PHIL GR9180: Topics in Moral Philosophy
Approaches to Applied Ethics—Philosophy of AI
Fall 2024, Monday 2:10-4pm
DRAFT SYLLABUS

1. Course Description

The philosophy of AI is an emerging field. Right now, AI is importantly concerned with LLMs. It is also concerned with the relation between natural and artificial intelligence. Researchers and public discourse ask whether AI can be "aligned" with values. Accordingly, key questions in Philosophy of AI relate to language, thought, and values.

The seminar starts with the widely debated alignment problem (Part I: weeks 1-4). Independent of how alignment works, it is by no means clear what the desired outcome is. People disagree about values. With which values should AI be aligned? At times the answer is: with "human values." Are "human values," in this context, the different and incompatible sets of values human beings have? If yes, what about the values that we should have?

AI researchers often focus on fairness, typically understood as the elimination of bias (Part II: weeks 5-6). We examine relevant notions of fairness and ask how fairness relates to other values, including and especially accuracy.

Next, we ask whether it makes sense to ascribe beliefs and intentions to AIs (Part III: weeks 7-11). Can AIs engage in reasoning, lie and be held responsible? We discuss "explainable AI," asking whether AI-outputs can be understood.

Finally, we examine questions about language as they apply to AIs (Part IV: weeks 12-14). How can LLMs cope with famously tricky components of language and thought, such as generics and implicature?

Part of the seminar are workshop sessions associated with the ValuesLab. Invited guest speakers come from a range of fields: research on AI in medical diagnostics, philosophy of AI, and computational linguistics. This seminar aims to contribute to dialogue between philosophers and AI developers, and to a shared vocabulary.

2. Outline of Readings and Topics

Part I: Introduction

As an introduction to the philosophy and ethics of AI, we cover questions such as "what is value alignment?" and "how can/do natural and artificial intelligences learn values?"

Week 1: Introduction Philosophy of AI and Ethics of AI

Readings: Peter Railton, "Ethical Learning: Natural and Artificial," in Matthew Liao (2020); Raphaël Millière, "The Alignment Problem in Context," arXiv (2023); Ruth Chang, "Value Alignment," in ed. David Edmonds, AI Morality (OUP 2024/forthcoming);
Two introductory videos and a historical sketch:

<https://www.youtube.com/watch?v=zjkBMFhNj_g> <https://www.youtube.com/watch?v=ISkAkiAkK7A>
Bruce Buchanan, A (Very) Brief History of Artificial Intelligence

Part II: Alignment and Value Pluralism
Debates about "alignment" of AI with "human values" try to recognize that there is no single set of values that people endorse. How should we respond to this? Should AI applications be tailored to the views users already have? We work toward an updated "map" of outlooks in normative ethics, as a starting point for questions about alignment, a plurality of evaluative outlooks, and questions about relativism and universal values.

Week 2: Familiar Options in Normative Ethics–Teleology versus Deontology, Consequentialism versus Kantian Ethics
Readings: John Rawls on teleology and deontology, "Classical Utilitarianism," Part I ch. 5, A Theory of Justice (1971); Jens Timmermann, "What's Wrong with 'Deontology'?" Proceedings of the Aristotelian Society, Volume 115 (2015); Mill, Utilitarianism II; Shelly Kagan, The Limits of Morality (OUP 1989), selection; Bernard Williams, selection from "Against Utilitarianism."

Week 3: Recent Additions to the "Map" of Options in Normative Ethics–Virtue Ethics based on Aristotle, Virtue Ethics based on Confucius, Buddhist ethics, Philosophy of Race, Cognitive Science, Ethics of Care
Readings: Julia Annas, "Ancient Eudaimonism and Modern Morality," in: Bobonich Cambridge Companion to Ancient Ethics (CUP, 2017); Amber Carpenter and Pierre-Julien Harter, "Introduction" in Crossing the Stream, Leaving the Cave (OUP 2024); Charles Mill "White Ignorance" In Shannon Sullivan & Nancy Tuana (eds.), Race and Epistemologies of Ignorance, State Univ of New York (2007), 11-38 and "Global White Ignorance," in: M. Gross & L. McGoey (Eds.), Routledge international handbook of ignorance studies, Routledge/Taylor & Francis Group (2015), 217-227; SEP "Implicit Bias" <https://plato.stanford.edu/entries/implicit-bias/>; Virginia Held, The Ethics of Care: Personal, Political, and Global. Oxford University Press, 2005, pp. 9-28; Carol Gilligan, In a Different Voice: Psychological Theory and Women's Development, Harvard University Press, 1982. Pp. 1-11; Annette Baier, "The Need for More than Justice," Canadian Journal of Philosophy 13 (1987).

