AI Ethics for Technology Leaders

EGN6933/AITL/29530
Class Periods: WEDNESDAYS, 4.05-7.05PM Eastern Time, Periods 9, 10, 11
Location: Zoom https://ufl.zoom.us/j/93677846404?pwd=oW7FxFdUPWkkGZaLFhXsRwrwLJ7wOU.1
Academic Term: Fall 2024

Instructor:
Sonja Schmer-Galunder
Email: s.schmergalunder@ufl.edu
Phone: 415.604.6293
Office Hours: Thursdays 4-5PM ET via zoom https://ufl.zoom.us/j/98717368862

Teaching Assistant/Peer Mentor/Supervised Teaching Student:
Please contact through the Canvas website

- Divyansh Singh, Email: divyansh.singh@ufl.edu, Phone: 352.740.6854
- Mondays, 3:30-5:30 PM
- Link: https://ufl.zoom.us/my/divyanshmeetingroom or MALA student village

Course Description
This course equips future technology leaders to understand the ethical considerations of developing and deploying AI systems - computational systems that use large datasets to train predictive models and act on their outputs. Students will learn core concepts of AI ethics and apply them to real-world scenarios where leaders have faced ethical challenges. We will assess the impact of AI within a global context, navigating complex issues while respecting diverse social and cultural values. By the course's end, students will understand AI as part of a larger socio-technical system and be able to evaluate its impact on society and individuals globally. The ultimate goal is to foster responsible development of safe and beneficial AI. Upon successful completion, you will receive 3 course credits.

Course Objectives:

- **Learn how AI systems operate from the perspective of technology leaders and be able to think critically about the impact of AI systems on society.** Students will learn to understand the impact of AI on global markets, diverse societies and small communities, and practice applying ethical thinking when making decisions.

- **Understand relevant ethical concepts and frameworks and how to apply them. I**n other words students should apply conceptual tools from the humanities and social sciences to engineering and computer science problems. Students will learn basic ethical theories and common ethical issues that technology leaders had to handle in the past, including those negatively impacting the well-being of individuals and society. Some case studies will allow students to wrestle with real-world decisions with no clear answers, or answers that may depend on the **socio-cultural context**.

- **Learn to understand frameworks and strategies for influencing public policy and company policies for the development of safe AI systems.** Students will examine research on how people make value tradeoffs and how those tradeoffs are influenced by safety, responsibility and transparency considerations. Increasing the awareness of specific ethical considerations during the development of AI systems will be highlighted such that students can apply those frameworks in their future careers and decisions.

- A critical skill set of any technology leader are **interpersonal skills**. This class should give everyone an opportunity to work effectively in teams, develop communication and presentation skills, engage in peer review and direct one's own learning.

Motivation
We live in what is perhaps the most exciting but also challenging period in human history. AI is reshaping the world, bringing both opportunities and challenges for leaders. This course equips leaders to navigate the ethical decisions they will have to make by introducing them to important domains related to harms and risks, AI safety, accountability, fairness, transparency, explainability and social impact, with a specific focus on cultural differences around the globe.

Practical Considerations

- This course is taught remotely, with the **possibility** to **meet in person** on **9/18 and 11/20.** We **will** meet **in person** on **12/11 for presentations.**
- The course starts at 4.05PM and ends at 7.05PM, roughly following this outline:
    - 4.05PM Start: Administrative, Questions, Attendance
    - 4.15-4.25PM: Quiz on Canvas - at the beginning of each class there will be a quiz that will test you on the reading materials listed in the syllabus. **You are not allowed to use AI.** The quiz counts towards your final grade. We use HonorLock.
    - 4.25-5.05PM: Lecture covering the reading materials related to the topic
    - 5.05-5.20PM: Breakout room for AI news reporting
    - 5.20-5.30: Break
    - 5.30-6.15: **Guest Lecture from industry technology leaders -** each week we'll hear from technologists and researchers, many from silicon valley and top AI companies (this is a great opportunity to ask smart questions - come prepared!), 30 min lecture, 15 min questions
    - 6.15-6.30: Breakout room discussions (buffer, if we run over)
    - 6.30-7.05: Summarizing today's content and answering any questions you may have
- You are asked to have your zoom camera on throughout the class, and participate as much as possible!
- All students are required to write a research paper (min 5, max. 10 pages) and present it at the end of the semester to the whole class. You will be working in groups of 4-6 students.
- Some etiquette: Please arrive on time, have your microphone muted and your camera on, with good lighting, use the "raise hand" feature in zoom if you have a question or comment.

Note: The online component of this course uses both synchronous (that is, live online) and asynchronous formats throughout the semester. Students are fully responsible for lecture material delivered during all "live online" meetings. Students will also work asynchronously, both individually and in teams, to complete course requirements. Synchronous class sessions to be held in WERT 370 will be scheduled by the instructor in advance.

Required Computer
You will need to have a computer that allows either Windows or Mac to participate in the class.

Project Work & Grading

Class Participation (30%): All students need to attend class each week. If you cannot attend, please contact the instructors immediately by email. Everyone is expected to participate verbally at least once every class session, coming prepared, having done all assigned readings/media, and contributing thoughtful ideas to all class discussions. Students are expected to be prepared for cold calls during class.
Class Quizzes (15%): We will administer a short two-question survey each week at the beginning of class that will ask questions related to the reading material and media you had to prepare before class.

In-Class Presentation (25%): We will start each class with "current events" - a 5-min news update about the fast-moving world of AI. You'll have a chance to sign up for a time slot and prepare your short (~5-10 minute) presentation in advance, deliver it to the class (or breakout group), and take a few questions. We see this as a great opportunity to bring all of your voices into the conversation and reserve some time for topics that most interest you.
Final Class Presentation - IN PERSON (30%): On **Wednesday, December 11th 2024**, **4.05-7.05PM ET, Malachowsky Hall,** we'll gather as a class for final presentations. Each student will work with a team towards a **class project**, where students will present their paper, but also discuss some of the ethical challenges, and make recommendations to ensure AI can be beneficial, without causing harm. More assignment details will be shared in Canvas and explained in class within the first few weeks.

