Fall 2023
HSS 404 - AI Ethics

Instructor: Dr. Daniel Estrada
E-mail: estrada@njit.edu
WebEx: njit.webex.com/meet/estrada
Office: Cullimore 419
Office Hours: T 3pm and by appt.
Discord: discord.gg/NxFvdH7

Class Meeting: MW 1:00pm - 2:20pm CKB 341

Course description: This course addresses contemporary issues, debates, and controversies in AI Ethics, with the ultimate goal of auditing popular AI software currently in use. The course begins with a historical introduction to foundational concepts in computer science and machine learning. This unit is designed for students with no prior experience in computer science, with the goal of developing some core intuitions on the development, use, and limitations of machine learning techniques. The next unit reviews recent literature in AI Ethics to introduce foundational concepts and issues. This unit will also introduce and motivate the idea of an AI audit through several case studies and prepared examples. In the final unit, students will conduct an informal, external audit and ethical review of some specific AI application, presenting research on the operation and social impact of the technology. The class will conclude with a class activity focused on developing guidelines, principles, and policy recommendations that encourage the safe and ethical use of AI technologies.

Prerequisites: HUM 102 and one from among Hum 211, Hum 212, Hist 213 or Hist 214 or their equivalents, all with a grade of C or better; completion of either the Lit/Hist/Phil/STS or the Open Elective in Humanities and Social Science, with a grade of C or better.

Course objectives:

- Introduce students to the history and current practices in AI and machine learning in order to develop some intuition for how ML models are implemented, trained, and deployed in a variety of real world applications
- Review important concepts, methods, debates, and controversies in AI Ethics by directly engaging with recent scholarship in the field
- Gain insight into the development and use of particular AI applications through an informal, external audit and ethical review of the technology and its social impact

Course overview
The course consists of the following assignments and expectations, which are explained in more detail below.
The course consists of the following assignments and expectations, which are explained in more detail in the syllabus.

- Regular classroom attendance is expected. Students can miss 3 classes without penalty. (10%)
- Participation in classroom discussions and in the audit project earns 10%
- 300+ words of “reading notes” due for each class reflecting on reading assignments in Lesson 1-9 (20%)
- One presentation on weekly readings in Units 1 or 2. Schedule your presentation on Canvas. (20%)
- Two papers (3-5 pages) are required for Unit 1 & 2. (15%)
- Audit project is worth 20% of your grade.
- Final reflection essay is 5%

Your grades depend on the following assignments:

Attendance: Regular class attendance is required, and earns up to 100 points of credit for the semester. Students can miss up to three class sessions before it impacts your grade. There are 28 total days of class, so 25 attendance days earn full credit. On time attendance counts for one day. Attendance is considered late if registered more than 5 minutes after class begins and earns 80% credit. Attendance is taken on the class Discord server. Please do not register attendance on Discord until you are actually in your seat in class. Students registering attendance without being in class will lose all attendance credit for the semester. Note that attendance for some classes also earns participation credit! See more info below.

Reading Notes: Students are expected to complete 300+ words of reading notes for each class during Units 1 and 2. Notes should be posted directly in Canvas in the appropriate discussion thread; uploaded files are not sufficient. Reading notes document a student’s engagement with the weekly readings. Notes can engage either required or supplemental readings. Notes don’t need to be structured as a formal essay. Scattered thoughts and reactions, bullet points, sketches of ideas, etc are fine. However, notes should be primarily in your own words. Quotes or direct paraphrasing from the source material do not count towards the word count for notes. You can include quotes you find important or interesting, but you should also explicitly explain and react to the quote in your own words. Notes will be scrutinized for plagiarism, so please be careful to write your notes in your own words, and to cite anything from the source! Notes are due by the start of class on Tuesday and Thursday. Notes submitted before class earn full credit. Notes submitted after class starts earn partial credit. Notes submitted after Friday are considered late and receive a late penalty. See the rubric on Canvas for details.

Presentations: Students must prepare a 10-15 minute presentation on one of the readings in class in Units 1 and 2. Slides are encouraged but not required. The presentation should offer a close reading of the text, summarizing and explaining (in the student’s own words) the main conclusions, concepts, and perspectives discussed in the readings. Presentations must engage the primary readings, but they can also engage with supplemental readings and independent research, provided that the primary reading is discussed sufficiently. Students can work individually or in pairs, but in either case students presenting on the same day should coordinate beforehand to ensure coverage of the material. Students will also have a series of

informal presentations associated with the audit project, which are graded as participation (see below).

