PHI3681: Ethics, Data, and Technology

3 | credits

NOTE: This course complies with all UF academic policies. For information on those policies and for resources for students, please see UF's "Academic Policies and Resources" web page.

I. General Information

Meeting days and times: T, H: 9:35-10:25
Class location: LIT 0109

Instructor(s):

Name: Cameron Buckner
Office Building/Number: FLO 330B
Phone: (352) 392-2084
Email: cameron.buckner@ufl.edu
Office Hours: T, R: 10:30-11:30 FLO 330B

Teaching Assistant(s):

Name: Jake Haun
Office Building/Number: FLO 200
Phone: (352) 392-2084
Email: jake.haun@ufl.edu
Office Hours: M/W 12:30-2:00

Name: jgarcia11@ufl.edu
Office Building/Number: FLO 316
Phone: (352) 392-2084
Email: jgarcia11@ufl.edu
Office Hours: T/H 11:00 AM – 12:30 PM

Course Description

This course will explore philosophical issues surrounding the development and deployment of emerging technologies, focusing especially on technological advances based on “deep learning” techniques in computer science. The primary focus will be on ethical and explanatory questions surrounding the use of these systems, which in just a few years have come to have pervasive effects in our daily lives—despite the fact that our understanding of their philosophical implications remains rudimentary. Questions we will explore are: in what senses are these systems biased, and when is their bias ethically problematic? Can we explain the workings of these vastly complex systems—containing billions of parameters and trained on Internet-scale datasets—in a way that answers to our existing scientific, legal, and ethical practices? Who is

responsible when these systems err? And finally: how can we adjust to the radical changes these systems are bringing to our social, political, and economic lives without losing our fundamental humanity, and can these systems be designed in a way to align with human values—as opposed to the pursuit of alien, machine objectives?

Prerequisites
None.

General Education Designation: none.

Course Materials

Materials will be available through the following means:
All readings will be posted on the course Canvas site
Materials Fee: N/A

II. Course Goals

Course Objectives
In this course we will:

- Review the basic architectures and training methods used in contemporary "deep" machine learning research, the differences between this approach and earlier methods in artificial intelligence, and their current applications to software systems in daily life.
- Provide a basic vocabulary for understanding and ethically evaluating these systems by drawing upon theories and concepts from philosophy.
- Review arguments for various positions in the ethical evaluation of cutting-edge technologies, represent them fairly and clearly, and evaluate their cogency.

Student Learning Outcomes
A student who successfully completes this course will be able to:

- Remember and understand technical terms to discuss artificial intelligence technologies and their ethical evaluation.
- Develop original arguments about contextually appropriate ethical frameworks to use across all aspects of AI, anticipate objections, and evaluate them in a conscientious manner.
- Speak and write persuasively on abstract and conceptually difficult issues at the intersection of philosophy and emerging technologies.

III. Graded Work

Graded Components
Weekly reading responses (10%): Responses to weekly readings will appear as quizzes on Canvas; they are normally due Tuesday at 11:59 the week they are assigned

Mid-term exam (20%): In-class mid-term exam with definition questions and short essays,

completed in a blue book in class

Final exam (20%): In-class final exam with definition questions and short essays, completed in a blue book in class

Group Case Studies (50%): Five group projects will be assigned and presented in discussion sections on Friday. They are graded via rubrics which will be available ahead of time. A full assignment submission consists of a group presentation presented in class, an issue brief submitted afterwards in Canvas, and a "discussant appraisal" which evaluates the presentation of another group.

TOTAL: 100%

Grading Scale

Letter Grade	Number Grade
A	100-92.5
A-	92.4-89.5
B+	89.4-86.5
B	86.4-82.5
B-	82.4-79.5
C+	79.4-76.5
C	76.4-72.5
C-	72.4-69.5
D+	69.4-66.5
D	66.4-62.5
D-	62.4-59.5
E	59.4-0

Note: A minimum grade of C is required to earn General Education credit.

IV. Calendar

Date	Topic	Readings and Work Due
Week 1 Jan 13	Basic Machine Learning #1: AI & DCNNs	Turing 1950: Computing Machinery and Intelligence Buckner 2018: Deep Learning: A Philosophical Introduction *Discussion Section Group Activity #1 Assigned:* Case Study on the Turing Test