Scheduling Conflicts:
Please notify us by email (divyansh.singh@ufl.edu) by the second week of the term about any known or potential extracurricular conflicts (such as religious observances, interviews, or other activities). We will try our best to help you with making accommodations, but cannot promise them in all cases. In the event there is no mutually-workable solution, you may be dropped from the class.

Attendance:
Attendance is **mandatory. One absence per semester** does not affect your grade. This includes doctor appointments, family emergencies, recruiting sessions, etc. If you have a second absence for a reason beyond your control, please email Divyansh Singh (divyansh.singh@ufl.edu), and cc the instructor (s.schmergalunder@ufl.edu) and submit a note from the doctor, etc., specifying the date of class and that you can't attend. **Showing up at the beginning of class and leaving in the middle will result in zero attendance points** unless you explain why in advance via email. If you explain in advance, you can get a half-point if you stay for more than half the class.
Excused absences must be consistent with university policies in the Graduate Catalog (https://catalog.ufl.edu/graduate/regulations) and require appropriate documentation. Additional information can be found here: https://gradcatalog.ufl.edu/graduate/regulations/

Students **must** be prepared to have their **CAMERAS ON** during Zoom sessions, with appropriate attire and backgrounds. Let the instructor know ahead of time if you have a technical issue.

Cheating:
Anyone caught cheating will receive a failing grade and will also be reported to the University Office of Student Conduct.

Plagiarism/Self-plagiarism/Use of AI tools:
You must be original in composing your work in this class. To copy text or ideas from another source (including your own previously, or concurrently, submitted coursework) without appropriate reference is plagiarism and will result in a failing grade for your assignment and usually further disciplinary action. For additional information on plagiarism, self-plagiarism, and how to avoid it, see, for example:
https://gradadvance.graduateschool.ufl.edu/media/gradadvancegraduateschoolufledu/OGPD_Plagiarism_Workshop_20221019.pdf

Note that cheating on exams and plagiarism are examples of violations in the realm of ethics and integrity. Honesty, integrity, and ethical behavior are of great importance in all facets of life.

Use of AI tools (based on Mollick): During this course, we may ask you to use AI (ChatGPT, or a similar tool provided by UF etc), and you may decide to use these tools on your own. **When you use AI tools, you must acknowledge them, and not include work created by a tool (even as bullets in your presentations, for example) without mentioning it.** We ask that you add a final slide to each of your presentations explaining what you used AI for and what prompts you used to get the results. You will still be responsible for errors in the work. **We may use Turnitin or other similar tools to compare your writing to existing work.** Be aware that information provided by AI tools can be false. You are responsible to provide correct information and original source citations of your work.

Course Schedule

Week	Date	Topic	Description
1	8/28/2024	Introductions & Goal Setting	In this session we will introduce the course, study materials, course structure, grading, motivation and goals. Students will learn about the course objectives, meet each other, instructors and learn basic definitions, concepts and terminology.
2	9/4/2024	AI Ethics and Leadership	We talk about AI Ethics as a field and why it is important. We discuss various theories relevant to the ethical development and use of technology and AI. We discuss and define what ethical leadership means and think about important tools for future AI leaders to make ethical decisions for the deployment of technologies.
3	9/11/2024	Artificial Intelligence and the Problem of Unintended Consequences	In this class we talk about unintended consequences, harms and risks from AI systems and what technology leaders should think about. We talk about various types of problems that can arise when optimizing for objective functions.
4	9/18/2024	Bias: Human vs. Algorithmic	We learn about various types of biases. Why are AI systems biased and what can we do about it? How can we measure bias? How can we mitigate biases and improve fairness?
5	9/25/2024	Data: What is data?	In this session students will learn about the role of data in AI with a focus on ethical considerations, e.g. representation, inclusion, bias, consent, processing, etc. We will discuss data annotation processes and concerns related to public information, as well as trade-offs between data quality and quantity.
6	10/2/2024	Values and Norms	We take a deeper dive into norms, values and virtues. We learn about deductive and non-deductive arguments, fallacies and risks in ethical discussions, and the value alignment problem.
7	10/9/2024	Privacy and Transparency	IWe learn about the ethical questions related to privacy and privacy protection.

8	10/16/2024	Truth and Misinformation	We will talk about mis- and disinformation, and the importance of truth and trust. We learn about the use of algorithms that are optimized to influence people, and technology to make information more trustworthy and safe.
9	10/23/2024	Society: The Politics of Algorithms	In this class we will learn about the global context for the use and deployment of AI, including, geopolitical impacts and national security risks.
10	10/30/2024	Culture: Human diversity in AI systems	In this session we talk about the importance of understanding the global market, value pluralism, why human diversity is important and how AI is a form of collective intelligence.
11	11/6/2024	Generative Models I: Large Language Models	We will talk about what generative models are, how they work and what type of ethical and social issues they pose. We will also discuss the risks and potential of LLMs and tradeoffs between open-source and closed models.
12	11/13/2024	Generative Models II: AI Safety and Catastrophic Risk	In this session we will discuss potential catastrophic risk from AI. We talk about Artificial General Intelligence (AGI) and different schools of thought related to AI Safety.
13	11/20/2024	Regulation and Governance of AI	We learn about the history and current state of AI regulation, why it is important to follow and understand AI regulation in the US and elsewhere.
THANKSGIVING WEEK - NO CLASS			
14	12/4	The Status and Future of AI in the workplace	In the last class we will discuss the future of the job market and how AI will transform future societies.
15	12/11	Finals: Project Work	Students will present projects they have been working on throughout the semester. The goal of the project is to show how ethical consideration will make AI development more fair, beneficial and safe.