Participation: Participation credit is earned by participating in audit project activities. This includes regular attendance during the audit project, as well as regular participation and engagement in audit presentations, discussions, and group work. Participation credit will also be evaluated by group members in the audit project, which will be factored into the final participation grade.

Papers are short writing assignments that substantively engage with the debates and ideas found in the readings and lectures for each unit of the course. Papers should demonstrate a clear understanding of the issues presented in the readings and careful critical analysis of the texts. Papers can be argumentative and defend a particular position or controversial thesis in the debate. Papers can also be clarificatory, seeking to elucidate some complex issue or concept through additional research and reflection. Papers will be developed over several activities at the end of the semester, during which proposals and drafts will be made available for peer review and feedback.

Audit project: Lessons 10-15 will develop an elaborate group project that will involve an informal external audit of some popular AI software online, likely ChatGPT and DALL-E. Students will be divided into groups for different AI systems, which will again be divided into “Scoping” and “Testing” groups. These groups will gather research and strategize an approach to auditing these systems by preparing a social impact assessment, a FMEA chart and testing schedule, and other critical components of a thorough audit. Students will compile these tools into a final report and review of the software. Students are expected to discuss and present on their team’s progress during class meetings, to collaborate with their group to complete their part of the project on time, and to contribute to the final report and assessment. Students will be asked to grade each other’s performance and contributions to the group project. Grades on the audit project will depend on a student’s presentations, participation in audit activities, and on the quality of the final report.

Reflection Essay: At the end of the semester, students are asked to reflect on their work in a short reflection essay. I’m specifically interested in feedback on how the audit project went, what worked or didn’t work about the project, and any insights students gained on the status and operation of AI systems by working on the project. Students can also reflect more generally on the state of AI Ethics, and reactions to the reading and lesson material this semester. Reflection essays should be 2-3 pages (600-900 words). See Canvas for details.

Accessibility policy: I want all students to succeed in this class, and I will gladly accommodate the special circumstances and needs of all students to make sure that happens. I understand that life doesn’t happen on the semester schedule, and that school work can’t always be a top priority. In pandemic conditions we all need to be more flexible with scheduling and difficult work conditions; I understand how medical issues or disability can complicate these challenges. If there is any issue impacting your performance in class, please come talk to me in office hours or

send me a message by email or on Canvas! Even if you're behind on assignments, drop me a message letting me know what's up, I'm sure we can figure something out =)

Late policy: Assignments earn a small late penalty for material submitted after the assignment due dates posted on Canvas. I'll allow a short (~30 min) grace period for assignments due at midnight; assignments received at 12:01am will not be marked as late, but assignments received at 2am will. After one week, the penalty is increased to 50%. At the start of Unit 2, no additional Unit 1 assignments will be accepted for credit.

Excused Absences: If you have a legitimate excuse that you know about in advance (an academic conference, athletic event, National Guard duty, expected delivery date, etc.), please make arrangements with me in advance. Extensions for anticipated issues must be arranged at least 48 hours before a deadline to avoid a late penalty. Unexpected emergencies (medical emergencies, deaths in the family, etc.) should be brought to the attention of the Dean of Students with the Student Concern Reporting Form. The Dean's office is equipped to verify your situation confidentially and provide the administrative support you need. The Dean's office can also coordinate with all your instructors for any issues that arise. After an emergency and when you are able to return to school work, let me know what's up (a short note will do). I'll recommend you contact the Dean with the form linked above if you haven't already, and we can discuss a plan for completing your missing work, and go from there.

Plagiarism Detection: Students are expected to submit their work to plagiarism detection on Canvas. Failure to submit to plagiarism detection will result in zero credit for these assignments, forfeiting all extra credit for the semester, and a referral to the Dean of Students for violation of the plagiarism policy. Repeated or extreme instances of plagiarism will be treated as a violation of the Student Code of Academic Integrity and referred to the Dean of Students with a recommendation to fail the course.

