ETHICS AND EPISTEMOLOGY OF MACHINE LEARNING (PHI 6406)

Spring 2024 Graduate Seminar
Tuesdays 6:35-9:35pm (Dodd 181)

Zina Ward ▪ zward@fsu.edu

Course Description: Recent breakthroughs in machine learning (ML) have led to major – and perhaps accelerating – advances in the capabilities of AI systems. Generative AI tools like ChatGPT have raised public awareness of these advances and prompted increasing scrutiny about their risks. This seminar will explore recent philosophical work on machine learning, covering topics related to ethics, epistemology, philosophy of science, philosophy of mind, political philosophy, and philosophy of action. Our focus will be on the near-term issues that AI systems raise, rather than issues that are more speculative. Topics to be discussed include: the opacity of ML systems; what AI can contribute to science; how ML models bear on the debate between rationalism and empiricism; the capabilities of large language models; how autonomous technologies create responsibility gaps; and the problem of machine bias.

Office Hours: I encourage you to email me to set up a meeting if you want to talk about the course content, paper ideas, or anything else. You can also come to my office hours, which are Thursdays from 1-2:45pm in Dodd 283.

Assessment: Your grade for the seminar will be determined by:

(1) ***Weekly Responses* (25%)*:*** Each week, you must submit a thought about one or more of the readings (e.g., an objection, a counter-example, a connection to an earlier reading, a re-working of an argument, an illuminating case study, a question). This should be 1-2 paragraphs in length. Please come to seminar prepared to discuss your thoughts, as I may ask you to share your contribution. The hope is that these weekly responses encourage you to read the assigned papers in an engaged, critical mode.

Your response must be submitted to Canvas by the end of the day on Monday before each seminar meeting. I don't care if you get it in by midnight – just make sure it's there when I wake up on Tuesday! Your response grade will be reduced if you miss more than two.

(2) ***Paper(s)* (75%):** Given the seminar's exploratory nature, you'll have the option of writing one longer research paper (perhaps connected to your antecedent interests/work) or three shorter papers that engage solely with course readings. That is, your two options are:

(i) A standard term paper of roughly 5,000-6,000 words. You should read additional work (beyond what's been assigned for this course) in order to engage more deeply with the current literature on your chosen topic. I would be happy to help point you to existing work if there's something in particular you'd like to write about. The deadline for a research paper is April 30. You must submit a paper proposal on Canvas by Sunday, April 7. This proposal should include a tentative thesis of one or two sentences, plus a paragraph-length description of how you intend to argue for the thesis (or an outline). I will give you written feedback on your proposal.

I suggest that you choose this route if you are interested in developing research in the philosophy of ML, or if there is an ML-related topic that is adjacent to your other philosophical or academic interests.

(ii) Three short papers of 1,200-1,600 words. Your short papers need not go beyond the texts assigned for this course (i.e., no additional research is required). It would

be reasonable for a short paper to engage with just one or two papers we have discussed in class. The first short paper is due on Feb. 23, the second on March 22, and the third on April 30. You do not need to submit paper proposals, but I am happy to talk to you about your ideas.

I suggest you choose this route if you do not intend to develop a research project in the area of philosophy of ML (i.e., if this seminar is just exploratory for you) or if you expect to have a particularly busy end-of-semester.

Note that you must decide early in the semester if you wish to pursue option (i) or (ii). If you miss the deadline for the first short paper, you're committed to writing a long research paper. No matter which option you choose, I recommend that you start a list of potential paper topics as soon as the semester begins. For either option, feel free to further develop an idea that you first raised in a weekly response. You may also write about a topic related to the philosophy of AI/ML that we have not covered in the course.

A Note on Breadth Requirements: This course could plausibly fulfill multiple breadth requirements for Philosophy MA/PhD students. Your one long research paper, or two out of three short papers, should be on topic(s) that fall under the category for which you wish to receive credit for the course. If you have any questions about this, or about the acceptability of a potential topic, send me an email. The default assumption will be that the course falls under (i), since the seminar code is PHI6406 (Philosophy of Science), so make sure you contact me (and think ahead about your paper topic) if you want credit in another area. The distribution areas/groupings are:

(i) Epistemology, Metaphysics, or Philosophy of Science;
(ii) Philosophy of Action, Language, or Mind;
(iii) Ethics, Social or Political Philosophy.

Summary of Deadlines:

every Monday night	weekly responses
Friday, Feb. 23	short paper #1 due [option (ii)]
Friday, March 22	short paper #2 due [option (ii)]
Sunday, April 7	long paper proposal due [option (i)]
Tuesday, April 30	long paper, short paper #3 due

Auditing: I welcome auditors to attend the seminar (subject to space constraints). However, auditors must also submit weekly responses so that they can be included in discussion. If you are an auditor and do not yet have access to the Canvas page, let me know so that I can add you.

Readings: All readings are available on Canvas.

General Resources: I would strongly encourage you to try to stay up to date with AI/ML-related advances and news this semester. Here are some resources that I recommend. You may have other suggestions; we'll make a collective list on the first day of class.