Course Schedule and Required Reading Materials by Week

Week 1: Introductions & Goal Setting - 8/28/2024

Description: In this session we will introduce the course and why it is important. Students will learn about the course structure, the objectives, meet each other, instructors and learn basic definitions, concepts and terminology.

Goals:
- Students should have a good understanding of what to expect in course
- Basic understanding of relevant terminology
- Answering any questions that students may have

Topics for Discussion:
- What are the expectations that students have? What would students like to learn?
- What are important concept definitions related to AI Ethics, e.g. what is the difference between morality and ethics?
- Why is Ethics relevant for tech?

Week 2: AI Ethics and Leadership - 9/4/2024

Description: We learn about different ethical theories relevant to the development and use of technology and AI. We discuss and define what ethical leadership means and think about important tools for future AI leaders to make ethical decisions for the deployment of technologies.

Goals:
- Learn about main schools of thoughts in AI Ethics
- We discuss various use cases when technical leaders were faced with ethical dilemmas
- Who are AI leaders or "thought leaders" today? And Why?

Topics for Discussion:
- Should AI make (moral) decisions?
- How can we guarantee human control over AI systems?
- AI leadership?
- What makes a decision an ethical one?
- How much should we take potential consequences into account when making an ethical choice?

Reading Materials:
- Van De Pol, Ibo and Royakkers, Lamber "Normative Ethics" in Ethics, Technology and Engineering, Wiley-Blackwell, 2011, **chapter on normative ethics, pages 65-108** https://cdn.prexams.com/6229/BOOK.pdf
- BBC Newsnight: "The trolley problem and ethics of driverless cars" – https://www.youtube.com/watch?v=FypPSJfCRFk (5 minutes)
- Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. "The moral machine experiment." *Nature* 563, no. 7729 (2018): 59-64, https://core.ac.uk/download/pdf/231922494.pdf & https://www.moralmachine.net/
- Optional: Liautaud, Susan, "How to make ethical decisions", Stanford University Human-Centered Artificial Intelligence (HAI), Directors' Conversation, 2021, https://www.youtube.com/watch?v=alzBQ24hD6s
- Optional: Times Magazine, Time 100/AI in 2023, https://time.com/collection/time100-ai/
- Optional: Gunkel, D. J., and J. Bryson. "Introduction to the special issue on machine morality: the machine as moral agent and patient. PhilosTechnol 27 (1): 5–8." (2014). https://link.springer.com/article/10.1007/s13347-014-0151-1

Guest Speaker: Amber Ross, Associate Professor, Department of Philosophy, University of Florida

Week 3: Artificial Intelligence and the Problem of Unintended Consequences - 9/11/2024

Description: We started to think about AI Ethics in recent years because we all experienced in one way or another how AI has changed our lives. In this class we talk about unintended consequences, harms and risks from AI systems and what technology leaders should think about.

Goals:
- Understand what Artificial Intelligence is, and have basic knowledge of its history
- Be able to critically think about ethical and social consequences of AI
- Learn about different types of types of harms

Topics for Discussion:
- AI, AGI or ML?
- Who is an AI Ethicist?

Reading Materials:
- Wiener, Norbert. "Some Moral and Technical Consequences of Automation (1960)." (2021), https://www.cs.umd.edu/users/gasarch/BLOGPAPERS/moral.pdf
- Koehrsen, Will. "How to Mind Goodhart's Law and Avoid Unintended Consequences | Built In," October 19, 2021. https://builtin.com/data-science/goodharts-law. (900 words, 4 min)
- Hassabis, Demis "Unreasonably Effective AI", Google DeepMind podcast, (August 2024, https://www.youtube.com/watch?v=pZybROKrj2Q (52 min)
- *Shelby, Renee, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari et al. "Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction." In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, pp. 723-741. 2023, https://dl.acm.org/doi/pdf/10.1145/3600211.3604673*
- Optional: Frank, Lily Eva, and Michał Klincewicz. "Uses and Abuses of AI Ethics." In *Handbook on the Ethics of Artificial Intelligence*, pp. 205-217. Edward Elgar Publishing, 2024, https://philarchive.org/archive/FRAUAA-4
- Optional: maybe too easy: Andrew Ng, "Machine Learning vs. AI: Differences, Uses, and Benefits" https://www.coursera.org/articles/machine-learning-vs-ai (1200 words, 5 min + 8 min video))
  Optional: Google Cloud, "What is Artificial Intelligence (AI)?, https://cloud.google.com/learn/what-is-artificial-intelligence

Guest Speaker: (tentatively) Renee Shelby, Google Inc.

Week 4: Bias: Human vs. Algorithmic - 9/18/2024

Description: We learn about various types of biases. Why are AI systems biased and what can we do about it? How can we measure bias? How can we mitigate biases and improve fairness?

Goals:
- Have a good understanding of different types of biases.
- Understand ethical considerations of predictive algorithms.
- Learn about algorithmic auditing

Topics for Discussion:
- Are biases always bad?
- What is fair?