Copying and pasting from the web is one form of plagiarism. **Failing to provide adequate citations is also a form of plagiarism. Copying from your own work counts as plagiarism. Changing a few words in an extensively quoted passage is a form of paraphrase and may constitute plagiarism.** If you are unsure of what constitutes plagiarism, please check this website for a detailed explanation of paraphrase and plagiarism. Any work you use should be given adequate citation so I can find and review your sources. Just as in mathematics, you need to show your work! If you use any source in your research, (including Wikipedia and other encyclopedias) *even if you don't quote them directly*, provide a citation.

NJIT Plagiarism Policy

"Academic Integrity is the cornerstone of higher education and is central to the ideals of this course and the university. Cheating is strictly prohibited and devalues the degree that you are working on. As a member of the NJIT community, it is your responsibility to protect your educational investment by knowing and following the academic code of integrity policy that is found at:

http://www5.njit.edu/policies/sites/policies/files/academic-integrity-code.pdf.

Please note that it is my professional obligation and responsibility to report any academic misconduct to the Dean of Students Office. ***Any student found in violation of the code by cheating, plagiarizing or using any online software inappropriately will result in disciplinary action. This may include a failing grade of F, and/or suspension or dismissal from the university.*** *If you have any questions about the code of Academic Integrity, please contact the Dean of Students Office at dos@njit.edu"*

Grades are explicitly calculated using the grading rubric below. For various reasons Canvas may show a different grade breakdown, but the rubric below is the official grading policy.

Grades:

Attendance 100
Participation 100
Reading notes
Introduction 10
Unit 1/2 notes: 190
12pts x 16
Unit 1/2 Presentation: 200
Unit 1 paper: 75
Midterm: 75

Audit presentations: 100
Audit participation: 100
Audit report: 100
Final reflection essay: 50

Reading Schedule

The first reading is required, and should be the focus of presentations and notes. Other readings are supplemental and can be included in presentations and notes.

Unit 1: Background on computing
Lesson 1: History of Computing
W 9/6 Race, Gender, and Technology

- Benjamin (2019) Race after technology Intro, Ch 3
- Mullaney et al (2021) Your Computer on Fire Intro, Ch 1, 6, 7
    - BobbyBroccoli (2022, YouTube) The image you can't submit to journals anymore
    - Edwards (1994) Computers in Society and Culture

Lesson 2: History of AI
M 9/11 Technological Redlining

- Noble (2018) Algorithms of Oppression Intro, Ch 1

    - Benjamin (2019) Race after technology Ch 2
    - Cave & Dihal (2020). The whiteness of AI

W 9/13 Your robot is a human

- Mullaney et al (2021) Your Computer on Fire Ch 2, 8, 9
    - Benjamin (2019) Race after technology Ch 1
    - Cave, Dihal, & Dillon (2020). *AI narratives* Ch 8, 9, 13, 15

Lesson 3: Neural Networks

M 9/18: Artificial Neural Networks

- Kriegeskorte & Golan (2019) Neural Network Models and Deep Learning
    - Marcus (2018) Deep Learning: A critical appraisal
    - LeCun, Bengio, & Hinton (2015) Deep Learning
    - Hinton (2007) The next generation of neural networks
    - Tensorflow Playground (demo)

W 9/20: LLMs and other foundation models

- Bender et al (2021) On the dangers of stochastic parrots

  3blue1brown: Neural Networks. video series (S3 E1-4)

  Computerphile: Neural Networks video series
    - How AI image generators work
    - Stable Diffusion in code
    - How GPT3 works
    - AI Language models and transformers

  Bommasani et al (2021). On the opportunities and risks of foundation models.

Lesson 4: Can machines think?