Week 2 Jan 20	Generative AI and Philosophy of Mind	Buckner 2018: Empiricism without Magic-Transformational Abstraction in DCNNs
Week 3 Jan 27	The Black Box Problem	Rudin 2019: Stop Explaining Black Box Machine Learning Models… Buckner 2023: Black Boxes or Unflattering Mirrors? Comparative Bias in the Science of Machine Behaviour *Discussion Section Group Activity #1* Presentations Jan 30 Case Study on the Turing Test
Week 4 Feb 3	Basic Machine Learning #2: Transformers	Millière & Buckner 2022: A Philosophical Introduction to Language Models Pt. 1 Block 1981: Psychologism and Behaviorism *Discussion Section Group Activity #1* Group Issue Briefs **and** Individual Discussant Appraisals Due Wed Feb 4 11:59 PM Case Study on the Turing Test *Discussion Section Group Activity #2 Assigned:* Case Study on Bias in DNNs
Week 5 Feb 10	Interventionist Interpretability Methods	Millière & Buckner 2023: Interventionist Methods for Interpreting Deep Neural Networks Vredenburgh 2022 - "The Right to Explanation"
Week 6 Feb 17	Algorithmic Bias 1	Fazelpour and Danks 2021: Algorithmic Bias—Senses, Sources, Solutions Julia Angwin 2016 – "Machine Bias" ProPublica Corbett-Davies et al. 2016: "A computer algorithm used for bail…" *Discussion Section Group Activity #2* Presentations Fri Feb 20 Case Study on Bias in DNNs

Week 7 Feb 24 **Feb 26:** ***Exam #1***	Algorithmic Bias 2	Johnson 2020 – Algorithmic Bias-on the implicit biases of social technology Creel and Hellman 2022 – The Algorithmic Leviathan *Discussion Section Group Activity #2* Group Issue Briefs **and** Individual Discussant Appraisals Due Wed Feb 25 11:59 PM: Case Study on Bias in DNNs *Discussion Section Group Activity #3* Assigned*:* Case Study on Responsibility for AI Accidents
Week 8 Mar 3	Responsibility	Mathias 2004 – The Responsibility Gap Tigard 2021 – There is no Techno-responsibility Gap
Week 9 Mar 10	Trust	Simion & Kelp 2023 – Trustworthy Artificial Intelligence Hevelke & Nida-Rumelin 2015 – Responsibility for Crashes of Autonomous Vehicles
March 16-20	Spring Break	No class
Week 10 Mar 24	The Alignment Problem	Anthropic Team 2022 – Constitutional AI Gabriel 2020 – Artificial Intelligence, Values, and Alignment *Discussion Section Group Activity #3* Presentations Mar 27: Case Study on Responsibility for AI Accidents
Week 11 Mar 31	Algorithms, Echo Chambers, and Mental Health	Nguyen 2020 – "Echo Chambers and Epistemic Bubbles" Munroe 2024 – "Echo Chambers, Polarization, and 'Post-Truth'- In Search of a connection" Wells et al. "Facebook knows Instagram is Toxic for Teen Girls" (WSJ)

		Lewis, "Our minds can be hijacked" (The Guardian) *Discussion Section Group Activity #3* Group Issue Briefs **and** Individual Discussant Appraisals Due Wed Apr 1 11:59 PM: Case Study on Responsibility for AI accidents *Discussion Section Group Activity #4 Assigned:* Case study on AI Alignment
Week 12 Apr 7	Generative AI, Art, and Intellectual Property	Vlaad 2024 – A Portrait of the Artist as Young Algorithm Kieval 2024 – Artificial Achievement Goetze 2024 – AI art is theft *Discussion Section Group Activity #4*: Presentations Apr 10 Case study on AI Alignment
Week 13 Apr 14	The Future of Work	Danaher 2017 – Will life be worth living in a world without work? Belic 2024 – Institutions, Automation, and Legitimate Expectations *Discussion Section Group Activity #4* Group Issue Briefs **and** Individual Discussant Appraisals Due Wed Apr 15 11:59 PM: Case Study on AI Alignment *Discussion Section Group Activity #5 Assigned:* Case Study on Generative AI
Week 14 Apr 21	**Final Exam In Class Apr 21**	**Final Exam In Class Apr 21**
Week 15 Apr 28		*Discussion Section Group Activity #5*

		SUBMITTED VIA VIDEO ON CANVAS Group Issue Briefs **Due Apr 28** 11:59 PM: Case Study on Generative AI No discussant appraisals for Group Activity #5

V. Procedure for Conflict Resolution

Any classroom issues, disagreements or grade disputes should be discussed first between the instructor and the student. If the problem cannot be resolved, please contact Dr. Jon Rick (jrick@ufl.edu, (352) 293-1807). Be prepared to provide documentation of the problem, as well as all graded materials for the semester. Issues that cannot be resolved departmentally will be referred to the University Ombuds Office (http://www.ombuds.ufl.edu; 352-392-1308) or the Dean of Students Office (http://www.dso.ufl.edu; 352-392-1261).