HSS 404 - AI Ethics

Fall 2024

Instructor: Dr. Daniel Estrada
E-mail: estrada@njit.edu
Zoom Meeting Link
Office: Cullimore 419
Office Hours: T 12-1pm & by appt.
Discord: discord.gg/NxFvdH7

Class Meeting:

HSS 404-061 TF 10:00am - 11:20am CKB 126
HSS 404-067 TF 1:00pm - 2:20pm CULL 110

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

Lesson Plan

Unit 1: Background on computing and AI

- Lesson 1: History of Computing
- Lesson 2: History of AI

- Lesson 3: Neural Networks
- Lesson 4: Can machines think?

Unit 2: Topics in AI Ethics

- Lesson 5: Introduction to AI Ethics
- Lesson 6: AI and Justice
- Lesson 7: Autonomous weapons and vehicles
- Lesson 8: AI Policy
- Lesson 9: Algorithmic Audits

Unit 3: AI Audit project

- Lesson 10: Planning and Research
- Lesson 11: Scoping
- Lesson 12: Testing
- Lesson 13: Testing 2
- Lesson 14: Scoping 2
- Lesson 15: Course wrap up

Assignments and expectations

- **Attendance:** Regular classroom attendance is **required**. Students can miss 3 classes without penalty. (10%)
- **Participation** includes the introduction, scheduling podcasts, and final reflection essay (10%)
- **Reading notes**: 300+ words of "reading notes" due for each class reflecting on reading assignments in Lesson 1-9. (25%)
- **Presentation**: One 10-15 min presentation with slides on readings in Units 1 or 2. Schedule your presentation on Canvas. (10%)
- **Midterm Paper:** Students are required to complete one 7-10 page paper on a topic of their choosing from Units 1 and 2. Will include a proposal and drafting round. (10%)
- **Podcasts:** Students are required to participate in a 1 hour recorded group conversation at the end of Unit 1, 2, and 3. (15%)
- **Audit Project:** Students will work in groups on the Audit project. Requires collaboration with group members on the project, participating in group presentations on the project, and contributing meaningfully to the final report. (20%)

Grading policy

Assignment details can be found in the syllabus and on Canvas

- **Attendance 100 pts**
    - 29 class meeting days
    - 3 allowed absences
    - **= 26 required attendance days**
    - **= 4 pts/day**
    - May earn extra credit for exceptional attendance

- **Participation 100 pts**
    - 20 pts Introductions
    - 30 pts podcast scheduling
    - 50 pts Final Thoughts
- **Reading Notes 250 pts**
    - 18 required notes
    - **= 15 pts each if submitted before class**
    - **= 12 pts each if submitted by Friday**
    - **= 10 pts each if submitted by the end of Unit**
- **Presentation 100 pts**
    - 10-15 minutes with slides covering some readings from Unit 1 or 2.
- **Midterm paper 100 pts**
    - Midterm grade includes drafting round
    - 10 pts Midterm proposal
    - 15 pts Midterm draft
    - 75 pts Midterm paper
- **Podcasts 150 pts**
    - 50 pts Unit 1 Podcast
    - 50 pts Unit 2 Podcast
    - 50 pts Unit 3 Podcast
- **Audit Project 200 pts**
    - 50 pts Presentation 1
    - 50 pts Presentation 2
    - 100 pts Final audit report

Total Grade Points = 1,000 pts

Grade Scale:

Final grades are calculated on the following scale:

A: 900+
B+: 850+
B: 800+
C+: 700+
C: 600+
D: 500+
F: < 500

There is a 5 point tolerance for bumping a grade to the next letter when calculating final grades.

Assignment Details

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 29 total days of class, so 26 attendance days earn full credit. On-time attendance counts for one day. Attendance is considered late if registered more than 10 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 physically in the classroom will lose all attendance credit for the semester.

Reading Notes: Students are expected to complete 300+ words of reading notes before 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 submitted before class earn full credit. Notes submitted after class starts earn partial credit. Notes are accepted with a late penalty until the end of the Unit. See the rubric on Canvas for grading details.

Presentations: Students must prepare one 10-15 minute presentation on one or more of the Lesson readings in class for Units 1 or 2. Must include informative slides, and students should not just read directly from the slides. 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 to some extend, but they can also engage with supplemental readings and independent research, provided that the primary reading and lesson themes are 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. If working in pairs, presentations should be 20-30 min. in length, and should divide that time between the students. Students will also have a series of informal presentations associated with the audit project, which are part of the audit project grades.