M 9/25: Predictive processing

- Clark (2013) Whatever next?
    - Chalmers (2012) Computational foundations of cognitive science
    - Turing (1950) Computing Machinery and Intelligence
    - Webb (2019) Insects as a cognitive edge case
    - van Rooij et al (2023) Reclaiming AI as a theoretical tool for cognitive science

W 9/27: Problematic analogies

- Bender (2022) Resisting dehumanization in the age of AI
    - Bender and Koller (2020) Climbing towards NLU
    - Baria and Cross (2021) The brain is a computer is a brain
    - Hayles (2019) Can computers create meanings?
    - Piccinini (2010) Mind as neural software?
    - Mitchell & Krakauer (2022) The Debate Over Understanding in AI's Large Language Models

F 9/29: **U1 paper due.**

Unit 2

Lesson 5: Introduction to AI Ethics

M 10/2: Transparency and Accountability

- Crawford (2021) *The atlas of AI* Intro, Ch 2, 3, 4
    - Kate Crawford and Vladan Joler (2018) Anatomy of AI

    - Licht & Licht (2020) AI, Transparency, and Public Decision-making
    - Doshi-Velez et al (2017) Accountability of AI Under the Law
    - Eschenbach (2021) Why we do not trust AI

W 10/4: Fairness

- Whittaker (2021) The steep cost of capture
    - Ruha Benjamin (2019) Race after technology Ch 2
    - Coded Bias (2020) documentary
    - NeurIPS workshop (2017, Vimeo) Fairness in machine learning
    - Bennett & Keyes (2020) What is the point of fairness?

Lesson 6: AI and Justice

M 10/9 Algorithmic policing

- Angwin et al (2016) Machine bias in sentencing
    - Asaro (2016) Hands up, don't shoot!
    - McGuire (2021) The laughing policebot
    - Berk et al (2021) Fairness in criminal justice risk assessments

W 10/11 Algorithmic injustice

- Birhane (2021) Algorithmic Injustice
    - Gabriel (2022) Towards a theory of justice for AI
    - Birhane et al (2022) The forgotten margins of AI Ethics
    - Keyes (2020) Automating autism
    - Ostrowski et al (2022) Ethics, equity, and justice in HRI

Lesson 7: Autonomous weapons and vehicles

M 10/16 Autonomous weapons

- Roff and Moyes (2016) Meaningful Human Control, Artificial Intelligence, and Autonomous Weapons
    - Amoroso & Tamburrini (2020) AWS and Meaningful Human Control
    - Sharkey (2018) AWS, Killer Robots, and Human Dignity
    - Asaro (2016) Autonomous weapons

W 10/18 Autonomous Vehicles

- Lin (2016) Why Ethics Matters for Autonomous Cars
    - Koopman and Wagner (2017) Autonomous Vehicle Safety
    - MIT: Moral Machine (2017 Publication)
    - Jacques (2019) Why The Moral Machine is a Monster
    - NHTSA Topic Overview: AV Safety

Lesson 8: AI Policy

M 10/23: Ethics frameworks

- Jobin et al (2019) The global landscape of AI Ethics guidelines
    - Floridi and Cowls (2022). A unified framework of five principles for AI in society.
    - Bollier (2019) Artificial Intelligence and the Good Society
    - Greene et al (2019) Better, nicer, clearer, fairer

W 10/25: Global AI Policy

- Hagendorff (2019) The Ethics of AI Ethics–An Evaluation of Guidelines
    - Arun (2019) AI and the Global South
    - Amrute et al (2022) A Primer on AI in/from the majority world

    - Keyes et al (2019) Human-computer insurrection: Notes on an anarchist HCI.

F 10/27 U2 paper proposal due

Lesson 9: Algorithmic Audits

M 10/30: **Gender Shades**

- Buolamwini and Gebru (2018) Gender Shades
    - Raji and Buolamwini, J. (2019). Actionable auditing
    - Costanza-Chock, Raji, & Buolamwini (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem

W 11/1: AI Audits

- Raji et al. (2020). Closing the AI accountability gap
    - Keyes and Austin (2022). Feeling fixes: Mess and emotion in algorithmic audits
    - Mokander et al (2023) Auditing large language models: a three-layered approach

Lesson 10: Research

M 11/6: **U2 Draft due. In class peer review.**

- Smile detection case study

W 11/8: Preliminary Research

F 11/10 U2 paper due

Lesson 11: Scoping

M 11/13 Group work & updates **LAST DAY TO WITHDRAW**

W 11/15 Social Impact Assessment presentations

Lesson 12: Testing

M 11/20 Group work & updates

THANKSGIVING BREAK

Lesson 13: Testing continued

M 11/27: Group work & updates

W 11/29: Completed FMEA chart presentations

Lesson 14: Conclusions

M 12/4: Final testing results

W 12/6: Summary Reports & Remediation plans

Lesson 15: Post-Audit Reflection

M 12/11: Open Discussion. Last Day of class

F 12/15 **Reflection Essay due, Final Audit Report due, Extra Credit assignment due.**

Readings:

Unit 1

Week 1: History of computers & AI

- Cave & Dihal (2020). The whiteness of AI.
- Turing (1950) Computing Machinery and Intelligence
- Computerphile: Babbage and Lovelace video series
- Wolfram: Untangling the tale of Ada Lovelace

Week 2: Computation and algorithms

- Piccinini (2010) Mind as neural software?