Reading Materials:

- Buolamwini, Joy, "Unmasking AI Book Launch," All Tech is Human, YouTube, Oct 31, 2023. (52 min) https://www.youtube.com/live/_DWx5XQr0xI?si=afG4ZfKXnXjgQ7Hy
- Yong, Ed. "A Popular Algorithm Is No Better at Predicting Crimes Than Random People," *The Atlantic*, January 17, 2018. (1200 words, 5 min) https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/.
- Larson, Jeff, Mattu, Surya, Kirchner, Lauren and Angwin, Julia "How we Analyzed the COMPAS Recidivism Algorithm", ProPublica, 2016 https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
- *Zink, Anna, Ziad Obermeyer, and Emma Pierson. "Race adjustments in clinical algorithms can help correct for racial disparities in data quality." Proceedings of the National Academy of Sciences 121, no. 34 (2024) https://www.pnas.org/doi/full/10.1073/pnas.2402267121*
- Optional: Friedman, Scott, Sonja Schmer-Galunder, Anthony Chen, and Jeffrey Rye. "Relating word embedding gender biases to gender gaps: A cross-cultural analysis." In *Proceedings of the first workshop on gender bias in natural language processing*, pp. 18-24. 2019 https://aclanthology.org/W19-3803.pdf
- Optional: Landers, Richard N., and Tara S. Behrend. "Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models." American Psychologist 78, no. 1 (2023): 36, https://psycnet.apa.org/fulltext/2022-30899-001.html?ref=salesforce-research
- Optional: Movie recommendation: Coded Bias, Directed by Shalini Kantayya, Netflix, 2020. An intimate view into the lives of key scholars and activists fighting against AI bias and discrimination.

Guest Speaker: Emma Pierson, University of California Berkeley

Week 5: What is data? - 9/25/2024

Description: In this session students will learn about the role of data in AI with a focus on ethical considerations, e.g. representation, inclusion, bias, consent, processing, etc. We will discuss data annotation processes and concerns related to public information, as well as trade-offs between data quality and quantity.

Goals:
- Understand the genealogy of data, and ethical consideration in collecting publicly available data
- Understand what data annotation is and how can work towards data annotations responsibly

Topics for Discussion:
- What types of data should be used in AI models?
- Who owns my data? How do users perceive data ownership?

Reading Materials:

- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92. https://dl.acm.org/doi/pdf/10.1145/3458723
- ACM short video: https://www.youtube.com/watch?v=R7s7_T4yXak
- Raghavan, B. and Schneier, B. "Seeing Like a Data Structure": https://www.belfercenter.org/publication/seeing-data-structure
- Optional: Schmer-Galunder, Sonja, Ruta Wheelock, Scott Friedman, Alyssa Chvasta, Zaria Jalan, and Emily Saltz. "Annotator in the Loop: A Case Study of In-Depth Rater Engagement to Create a Bridging Benchmark Dataset." https://doi.org/10.48550/arXiv.2408.00880 (2024).

- Optional: *Denton, Emily, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. "On the genealogy of machine learning datasets: A critical history of ImageNet." Big Data & Society 8, no. 2 (2021): 20539517211035955.* *https://journals.sagepub.com/doi/full/10.1177/20539517211035955?s=09*
- Optional: Boyer, Pascal. "Ownership Psychology as a Cognitive Adaptation: A Minimalist Model." Behavioral and Brain Sciences, October 18, 2022, 1–35.

Guest Speaker: Andrew Smart, Google Inc.

Week 6: Values and Norms - 10/2/2024

Description: We take a deeper dive into norms, values and virtues. We learn about how to bring human values to AI systems, risks and the value alignment problem.

Goals:
- Students should have basic understanding of values, norms, codes and standards in AI Ethics
- Understand the value alignment problem

Topics for Discussion:
- What and whose values should AI align with?

Reading Material:

- The Alignment Problem: Machine Learning and Human Values with Brian Christian, 2022. https://www.youtube.com/watch?v=z6atNBhItBs. (1 hour 13 min)
- Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen et al. "Constitutional ai: Harmlessness from AI feedback." arXiv preprint arXiv:2212.08073 (2022). https://arxiv.org/pdf/2212.08073
- Abernethy, Jacob, Francois Candelon, Theodoros Evgeniou, Abhishek Gupta, and Yves Lostanlen. "Bring Human Values to AI." Harvard Business Review 103, no. 3-4 (2024): 59-68, https://hbr.org/2024/03/bring-human-values-to-ai
- Optional: Boddington, Paula. "Normative Modes: Codes and Standards." In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das. Oxford University Press, 2020. https://doi.org/10.1093/oxfordhb/9780190067397.013.7 (may need to log in to library)
- Optional: Johnson, G. M. (2023). Are algorithms value-free?: Feminist theoretical virtues in machine learning. Journal of Moral Philosophy, 1(aop), 1-35. https://www.gmjohnson.com/uploads/5/2/5/1/52514005/are_algorithms_value_free_.pdf)

Guest Speaker: Adam Russell, Information Science Institute, University of California, Los Angeles

Week 7: Privacy and Transparency - 10/2/2024

Description: We learn about the ethical questions related to privacy and privacy protection.

Goals:
- Learn about privacy risks and problems related to systematic digital surveillance
- Increase awareness of invasions of privacy

Topics for Discussion:
- Who owns the data that is used for training generative models?
- Do we need to re-think privacy?

Reading Materials:

- Caltrider, Jen, Ryko, Misha and McDonald, Zoe, "It's official: Cars Are the Worst Product Category We have Ever Reviewed for Privacy." Mozilla Foundation Review, 2023 https://foundation.mozilla.org/en/privacynotincluded/articles/its-official-cars-are-the-worst-product-category-we-have-ever-reviewed-for-privacy/
- Lyngaas, Sean, "Mental health startup exposes the personal data of more than 3 million people," CNN, March 10, 2023 (500 words, 2 min) https://www.cnn.com/2023/03/10/politics/cerebral-mental-health-privacy-data-exposure/index.html
- Véliz, Carissa, "Privacy is Power", Harvard Carr Center for Human Rights Policy, 2021, https://www.youtube.com/watch?v=_zMwFfw9rjU
- Jon Berkely, "What Machines Can Tell from you Face", Economist, 2017, https://drive.google.com/file/d/1eDvkTTgbaf4hHuhMXFC1Oj6kaDP9ZtD8/view
- Brittain, Blake, "OpenAI, Microsoft hit with new US consumer privacy class action," Reuters, September 6, 2023. (400 words, 2 min) https://www.reuters.com/legal/litigation/openai-microsoft-hit-with-new-us-consumer-privacy-class-action-2023-09-06/
- Optional: "Joan is Awful," Black Mirror TV Episode, Netflix, Season 6, 2023. (subscription required, free trials available)
- Optional: Podcast: https://www.nytimes.com/2023/10/06/podcasts/hard-fork-google-trial.html? Section about "How to Wear A.I."