Week 4: Alignment–with what?
Readings: Google's values for responsible AI: <https://www.tensorflow.org/responsible_ai>; Oliver Klingefjord, Ryan Lowe, Joe Edelman, "What are human values, and how do we align AI to them?" (2024) <https://arxiv.org/abs/2404.10636>; additional readings TBD.

Part III: Fairness and Accuracy
AI researchers seem to think of fairness as the opposite of bias. Fairness is discussed with a view to at least two kinds of tasks: descriptive assessments of what is the case, and predictive-probabilistic outputs. We discuss how fairness relates to accuracy in both contexts. We ask how the focus on fairness fares w.r.t. to values that are emphasized in a range of approaches in ethics, for example, feminist ethics of care and virtue ethics.

Week 5: Ethical AI, Fairness, and Accuracy
Readings: Richard Zemel et al, "Learning Fair Representations," <https://proceedings.mlr.press/v28/zemel13.html>; Talia Gillis, Bryce McLaughlin, Jann Spiess "On the Fairness of Machine-Assisted Human Decisions" (2023); Drago Plečko and Elias Bareinboim (2024), "Causal Fairness Analysis", Foundations and Trends in Machine Learning: Vol. 17, No. 3, pp 1–238. DOI: 10.1561/2200000106; Iason Gabriel, "Toward a Theory of Justice for Artificial Intelligence," 2022 by the American Academy of Arts & Sciences; Kosuke Imai and Zhichao Jiang, "Principal Fairness for Human and Algorithmic Decision-Making," (2022).

Week 6: ValuesLab Workshop on Fairness in AI (Date TBD)
Guest: Joshi Shalmali (CU Medical School) on her work on fairness in AI.
Commentator: TBD

Part III: Mental States?
It is tempting to speak of what an AI "believes," "intends," and so on. This is especially tempting when one interacts with robots. Whether these ascriptions are plausible depends, in part, on what we think beliefs, intentions, and so on, are. People often speak of AIs as a black box, signaling that it seems opaque how an AI arrives at its output. We ask what explainability amounts to, and how it relates to traditional ideas in philosophy of mind and epistemology.

Week 7: Belief-Ascriptions?
Readings: Murray Shanahan, "Talking about Large Language Models," (2023); Bernard Williams, "Deciding to Believe" (1973); Beba Cibralic and James Mattingly, "Machine agency and representation," (2021).
Additional readings: Timothy Williamson, selections from "Is Knowing a Mental State?" (1995) and "Knowledge First Epistemology," (2011); Harvey Lederman and Kyle Mahowald, "Are Language Models More Like Libraries or Like Librarians? Bibliotechnism, the Novel Reference Problem, and the Attitudes of LLMs," (2024).

Week 8: Robots
Readings: Shuai Yuan, Simon Coghlan, Reeve Lederman, Jenny Waycott, "Ethical Design of Social Robots in Aged Care: A Literature Review Using an Ethics of Care Perspective," International Journal of Social Robotics; Kiesler and Hinds, "Introduction to This Special Issue on Human–Robot Interaction," Human-Computer Interaction, 2004, Volume 19, pp. 1–8; Sung et al., 2007: "My Roomba Is Rambo: Intimate Home Appliances," In: Krumm, J., Abowd, G.D., Seneviratne, A., Strang, T. (eds) UbiComp 2007: Ubiquitous Computing. UbiComp 2007. Lecture Notes in Computer Science, vol. 4717. Springer, Berlin, Heidelberg.