- The NYTimes' Hard Fork podcast covers tech news weekly.
- MIT Technology Review has a lot of great, accessible resources. I recommend signing up for their daily email digest of tech news, The Download.
- For technical overviews of topics like explainability and fairness, check out the last few years of NeurIPS video tutorials (here are the tutorials from 2021, 2022, and 2023).
- Hi Phi Nation, the wonderful (general audience) philosophy podcast, dedicated their most recent season to "The Ethics of our Digital Futures."

Attendance and Illness: You shouldn't attend class in person if you are sick. I would appreciate it if you would let me know if you miss seminar because of illness or family emergency.

Use of Generative AI: We'll be talking quite a bit about generative AI in this class. I hope it goes without saying in a graduate seminar like this one that all work you submit must be your own. Although there are subtle and difficult issues here, as a general rule, I treat AI-based assistance like ChatGPT the same way I treat collaboration with other people. You are welcome to talk about your ideas and work with other people, both inside and outside the class, as well as with AI-based assistants. However, you should never include in your assignment anything that was not written directly by you without proper citation (h/t David Joyner for this formulation). Including anything you did not write in your papers without proper citation is a violation of the honor policy. If you intend to use ChatGPT in a particular way, but aren't sure whether it's compatible with the honor policy, please come talk to me – or better yet, bring up the question in seminar!

University Policies:

1. *University Attendance Policy*

Excused absences include documented illness, deaths in the family and other documented crises, call to active military duty or jury duty, religious holy days, and official University activities. These absences will be accommodated in a way that does not arbitrarily penalize students who have a valid excuse. Consideration will also be given to students whose dependent children experience serious illness.

2. *Academic Honor Policy*

The Florida State University Academic Honor Policy outlines the University's expectations for the integrity of students' academic work, the procedures for resolving alleged violations of those expectations, and the rights and responsibilities of students and faculty members throughout the process. Students are responsible for reading the Academic Honor Policy and for living up to their pledge to "...be honest and truthful and... [to] strive for personal and institutional integrity at Florida State University." (For more details see the FSU Academic Honor Policy and procedures for addressing alleged violations.)

3. *Americans With Disabilities Act*

Students with disabilities needing academic accommodation should:
(1) register with and provide documentation to the Office of Accessibility Services; and
(2) bring a letter to the instructor indicating the need for accommodation and what type.

Please note that instructors are not allowed to provide classroom accommodation to a student until appropriate verification from the Office of Accessibility Services has been provided. This syllabus and other class materials are available in alternative format upon request. For more information about services available to FSU students with disabilities, contact:

Office of Accessibility Services
874 Traditions Way
108 Student Services Building
Florida State University
Tallahassee, FL 32306-4167

(850) 644-9566 (voice)
(850) 644-8504 (TDD)
Email: oas@fsu.edu
https://dsst.fsu.edu/oas

Reading Schedule

Week 1 (Jan. 9) Introduction & Logistics

Zerilli et al. (2021), *A Citizen's Guide to Artificial Intelligence,* Chapter 1, "What is Artificial Intelligence?"
Nature editorial, "It's time to talk about the known risks of AI"
Optional: Piper (2022), "There are two factions working to prevent AI dangers. Here's why they're deeply divided."

Week 2 (Jan. 16) Background: Neural Networks and Reinforcement Learning

3Blue1Brown series on Neural Networks (3 videos; you can skip the 4th in the series)
LeCun et al. (2015), "Deep Learning"
Sutton & Barto (2018), *Reinforcement Learning*, Chapter 1, "Introduction"
Optional: Butlin (2023), "Reinforcement Learning and Artificial Agency"
Optional: 3Blue1Brown video, "But what is a convolution?"

Week 3 (Jan. 23) The Value of Explanation

Knight (2017), "The Dark Secret at the Heart of AI"
Vredenburgh (2021), "The Right to Explanation"
Colaner (2022), "Is Explainable Artificial Intelligence Intrinsically Valuable?"
Optional: Kafka (1925), *The Trial*
Optional: Marcus & Southen (2024), "Generative AI Has a Visual Plagiarism Problem"

Week 4 (Jan. 30) Transparency and Explainability

Creel (2020), "Transparency in Complex Computational Systems"
Zednik (2021), "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence"
Fleisher (2022), "Understanding, Idealization, and Explainable AI"
Optional: Hancox-Li (2020), "Robustness in Machine Learning: Does it Matter?"
Optional: Rudin (2019), "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead"

Week 5 (Feb. 6) Machine Learning in Science

Duede (2022), "Instruments, Agents, and Artificial Intelligence: Novel Epistemic Categories of Reliability"
Duede (2023), "Deep Learning Opacity in Scientific Discovery"
Ward (manuscript), "Natural Kinds and Machine Learning: The Case of Male and Female Brains"
Optional: Ezra Klein podcast (July 2023), "A.I. Could Solve Some of Humanity's Hardest Problems. It Already Has."
Optional: Lockhart (2023), "Because the Machine can Discriminate: How Machine Learning Serves and Transforms Biological Explanations of Human Difference"