Guest Speaker: TBD

Week 8: Truth and Misinformation - 10/16/2024

Description: We will talk about mis- and disinformation, and the importance of truth and trust. We learn about the use of algorithms that are optimized to influence people, and technology to make information more trustworthy and safe.

Goals:
- Understand the difference between mis- and disinformation
- Learn about algorithm optimization
- Learn about technological solutions to decrease manipulation and increase trust

Topics for Discussion:
- Why should companies care about misinformation?
- Is targeting users always bad?

Reading Materials:

- Allen, Jeff. "Misinformation Amplification Analysis and Tracking Dashboard." Integrity Institute, October 13, 2022. (3400, 15 min) https://integrityinstitute.org/blog/misinformation-amplification-tracking-dashboard.

- PBS, Fake: Searching for Truth in the Age of Misinformation, 2020. (56 min) https://www.pbs.org/video/fake-searching-for-truth-in-the-age-of-misinformation-jczifr/
- Satariano, Adam, and Paul Mozur. "The People Onscreen Are Fake. The Disinformation Is Real." The New York Times, February 7, 2023. https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html.
- Ressa, Maria, "Keynote Address by Maria Ressa at the UNESCO Global Conference "Internet for Trust," UNESCO, 2023. (18 min video) https://www.youtube.com/watch?v=E6-qcA5e-a4
- llyn, Bobby, "'New York Times' considers legal action against OpenAI as copyright tensions swirl," NPR, August 16, 2023. https://www.npr.org/2023/08/16/1194202562/new-york-times-considers-legal-action-against-openai-as-copyright-tensions-swirl (1200 words, 5 min)
- Optional: Farid, Hany. "Creating, Using, Misusing, and Detecting Deep Fakes." Journal of Online Trust and Safety 1, no. 4 (September 20, 2022). (13,000 words, 60 min) https://tsjournal.org/index.php/jots/article/view/56/36
- Optional: DiResta, Renee, "The Invisible Rulers Turning Lies Into Reality", Talk at the Commonwealth Club, June 2024, https://www.youtube.com/watch?v=Ad2gjdN_k5Y
- Optional: Stray, Johnathan, Ravi Iyer, and Helena Puig Larrauri. "The Algorithmic Management of Polarization and Violence on Social Media." The Knight First Amendment Institute, Aug 22, 2023. (11,000 words, ~45 min) https://s3.amazonaws.com/kfai-documents/documents/fced667a59/Algorithmic-Management.pdf
- Optional: Atlantic Council, "Narrative Warfare": https://www.atlanticcouncil.org/in-depth-research-reports/report/narrative-warfare/

Guest Speaker: TBD

Week 9: Society: The Politics of Algorithms - 10/23/2024

Description: In this class we will learn about the global context for the use and deployment of AI, including, geopolitical impacts and national security risks.

Goals:
- Geopolitical implications of AI
- Understand the global context (or production cycle) of AI
- Be able to think critically about national security risk

Topics for Discussion:
- Is the innovation competition between the U.S. and China a race to the bottom?

Reading Materials:

- Crawford, Kate, and Vladan Joler. "Anatomy of an AI System." Anatomy of an AI System (2018). https://anatomyof.ai/ and NEW! https://calculatingempires.net/
- Winner, Langdon. "Do artifacts have politics?." In Computer Ethics, pp. 177-192. Routledge, 2017. https://eclass.uoa.gr/modules/document/file.php/PHS541/Winner%20--%20Do%20Artifacts%20Have%20Politics.pdf
- World Economic Forum, "The Geopolitical Impacts of AI", https://intelligence.weforum.org/topics/a1Gb0000000pTDREA2/key-issues/a1Gb00000017LCAEA2

- Coeckelbergh, Mark. The political philosophy of AI: an introduction. John Wiley & Sons, 2022, video summary: https://www.youtube.com/watch?v=jqpetqc5tig

Guest Speaker: TBD

Week 10: Culture: Human diversity in AI systems - 10/30/2024

Description: The influence and distribution of AI does not stop at national borders. Still, different people speak different languages and have different norms and values. In this session we talk about the importance of understanding the global market, value pluralism, why human diversity is important and how AI is a form of collective intelligence.

Goals:
- Learn about different ways of seeing and thinking
- Understand what a WEIRD bias is and why we find it so often in AI systems

Topics for Discussion:
- Who is "the Human" when we talk about human-compatible AI systems?

Reading Materials:

- Turk, Victoria, How AI reduces the world to stereotypes, Rest of the World, 2023, https://restofworld.org/2023/ai-image-stereotypes/
- Atari, Mohammad, Mona J. Xue, Peter S. Park, Damián Blasi, and Joseph Henrich. "Which humans?." (2023), https://hmpa.hms.harvard.edu/sites/projects.iq.harvard.edu/files/culture_cognition_coevol_lab/files/which_humans_09222023.pdf
- Rudschies, Catharina, Ingrid Schneider, and Judith Simon. "Value pluralism in the AI ethics debate–Different actors, different priorities." The International Review of Information Ethics 29 (2020), https://informationethics.ca/index.php/irie/article/view/419/396
- Carl Miller, "Can Taiwan reboot democracy?", BBC documentary, 2021, https://www.youtube.com/watch?v=VbCZvU7i7VY&t=1s

Guest Speaker: Scott Friedman, Principal Research Scientist, Smart Information Flow Technologies

Week 11: Regulation and Governance of AI - 11/20/2024

Description: We learn about the history and current state of AI regulation, why it is important to follow and understand AI regulation in the US and elsewhere.