Podcasts: Podcasts are 1 hour live, recorded conversations among students in a podcast group. Podcasts are required at the end of each unit. Students are responsible for scheduling their conversation outside of class. Students should record their podcasts using video conferencing software like WebEx. Video is optional but clear audio with a working microphone is required. The podcast format is divided into two segments. In the first segment, students will

review the material from that Unit. In the second segment, students will have a debate or discussion prompt to engage with. See details on Canvas.

Participation: Participation credit is earned for the Introductions and Reflection Essay thread that bookend the semester. See Canvas for details. Participation also depends on timely scheduling of podcasts at the end of the Unit with Podcast groups, and on cooperating effectively with audit groups for the final audit project.

Midterm Paper is a 7-10 page scholarly writing assignment that substantively engages 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 in Unit 2, 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. 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: Reading notes are due before class according to the reading schedule on Canvas. Reading notes will earn a small late penalty if submitted after class before Friday. Reading notes will earn a larger penalty if submitted after Friday but before the end of the unit. **Reading Notes will not be accepted after the unit has concluded.** Other assignments, such as the Midterm paper and reflection essays, are typically due on Canvas by midnight on the day posted, and may be subject to a late penalty if posted after this deadline. See Canvas for details.

Excused absences: In an emergency situation or unplanned special circumstances that disrupt your capacity for school work, please attend to the emergency situation as a top priority! When you are ready for school work again, contact the Dean of Students through the links above to schedule an appointment where you can explain your situation. You don't need to share doctors notes or other personal information with me; my policy is the same regardless of the details of your situation. When you contact me, I'll work with you to plan out a way to make up missing assignments and recover your grade. **When I hear from the Dean of Students, I will waive any late penalties that might have accrued.**

For any non-emergency events, such as athletic events, academic conferences, job fairs, military service, or busy schedules around midterms and finals, I ask that you contact me at least 2 days in advance of the event to reschedule your assignments. In other words, **extensions will not be granted on the day an assignment is due.** If you contact me at least 2 days ahead of an event, we can arrange some rescheduling of assignments to accommodate your event.

Plagiarism Policy

Plagiarism Slides

Plagiarism means using work that you did not produce, but presenting it as if it is your own work in assignments. If you did not write the words yourself, you must clearly distinguish that work from your own with quotes and citations. When I am scanning for plagiarism I am looking for long blocks of text that are clearly taken from other sources (possibly with minor modifications) without proper attribution and without distinguishing it from the students own work. Passing off the work of another for credit is plagiarism, and it will not be tolerated in an ethics course.

Copying and pasting from the web is a form of plagiarism. Changing a few words in an extensively quoted passage is a form of plagiarism. Using AI text generators like chatGPT is a form of plagiarism. Failing to provide adequate citations is a form of plagiarism. Copying from your own work (including work from previous semesters) without acknowledgement counts as plagiarism. *In general, you should never copy large blocks of text from any other source and present it in your own essay as if it were your own words. That includes copying text from online text generators or language translators. Check*

this link for a detailed explanation of legitimate paraphrase and illegitimate plagiarism. Any work you use should be given adequate citation so your readers 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 dictionaries, Wikipedia and other encyclopedias, and translation tools) even if you don't quote it directly, provide a citation.

To avoid plagiarism, you must clearly distinguish your work from the work of others. Any work taken from others must be identified with "quotation marks" and explicit citation. Changing a few words in a quote does not make it your work. *If you use online text generators (like chatGPT, Grammarly, or other text sources), you must explicitly identify that text as not being your own work. You must also cite the explicit generator used, including the version and dates it was used. If you use AI text generators at all, you must also supply the full prompt history generating that text as an appendix to your assignment. If you wrote the essay in another language and then used a translator, you should provide the original text in the original language with your submission. If you read a script in any presentation, you must include the text of that script to the plagiarism detection software on Canvas. If you translate your essay from another language, you must include the original untranslated text for comparison. Failure to do so will not earn credit.*

Suspected cases of plagiarism will be given zero credit for the assignment with a warning about the plagiarism policy. Students found plagiarizing will also forfeit all extra credit opportunities for the semester. *Repeated or extreme instances of plagiarism will be reported directly to the Dean of Students as a violation of the Student Code of Academic Integrity* ***Note:*** *the research project is a honeypot for cheaters, and typically results in multiple instances of plagiarism in each section. I won't hesitate to fail students who cheat in my ethics course. Consider this your first warning.*

I have substantially reorganized my class around group discussions and presentations to discourage the use of AI text generators. None of the writing assignments in class are "busy work". They all ask you to demonstrate direct engagement with the readings and with the ideas and perspectives of your fellow students. Please take this opportunity to engage your peers in discussions on ethics seriously!