- Computerphile: Turing machines video series
- Numberphile: Godel's incompleteness theorem
- Aaronson, S. (2013). Why philosophers should care about computational complexity. *Computability: Turing, Gödel, Church, and Beyond*, *261*, 327.

Week 3: Machine learning basics

- 3blue1brown: Neural Networks video series (S3 E1-4)
- Computerphile: Neural Networks video series
    - How AI image generators work
    - Stable Diffusion in code
    - How GPT3 works
    - AI Language models and transformers
- Chalmers (2012) Computational foundations of cognitive science
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Week 4: Can machines think?

- Turing (1950) Computing Machinery and Intelligence
- Haugeland (1994) On the nature and plausibility of cognitivism
- van Rooij, I. (2008). The Tractable Cognition thesis. *Cognitive Science*, 32, 939-984.
- Baria and Cross (2021) The brain is a computer is a brain

Supplemental: Social robots and robot rights

- Rini (2017) Raising good robots
- Gunkel (2016) Can and should robots have rights?
- Gunkel (2018) Robot Rights
- Danaher (2017) Should Robots have Rights? Four perspectives
- Danaher (2020) Welcoming robots into the moral circle: a defense of ethical behaviorism
- Darling (2015) "Who's Johnny?" Anthropomorphic framing in human-robot interaction, integration, and policy
- Darling (2016) Extending legal protection to social robots
- Bryson (2010) Robots should be slaves
- Bryson (2017) Of by and for the people: the legal lacuna of synthetic persons

Unit 2

Week 5: AI Ethics overview

- Virginia Eubanks (2018) Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press
- Ruha Benjamin (2019) Race after technology: Abolitionist tools for the new jim code. John Wiley & Sons.
- Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence.* Yale University Press.
- Cave, S., Dihal, K., & Dillon, S. (Eds.). (2020). *AI narratives: A history of imaginative thinking about intelligent machines*. Oxford University Press.
- Asaro (2006) What should we want from a robot ethic?
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and machines*, *28*(4), 689-707.
- Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. *Machine Learning and the City: Applications in Architecture and Urban Design*, 535-545.

Week 6: Autonomous weapons and vehicles

- Haselager (2005) Robotics, philosophy, and the problems of autonomy
- Lin (2016) Why Ethics Matters for Autonomous Cars
- MIT: Moral Machine (2017 Publication)
- Jacques (2019) Why The Moral Machine is a Monster
- Roff and Moyes (2016) Meaningful Human Control, Artificial Intelligence, and Autonomous Weapons
- Asaro (2016) Autonomous weapons
- P. Koopman and M. Wagner, "Autonomous Vehicle Safety: An Interdisciplinary Challenge," in *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 1, pp. 90-96, Spring 2017. doi: 10.1109/MITS.2016.258349 URL:
- NHTSA Topic Overview: AV Safety — https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety

Week 7: AI Principles: Fairness, Transparency, Accountability

- Hagendorff, Thilo. "The Ethics of AI Ethics–An Evaluation of Guidelines." arXiv preprint arXiv:1903.03425 (2019).
- Bennett, C. L., & Keyes, O. (2020). What is the point of fairness? Disability, AI and the complexity of justice. *ACM SIGACCESS Accessibility and Computing*, (125), 1-1.
- Keyes, O., Hoy, J., & Drouhard, M. (2019, May). Human-computer insurrection: Notes on an anarchist HCI. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1-13).
- Keyes, O. (2020). Automating autism: Disability, discourse, and artificial intelligence. *The Journal of Sociotechnical Critique*, *1*(1), 8.
- Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature Machine Intelligence 1, no. 5 (2019): 206-215. https://doi.org/10.1038/s42256-019-0048-x

Week 8: Case studies

- Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner (2016) Machine bias in sentencing Propublica
- Kate Crawford and Vladan Joler (2018) Anatomy of AI
- Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Week 9: AI Audits literature

- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33-44).
- Raji, I. D., & Buolamwini, J. (2019, January). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 429-435).
- Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022, June). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1571-1583).
- Keyes, O., & Austin, J. (2022). Feeling fixes: Mess and emotion in algorithmic audits. *Big Data & Society*, *9*(2), 20539517221113772.