Goals:
- Go over use cases that lead to regulation of AI
- Understand current regulatory frameworks in the EU and the U.S.

Topics for Discussion:
- How do European laws like the GDPR, Digital Services Act and the EU AI Act shape the global tech landscape?
- What can be done from inside a company?

Reading Materials:

- Bryson, Joanna J. "The artificial intelligence of the ethics of artificial intelligence." The Oxford handbook of ethics of AI (2020): 1-25. May require library login
- Hoffman, Mia, "The EU AI Act: A Primer," Center for Security and Emerging Technology, September 26, 2023. https://cset.georgetown.edu/article/the-eu-ai-act-a-primer/
- NIST's Responsibilities Under Executive Order 14110 On Safe, Secure and Trustworthy AI," Nov 9, 2023. (25 min video) https://youtu.be/l2KDgSFw8Z4?si=B-Etm5KlSVP98AW5
- Optional: Ovadya, Aviv, "Democratic Policy Development using Collective Dialogues and AI," Arxiv, Nov 3, 2023. https://arxiv.org/pdf/2311.02242.pdf
- Optional: NIST, "Participating in White House Summit on Standards for Critical and Emerging Technology", https://www.nist.gov/news-events/news/2024/07/nist-participates-white-house-summit-standards-critical-and-emerging (August, 2024)
- Optional: Husovec, Martin, "How Does the EU's Digital Services Act Regulate Content Moderation? And Will it Work?" Stanford Cyber Policy Center, Apr 11, 2023. (1 hour) https://www.youtube.com/watch?v=np05wM3h2mc)
- Optional: Perrigo, Billy. "How Frances Haugen's Team Forced a Facebook Reckoning." Time, October 7, 2021. (3700 words, 30 min) https://time.com/6104899/facebook-reckoning-frances-haugen

Guest Speaker: Joanna Smolinka, Deputy Head of the EU Office in San Francisco

Week 12: Generative Models: Large Language Models - 11/6/2024

Description: We will talk about what generative models are, how they work and what type of ethical and social issues they pose. We will also discuss the risks and potential of LLMs and tradeoffs between open-source and closed models.

Goals:

- Understand basics of LLMs, their limitations and ethical considerations related to open and closed generative models
- Learn about Theory of Mind and be able to critically think about outputs of LLMs

Topics for Discussion:

- Why do we anthropomorphize LLMs?
- Are LLMs intelligent?

Reading Materials:

- Roose, Kevin. "Bing's A.I. Chat: 'I Want to Be Alive. 😈,'" February 16, 2023. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html
- Chiang, Ted, "ChatGPT is a blurry JPEG of the Web," The New Yorker, Feb 9, 2023. https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web (3300 words, 15 minutes)
- Kosinski, M. (2023). Theory of mind might have spontaneously emerged in large language models. arXiv preprint arXiv:2302.02083, https://arxiv.org/pdf/2302.02083
- Solaiman, Irene. "The gradient of generative AI release: Methods and considerations." In Proceedings of the 2023 ACM conference on fairness, accountability, and transparency, pp. 111-122. 2023,

https://dl.acm.org/doi/pdf/10.1145/3593013.3593981?casa_token=uv0Z29e-1tMAAAAA:PTOCI15AeYEn47MFTcaRheVLOSu-bD91eBG9TOayN7lSI1jJHEc1kuqqQkbS0Zoj2tzXglR-Mee4
- Lawton, George, "Attributes of Open vs. Closed AI Explained, Tech Target, https://www.techtarget.com/searchenterpriseai/feature/Attributes-of-open-vs-closed-AI-explained
- Optional: Podcast This American Life, "Greetings, People of Earth", https://www.thisamericanlife.org/803/greetings-people-of-earth
- Optional: Technical DeepDive including history of LLMs: Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., ... & Azam, S. (2024). A review on large Language Models: Architectures, applications, taxonomies, open issues and challenges. IEEE Access: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10433480)

Guest Speaker: TBD

Week 13: Generative Models: AI Safety and Catastrophic Risk - 11/13/2024

Description: Despite global consensus on the necessity of developing safe AI, there are currently no systems in place that address deliberate or unintentional harm or that ensure accountability during the development process of such systems. In this session we will discuss potential catastrophic risk from AI. We talk about Artificial General Intelligence (AGI) and different schools of thought related to AI Safety.

Goals: Understand concerns about AI safety and be aware of safety concerns when building and deploying AI systems in markets.

Topics for Discussion:
- Why are they building AI while also warning of the potential for catastrophic harm?
- How can we ensure AI is safe?
- Doomers vs. Accelerationists?

