Zara Stone (Forbes, 2017).

Week 7: Issues in Transparency

- This unit will cover the issue of the problem with understanding how AI fundamentally makes their decisions. Discussion topics will include a debate on a case study in the Foreseeability Problem of AI regulation.

- Readings:

    1. "The Foreseeability of Human–Artificial Intelligence Interactions" by Weston Kowert (Texas Law Review, 2017). (20 min)
    2. "The Artificial Intelligence Black Box and the Failure of Intent and Causation" by Yavar Bathaee (Harvard Journal of Law and Technology, Spring 2018). Read only pp. 906 to pp. 918. (30 min)
    3. "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability" by Mike Ananny and Kate Crawford (New Media and Society, 2016). Read only intro and conclusion. (10 min)

- Optional readings:

    1. "Our Machines Now Have Knowledge We'll Never Understand" by David Weinberger (Backchannel, 2017).
    2. "Algorithmic Transparency for the Smart City" by Robert Brauneis and Ellen P. Goodman (Yale Journal of Law and Technology, 2017)
    3. "Is Artificial Intelligence Permanently Inscrutable?" by Aaron M. Bornstein (Nautilus, 2016).
    4. "Is Effective Regulation of AI Possible? Eight Potential Regulatory Problems" by John Danaher (Philosophical Disquistions, 2015).
    5. "What happens when the robots get it wrong?" by LexisPSL (LexisNexis, March 3, 2017).

Week 8: Issues in Accountability

- This unit covers how it could be possible to make AI systems accountable. Discussion topics will include a debate on why it's necessary to ensure that AI systems are designed ethically and responsibly, and a case study on the self driving car.

- Readings:

    1. "Whose Life Should Your Car Save?" by Azim Shariff et al. (New York Times, 2016). (5 min)
    2. "You Can't Sue a Robot: Are Existing Tort Theories Ready for Artificial Intelligence?" by Matthew O. Wagner (FrostBrownTodd, LLC, February 7, 2018). (5 min)
    3. "When your self-driving car crashes, you could still be the one who gets sued" by Madeleine Claire Elish and Tim Hwang (Quartz, 2015). (10 min)

- Optional readings:

    1. "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms" by Nicholas Diakopoulos et al. (Fairness, Accountability, and Transparency in Machine Learning)
    2. "Regulating the Loop: Ironies of Automation Law" by Meg Leta Ambrose (We Robot, 2014)
    3. "Machines without Principals (sic): Liability Rules and Artificial Intelligence" by David C. Vladeck (Washington Law Review, 2014).
    4. "Machine Learning and the Law: Five Theses" by Thomas Burri (Machine Learning and the Law Conference, 2017).
    5. "Computer says no: why making AIs fair, accountable and transparent is crucial" by Ian Sample (Guardian, 2017).
    6. "Executive Summary: The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems" (IEEE).

Week 9: Issues in Reliability and Safety

- This unit will cover how we can create AI in a safer and more reliable manner. Discussion topics will include a special emphasis on the usage of AI in medicine and how diagnostic accuracy and reliability is a primary concern, as well as a secondary case study on the self driving car.
- Readings:

    1. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction" by M.C. Elish (We Robot, 2016). (30 min)
    2. A.I. Versus M.D. By Siddhartha Mukherjee (The New Yorker, 2017). (15 min)

3. "Trusting self-driving cars is going to be a big step for people" by Jonathan O'Callaghan (European Commission, April 2, 2019). (5 min)

- Optional readings:

1. "Machines Treating Patients? It's Already Happening" by Alice Park (TIME, March 21, 2019).
2. "Artificial intelligence in medicine: current trends and future possibilities" by V. Buch, et al. (British Journal of General Practice, 2018).

Week 10: Issues in Privacy and Security

- This unit will cover how the dangers of allowing AI to invade the privacy of individuals. Discussion topics will include a special emphasis and discussion on China's upcoming AI controlled Social Credit System, and a case study on Facebook Privacy

- Readings:

1. "The Threat of Algocracy: Reality, Resistance and Accommodation" by John Danaher (Philosophy and Technology, 2016) (Preprint 1: Preprint 2) Read "What Is the Threat of Algocracy?" section. (10 min)
2. "Principles of AI Governance and Ethics Should Apply to All Technologies" by Herb Lin (Lawfare, April 12, 2019). (5 min)
3. "The complicated truth about China's social credit system" by Nicole Kobie (Wired, January 21, 2019). (10 min)

- Optional readings:

1. "How Facebook's Algorithm Suppresses Content Diversity (Modestly) and How the Newsfeed Rules Your Clicks" by Zeynep Tufekci (Medium, 2015)
2. "Facebook Figured Out My Family Secrets, And Won't Tell Me How" by Kashmir Hill (Gizmodo, 2017).
3. "Artificial Intelligence Pushes the Anti-Trust Envelope" by Michaela Ross (Bloomberg BNA, 2017)
4. "Rethinking Privacy For The AI Era" by Forbes Insight Team (Forbes, March 27, 2019).