Week 10: Smile detection case study

Supplemental readings: Robot Rights

Gunkel (2018) Robot rights

Gunkel (2016) Can and should robots have rights?

Danaher (2017) Should Robots have Rights? Four perspectives

Danaher (2020) Welcoming robots into the moral circle: a defense of ethical behaviorism

Salvini (2016) How safe are robots in urban environments? Bullying a service robot

Bryson (2010) Robots should be slaves

Bryson (2017) Of by and for the people: the legal lacuna of synthetic persons

Birhane and van Dijk (2020) Robot rights? Let's talk about Human Welfare instead

Darling (2015) "Who's Johnny?" Anthropomorphic framing in human-robot interaction, integration, and policy

Darling (2016) Extending legal protection to social robots

Estrada (2020) Human supremacy as posthuman risk

Estrada (2017) Robot rights: cheap yo!

Estrada (2017) Alignment, fair play, and the rights of service robots

Estrada (2018) Sophia and her critics

Project 1: Automate a decision
Pick some routine task or decision in your life that you perform on a regular basis (for example, deciding what to eat for dinner), and sketch an algorithm that attempts to automate the decision making process by breaking the task down into simple and repeatable subtasks that can be executed by a computing machine. What are the inputs or initial states, and what are the outputs or final states of the algorithm? What properties must be measured or evaluated? How would you train an algorithm, and where do you find the training data? How reliably would your algorithm yield the correct or desired results? Would such machines be realistic or useful? Describe the algorithm as fully as possible, as well any technical challenges faced in automating the task. Finally, imagine your algorithm was implemented in a consumer product. What might be the social, political, and ethical consequences of bringing your algorithm into the world?

Students will prepare a short (3-5 pg) paper discussing their algorithm, potential challenges in its implementation, and offering an ethical analysis of its development and use.

Unit 2: Topics in AI Ethics

Week 5: AI Ethics overview
Week 6: Autonomous weapons and vehicles
Week 7: AI Principles: Fairness, Transparency, Accountability
Week 8: Case studies: Anatomy of AI, Gender Shades
Week 9: AI Audits literature

Project 2: Presentations

Students will present and discuss readings from this unit in class. Students can present individually or in pairs. Presentations should include slides, and should be 10-15 minutes in length, with another 5-10 minutes for Q&A.

Unit 3: Audit an AI

Week 10: Audit planning and preparation. Smile detection
Week 11: Scoping
Week 12: Mapping and Artifact Collection
Week 13: Testing
Week 14: Reflection
Week 15: Workshop activity

Following the auditing literature from Raji et al, the class will engage in a group project designed to complete an informal audit and ethical review of some existing AI technology. Audits consist of several stages, from scoping and mapping to artifact collection, testing, and reflection. These stages will be distributed among the class members to tackle various stages of this process. Students are each responsible for contributing to and presenting on their designated part of the audit process. On the basis of these audits, students will engage in a class workshop activity to develop a set of general guidelines and recommendations for the safe and ethical use of the technology. Together, the audit documents will be compiled into a final report assessing the technology.

Students are asked to write a short reflection essay at the end of the semester summarizing their work.

Project 3: Implement the Guidelines

For the final project, students are asked to prepare a report (10-15 pgs) highlighting the key results of their AI audit. Then, students should consider the guidelines and recommendations developed in the class workshop, and how these guidelines might address (or fail to address) the ethical challenges identified by their audit. To what extent do the class guidelines address the important ethical issues at stake? To what extent can these guidelines be enforced to ensure the safe and ethical use of AI?

Grades:

Project 1 paper: 20%
Project 2 presentation: 20%

Project 3 Audit Contribution: 20%
Final reflection essay: 10%
Participation & Attendance: 30%