See these Plagiarism Slides with detailed information on the NJIT and course policies on plagiarism, including examples of legitimate and illegitimate paraphrase, to help you understand the plagiarism policy.

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"*

Readings and Assignment Schedule

The first reading is required, and should be the focus of presentations. Other readings are supplemental and can be included in presentations and notes. **Notes are required by the start of class** for every lesson with reading assignments. Other assignment due dates are highlighted in **bold**. See the syllabus and Canvas for details.

Unit 1: Background on computing
Lesson 1: History of Computing
History of computing and AI

- Mullaney et al (2021) Your Computer on Fire Intro, Ch 1, 6, 7
    - Estrada (2023) History of AI audio lecture and slides

Race, gender, and technology

- Benjamin (2019) Race after technology Intro, Ch 2, Ch 3
    - BobbyBroccoli (2022, YouTube) The image you can't submit to journals anymore
    - Noble (2018) Algorithms of Oppression Intro, Ch 1
    - Cave & Dihal (2020). The whiteness of AI

Lesson 2: History of AI
Technological Redlining

- Eubanks (2018) Automating Inequality Intro, Ch 1, 3, 5
    - O'Neil (2016) Weapons of Math Destruction Intro, Ch 1, 5, 8
    - Benjamin (2019) Race after technology Intro, Ch 1, 2, 3

Your robot is a human

- Coded Bias (2020) documentary
    - Mullaney et al (2021) Your Computer on Fire Ch 2, 8, 9
    - Cave, Dihal, & Dillon (2020). *AI narratives* Ch 8, 9, 13, 15

Lesson 3: Neural Networks
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)
    - Tensorflow Embedding Projector (demo)

LLMs and other foundation models

- Vaswani et al (2017) Attention is all you need
- Bender et al (2021) On the dangers of stochastic parrots

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

Art of the Problem: How neural networks learned to talk, how NNs learn, how NNs learn concepts

Computerphile: Neural Networks video series

- How AI image generators work
- Stable Diffusion in code
- How GPT3 works
- AI Language models and transformers work

Lesson 4: Can machines think?

Predictive processing

- Clark (2013) Whatever next?
    - Chalmers (2012) Computational foundations of cognitive science
    - Chalmers (2022) Could a large language model be conscious
    - 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

Problematic analogies

- Mitchell & Krakauer (2022) The Debate Over Understanding in AI's Large Language Models
- 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?
    - Estrada (2024) Can a robot hand grasp?

U1 podcast due

Unit 2

Lesson 5: AI Ethics

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

Fairness

- Whittaker (2021) The steep cost of capture

    - NeurIPS workshop (2017, Vimeo) Fairness in machine learning
    - Bennett & Keyes (2020) What is the point of fairness?
    -

Lesson 6: AI and Justice

Algorithmic policing

- Angwin et al (2016) Machine bias in sentencing
    - ONeil (2018) Weapons of math destruction Ch 1, 3, 5
    - Asaro (2016) Hands up, don't shoot!
    - McGuire (2021) The laughing policebot
    - Berk et al (2021) Fairness in criminal justice risk assessments

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

Midterm Paper proposal due

Lesson 7: Autonomous weapons and vehicles

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

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

Ethics frameworks

- **Midterm draft due in class for peer review**
- 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

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.

Lesson 9: Algorithmic Audits

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

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

U2 Podcast due
Lesson 10: Planning and Research
Audit proposals due in class
Audit preliminary research

- Smile detection case study

Lesson 11: Scoping
LAST DAY TO WITHDRAW
Group work & updates
Scoping group presentations
Lesson 12: Testing
Group work & updates
Testing group presentations
Lesson 13: Group work
Group work & updates
Intra-group assessments
Lesson 14: Conclusions
Final Testing group presentations
Final Scoping group presentations

Lesson 15: Post-Audit Reflection
Open Discussion. Last Day of class
U3 Podcast Due
Reflection Essay due, Final Audit Report due.

Supplemental readings: Robot Rights

Gunkel (2023) Person, Thing, Robot
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