Week 6 (Feb. 13) DNNs: Rationalism vs. Empiricism

Buckner (2023), *From Deep Learning to Rational Machines*
Chapter 1, "Moderate Empiricism and Machine Learning"

Chapter 2, “What is Deep Learning, and How Should We Evaluate its Potential?”
Chapter 6, Section 6.6, “Interest and Innateness,” pp. 285-295

Week 7 (Feb. 20) DNNs: Perception and Imagination

Buckner (2023), *From Deep Learning to Rational Machines*
Chapter 3, “Perception”
Chapter 5, “Imagination”
Optional: Yiu et al. (2023), “Transmission Versus Truth, Imitation Versus Innovation: What Children Can Do That Large Language and Language-and-Vision Models Cannot (Yet)”
Optional: Vong et al. (2024), “Grounded Language Acquisition through the Eyes and Ears of a Single Child”

Week 8 (Feb. 27) Large Language Models

Karpathy (2023), “Intro to Large Language Models” (1 hour video)
Wolfram (2023), “What Is ChatGPT Doing… and Why Does It Work?” pp. 1-105 (don’t worry, the “pages” are very short!)
Chiang (2023), “ChatGPT is a Blurry JPEG of the Web”
Bender et al. (2021), “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”
Optional: Schwitzgebel et al. (2023), “Creating a Large Language Model of a Philosopher”
Optional: Daily Nous forums, “Philosophers on GPT-3,” “Philosophers on Next-Generation Large Language Models”

Week 9 (March 5) Symbol Grounding in LLMs

Bender & Koller (2020), “Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data”
Chalmers (2023), “Does Thought Require Sensory Grounding? From Pure Thinkers to Large Language Models”
Molle & Millière (2023), “The Vector Grounding Problem”
Optional: Harnad (1990), “The Symbol Grounding Problem”
Optional: Pavlick (2023), “Symbols and Grounding in Large Language Models”

SPRING BREAK

Week 10 (March 19) Bias and Machine Learning

Sunstein (2022), “Governing by Algorithm? No Noise and (Potentially) Less Bias”
Barocas & Selbst (2016), “Big Data’s Disparate Impact,” Introduction and Section 1 (pp. 671-93 only)
Deery & Bailey (2022), “The Bias Dilemma: The Ethics of Algorithmic Bias in Natural-Language Processing”
Optional: Bolukbasi et al. (2016), “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embedding”

Week 11 (March 26) Fairness and the ML Impossibility Theorems

Angwin et al. (2016), “Machine Bias”
Corbett-Davies et al. (2016), “A computer program used for bail and sentencing decisions was labeled biased against blacks. It’s actually not that clear.” (figures online)

MIT Tech Review interactive explainer (2019), "Can you make AI fairer than a judge? Play our courtroom algorithm game"
Grant (2023), "Equalized Odds is a Requirement of Algorithmic Fairness"
Optional: Barocas et al. (2020), *Fairness and Machine Learning,* Chapter 3, "Classification"
Optional: Hedden (2021), "On Statistical Criteria of Algorithmic Fairness"

Week 12 (April 2) Limits of Formal Approaches to Fairness

Castro et al. (2023), "Egalitarian Machine Learning"
Green & Hu (2018), "The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning"
Hellman (2023), "Big Data and Compounding Injustice"
Optional: Hu (2020), "Direct Effects: How Should We Measure Racial Discrimination?"
Optional: Binns (2017), "Fairness in Machine Learning: Lessons from Political Philosophy"

Week 13 (April 9) Recommender Systems: Fairness and Autonomy

Art of the Problem, "How Recommender Systems Work" (8 min video)
Stinson (2022), "Algorithms are Not Neutral"
Buss & Westlund (2018), "Personal Autonomy" (SEP Entry)
Bartmann (2023), "Reasoning with Recommender Systems? Practical Reasoning, Digital Nudging, and Autonomy"
Optional: Susser et al. (2019), "Online Manipulation: Hidden Influences in a Digital World"
Optional: Burr et al. (2018), "An Analysis of the Interaction Between Intelligent Software Agents and Human Users"

Week 14 (April 16) Algorithmic Monoculture, Arbitrariness, and Discretion

Creel and Hellman (2022), "The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems"
Kleinberg and Raghavan (2021), "Algorithmic Monoculture and Social Welfare" (no need to read the formal stuff; focus on the introduction and conclusion)
Vredenburgh (2023), "AI and Bureaucratic Discretion"

Week 15 (April 23) Automated Technologies and Responsibility Gaps

Matthias (2004), "The responsibility gap: Ascribing responsibility for the actions of learning automata"
Dahaner (2022), "Tragic Choices and the Virtue of Techno-Responsibility Gaps"
Hindriks & Veluwenkamp (2023), "The Risks of Autonomous Machines: From Responsibility Gaps to Control Gaps"
Optional: Sparrow (2007), "Killer Robots"
Optional: Tigard (2021), "There Is No Techno-Responsibility Gap"