Reading Materials:

- [Warning: simulated violent imagery] "Slaughterbots," Space Digital, 2017. https://www.youtube.com/watch?v=9CO6M2HsoIA. (8 minutes)
- Robinson-Early, Nick, "AI's 'Oppenheimer moment': autonomous weapons enter the battlefield", The Guardian, July 2024, https://www.theguardian.com/technology/article/2024/jul/14/ais-oppenheimer-moment-autonomous-weapons-enter-the-battlefield
- Miles, Robert. "10 Reasons to Ignore AI Safety," *Robert Miles AI Safety, YouTube*. https://www.youtube.com/watch?v=9i1WlcCudpU (17 minutes)
- Russell, Stuart. "If We Succeed." Daedalus: AI and Society, Spring 2022. https://www.amacad.org/publication/if-we-succee
- White House Memo, Ensuring Safe, Secure and Trustworthy AI, https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf
- Spiers, E. (2023). Tech Overlord's Horrifying, Silly Vision for Who Should Rule the World. The New York Times. https://www.nytimes.com/2023/10/28/opinion/marc-andreessen-manifesto-techno-optimism.html
- Kasirzadeh, A. (2024). Two Types of AI Existential Risk: Decisive and Accumulative. arXiv preprint, https://arxiv.org/pdf/2401.07836
- Optional: Coulter, Martin and Paul Sandle, "At UK's AI Summit developers and govts agree on testing to help manage risks," Reuters, November 2, 2023. (740 words, 3 min) https://www.reuters.com/world/uk/uk-pm-sunak-lead-ai-summit-talks-before-musk-meeting-2023-11-02/

- Optional: Elton, Dan, "What AI ethicists get wrong – and right – about AI safety," More is Different, Mar 19, 2023. (2400 words, 10 min) https://moreisdifferent.substack.com/p/what-ai-ethicists-get-wrong-and-right
- Optional: Roose, Kevin and Newton, Casey, "Dario Amodei, CEO of Anthropic, on the Paradoxes of AI Safety," Hard Fork podcast, July 21, 2023. https://www.nytimes.com/2023/07/21/podcasts/dario-amodei-ceo-of-anthropic-on-the-paradoxes-of-ai-safety-and-netflixs-deep-fake-love.html (first 57 min)

Guest Speaker: Atoosa Kazirzadeh, Schmidt Science Fellow, Professor of Philosophy at Carnegie Mellon University

Week 14: The Status and Future of AI - 12/4/2024

Description: In the last class we will discuss the future of the job market and how AI will transform future societies.

Goals:
- Understand how AI is transforming society and markets
- Be able to think critically technological development and how they may impact next generations

Topics for Discussion:
- Will machines replace us?
- What types of skills will not go away any time soon?

Reading Materials:

- World Economic Forum, "The Future of Jobs Report", https://www.weforum.org/publications/the-future-of-jobs-report-2023/
- Graeber, D. "On the Phenomenon of Bullshit Jobs: A Work Rant. https://web.archive.org/web/20180807024932/http://strikemag.org/bullshit-jobs/
- Behrend, Tara S., Daniel M. Ravid, and Cort W. Rudolph. "Technology and the changing nature of work." Journal of Vocational Behavior (2024): 104028, https://www.sciencedirect.com/science/article/pii/S0001879124000691?casa_token=gUXe962I8ZwAAAAA:5tI4PepirASBVAzrzYM_gwO4BP7AnPM1oubtARPbzfzQMwtwit-CB8_8n_VhTiOAZSqjlps21A
- Optional: Yuval Noah Harari, Lecture on AI and the future of humanity, https://www.youtube.com/watch?v=LWiM-LuRe6w

Guest Speaker: Tara Behrend, Michigan State University

Week 15: Final Project Presentations

Description: For our final presentations, we will convene in-person at Malachowsky Hall. Throughout the semester, students will work in teams of 4-6 on one of our class topics. The culmination of this effort will be a paper exploring the ethical, social, and legal implications of a socio-technical AI system. Each team will present their findings to the entire class. When selecting your topic, consider both its innovative potential to address societal challenges and its ethical implications, including long-term and unintended consequences. Choose a subject that genuinely interests you and your team, covering one of the 13 topics we will go over in class. Presentations will be allotted 10 minutes, with an additional 3 minutes reserved for questions from the audience.

Evaluation of Grades
The total points for each assignment are weighted to represent the given percentage.

Assignment	Total Points	Percentage of Final Grade
Attendance and Participation	100	20%
Quizzes	100	20%
In-class presentations (news)	100	15%
Final Paper	100	20%
Final Presentation	100	25%
Total	500	100%

Cumulative assignment score will be computed as a weighted average of the individual assignment scores (each on the 100-point scale) using the above weights. Letter grades will be obtained by thresholding as follows:

Grading Policy

Percent	Grade	Grade Points
93.4 - 100	A	4.00
90.0 - 93.3	A-	3.67
86.7 - 89.9	B+	3.33
83.4 - 86.6	B	3.00
80.0 - 83.3	B-	2.67
76.7 - 79.9	C+	2.33
73.4 - 76.6	C	2.00
70.0 - 73.3	C-	1.67
66.7 - 69.9	D+	1.33
63.4 - 66.6	D	1.00
60.0 - 63.3	D-	0.67
0 - 59.9	E	0.00

More information on UF grading policy may be found at:
http://gradcatalog.ufl.edu/content.php?catoid=10&navoid=2020#grades

Students Requiring Accommodations
Students with disabilities who experience learning barriers and would like to request academic accommodations should connect with the disability Resource Center by visiting https://disability.ufl.edu/students/get-started/. It is important for students to share their accommodation letter with their instructor and discuss their access needs, as early as possible in the semester.

Course Evaluation
Students are expected to provide professional and respectful feedback on the quality of instruction in this course by completing course evaluations online via GatorEvals. Guidance on how to give feedback in a professional and respectful manner is available at https://gatorevals.aa.ufl.edu/students/. Students will be notified when the evaluation period opens, and can complete evaluations through the email they receive from GatorEvals, in their Canvas course menu under GatorEvals, or via https://ufl.bluera.com/ufl/. Summaries of course evaluation results are available to students at https://gatorevals.aa.ufl.edu/public-results/.