Week 9: Lying and Hallucinating
Readings: James Mahon, "The Definition of Lying and Deception," SEP (2015) <https://plato.stanford.edu/entries/lying-definition/>; Andreas Stokke, "Lying and Asserting," Journal of Philosophy (2013); Levinstein, B. A., & Herrmann, D. A. (Accepted/In press). "Still no lie detector for language models: probing empirical and conceptual roadblocks," Philosophical Studies; James Campbell et al, "Localizing Lying in Llama: Understanding Instructed Dishonesty on True-False Questions Through Prompting, Probing, and Patching."

Week 10: ValuesLab Workshop
Guest: Phillip Koralus, Oxford, topic TBD
Commentator: TBD

Week 11: Explainable AI (XAI), Competence, and Reliability
Readings: The General Data Protection Regulation, issued by the European Union in 2018, speaks of a "right to explanation"; "Is explanation ex post rationalization?" <https://ykulbashian.medium.com/pragmatics-precedes-semantics-c8fb67a19ead>; Federica Russo, Eric Schliesser, Jean Wagemans, "Connecting Ethics and Epistemology of AI" (2023); Alan Goldman, "What is Justified Belief?" (selection, 1979); Ernest Sosa, A Virtue Epistemology (2007), "Lecture 2: A Virtue Epistemology"; Herman Cappelen and Joshua Dever, Making AI Intelligible: Philosophical Foundations (OUP 2021), "Introduction."

Part IV: Language
Much of the current debate about AI is about LLMs. This should put questions about language front and center in the discussion of ethical AI. We turn to two pervasive dimensions of language–and thought–that have recently received much attention: the semantics of generics and conversational implicature. Similarly, we look at the notion of linguistic communities and its role in philosophy of language and computational linguistics.

Week 12: ValuesLab Workshop
Guest: TBA
Commentator: TBD

Week 13: Generics and Implicature
Readings: Emily Allaway et al on generics as incomplete information; "Social and Political Aspects of Generic Language and Speech." McKeever and Sterken; Karen Lewis, "Gricean Pragmatics," Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, Edward Grefenstette, "The Goldilocks of Pragmatic Understanding: Fine-Tuning Strategy Matters for Implicature Resolution by LLMs," ArXiv (2022).
Background reading: Raphaël Millière & Cameron Buckner, "A Philosophical Introduction to Language Models – Part I: Continuity With Classic Debates," arXiv (2024).

3. Requirements

Readings: For each class, we have a range of readings. Students are expected to come to class having read the material carefully and ready to participate in discussion.

Presentations: Students are expected to participate in team presentations (5 minutes per team/paper). Presentations go along with a 1-page handout that clearly lays out key concepts and arguments. Students are expected to run the handout by the instructor at the latest 2 days prior to the presentation. They are invited to come to office hours to discuss the material.

E-credit writing requirements: Regular contributions to in-class presentations and team work in preparing presentations. Two 6-page papers or one 12-page paper.

Presentation/paper ratio: If a student is prepared to take on 3 presentations, each of which comes with a clear and detailed handout, the writing requirement is reduced to two 5-page papers, expanding on the handouts, or one 10-page paper. Due dates: TBD

R-credit requirements: Careful reading in preparation for class and participation in class.

4. Academic Integrity and Honor Code

Please consult Columbia University's policies on academic integrity as well as Columbia's honor code:

http://www.college.columbia.edu/academics/academicintegrity
https://www.college.columbia.edu/ccschonorcode
http://bulletin.columbia.edu/general-studies/undergraduates/academic-policies/academic-integrity-community-standards/

These policies explain Columbia University's academic regulations and how you can safeguard the integrity of your original work. Plagiarism and other forms of academic dishonesty are serious offenses. Please take the time to familiarize yourself with the details of what constitutes plagiarism and academic dishonesty. You are expected to confirm to these policies in your academic work. It is important that you understand that academic dishonesty can lead to disciplinary action, including failure in the course and suspension, or even expulsion, from the University.

5. Accommodations for Students with Disabilities

In order to receive disability-related academic accommodations, students must first be registered with Disability Services. More information on the Disability Services registration process is available online at <www.health.columbia.edu/ods>. Registered students must present an Accommodation Letter to the professor before an exam or other accommodations can be provided. Students who have, or think they may have, a disability are invited to contact Disability Services for a confidential discussion at (212) 854-2388 (Voice/TTY) or by email at <disability@columbia.edu>.