In-Class Recording

Students are allowed to record video or audio of class lectures. However, the purposes for which these recordings may be used are strictly controlled. The only allowable purposes are (1) for personal educational use, (2) in connection with a complaint to the university, or (3) as evidence in, or in preparation for, a criminal or civil proceeding. All other purposes are prohibited. Specifically, students may not publish recorded lectures without the written consent of the instructor.
A "class lecture" is an educational presentation intended to inform or teach enrolled students about a particular subject, including any instructor-led discussions that form part of the presentation, and delivered by any instructor hired or appointed by the University, or by a guest instructor, as part of a University of Florida course. A class lecture does not include lab sessions, student presentations, clinical presentations such as patient history, academic exercises involving solely student participation, assessments (quizzes, tests, exams), field trips, private conversations between students in the class or between a student and the faculty or lecturer during a class session. Publication without permission of the instructor is prohibited. To "publish" means to share, transmit, circulate, distribute, or provide access to a recording, regardless of format or medium, to another person (or persons), including but not limited to another student within the same class section. Additionally, a recording, or transcript of a recording, is considered published if it is posted on or uploaded to, in whole or in part, any media platform, including but not limited to social media, book, magazine, newspaper, leaflet, or third party note/tutoring services. A student who publishes a recording without written consent may be subject to a civil cause of action instituted by a person injured by the publication and/or discipline under UF Regulation 4.040 Student Honor Code and Student Conduct Code.

University Honesty Policy
UF students are bound by The Honor Pledge which states, "We, the members of the University of Florida community, pledge to hold ourselves and our peers to the highest standards of honor and integrity by abiding by the Honor Code. On all work submitted for credit by students at the University of Florida, the following pledge is either required or implied: "On my honor, I have neither given nor received unauthorized aid in doing this assignment." The Honor Code (https://sccr.dso.ufl.edu/process/student-conduct-code/) specifies a number of behaviors that are in violation of this code and the possible sanctions. Furthermore, you are obligated to report any condition that facilitates academic misconduct to appropriate personnel. If you have any questions or concerns, please consult with the instructor or TAs in this class.

Commitment to a Safe and Inclusive Learning Environment
The Herbert Wertheim College of Engineering values varied perspectives and lived experiences within our community and is committed to supporting the University's core values, including the elimination of discrimination. It is expected that every person in this class will treat one another with dignity and respect regardless of race, creed, color, religion, age, disability, sex, sexual orientation, gender identity and expression, marital status, national origin, political opinions or affiliations, genetic information, and veteran status.

If you feel like your performance in class is being impacted by discrimination or harassment of any kind, please contact your instructor or any of the following:
• Your academic advisor or Graduate Coordinator
• HWCOE Human Resources, 352-392-0904, student-support-hr@eng.ufl.edu
• Pam Dickrell, Associate Dean of Student Affairs, 352-392-2177, pld@ufl.edu
• Toshikazu Nishida, Associate Dean of Academic Affairs, 352-392-0943, nishida@eng.ufl.edu

Software Use
All faculty, staff, and students of the University are required and expected to obey the laws and legal agreements governing software use. Failure to do so can lead to monetary damages and/or criminal penalties for the individual violator. Because such violations are also against University policies and rules, disciplinary action will be taken as appropriate. We, the members of the University of Florida community, pledge to uphold ourselves and our peers to the highest standards of honesty and integrity.

Student Privacy
There are federal laws protecting your privacy with regards to grades earned in courses and on individual assignments. For more information, please see: https://registrar.ufl.edu/ferpa.html

Campus Resources:

Health and Wellness

U Matter, We Care:
Your well-being is important to the University of Florida. The U Matter, We Care initiative is committed to creating a culture of care on our campus by encouraging members of our community to look out for one another and to reach out for help if a member of our community is in need. If you or a friend is in distress, please contact umatter@ufl.edu so that the U Matter, We Care Team can reach out to the student in distress. A nighttime and weekend crisis counselor is available by phone at 352-392-1575. The U Matter, We Care Team can help connect students to the many other helping resources available including, but not limited to, Victim Advocates, Housing staff, and the Counseling and Wellness Center. Please remember that asking for help is a sign of strength. In case of emergency, call 9-1-1.

Counseling and Wellness Center: https://counseling.ufl.edu, and 392-1575; and the University Police Department: 392-1111 or 9-1-1 for emergencies.

Sexual Discrimination, Harassment, Assault, or Violence
If you or a friend has been subjected to sexual discrimination, sexual harassment, sexual assault, or violence contact the **Office of Title IX Compliance**, located at Yon Hall Room 427, 1908 Stadium Road, (352) 273-1094, title-ix@ufl.edu

Sexual Assault Recovery Services (SARS)
Student Health Care Center, 392-1161.

University Police Department at 392-1111 (or 9-1-1 for emergencies), or http://www.police.ufl.edu/.

Academic Resources

E-learning technical suppor*t*, 352-392-4357 (select option 2) or e-mail to Learning-support@ufl.edu. https://elearning.ufl.edu/.

Career Connections Center, Reitz Union, 392-1601. Career assistance and counseling; https://career.ufl.edu.

Library Support, http://cms.uflib.ufl.edu/ask. Various ways to receive assistance with respect to using the libraries or finding resources.

Teaching Center, Broward Hall, 392-2010 or 392-6420. General study skills and tutoring. https://teachingcenter.ufl.edu/.

Writing Studio, 302 Tigert Hall*,* 846-1138. Help brainstorming, formatting, and writing papers. https://writing.ufl.edu/writing-studio/.

Student Complaints Campus*:* https://sccr.dso.ufl.edu/policies/student-honor-code-student-conduct-code/;https://care.dso.ufl.edu.

On-Line Students Complaints*: https://distance.ufl.edu/getting-help/; https://distance.ufl.edu/state-authorization-status/#student-complaint.*