Philosophy 22962/32962

The Epistemology of Deep Learning

Spring 2021

TR 11:20–12:40

Instructors:	Anubav Vasudevan	Malte Willer
Email:	anubav@uchicago.edu	willer@uchicago.edu
Office hours (via Zoom):	R 1 –2:30	R 1– 2:30

Course Assistant Andrew Stone
Email: avstone@uchicago.edu

Course Description

Philosophers have long drawn inspiration for their views about the nature of human cognition, the structure of language, and the foundations of knowledge, from developments in the field of artificial intelligence. In recent years, the study of artificial intelligence has undergone a remarkable resurgence, in large part owing to the invention of so-called "deep" neural networks, which attempt to instantiate models of cognitive neurological development in a computational setting. Deep neural networks have been successfully deployed to perform a wide variety of machine learning tasks, including image recognition, natural language processing, financial fraud detection, social network filtering, drug discovery, and cancer diagnoses, to name just a few. While, at present, the ethical implications of these new and powerful systems are a topic of much philosophical scrutiny, the epistemological significance of deep learning has garnered significantly less attention.

In this course, we will attempt to understand and assess some of the bold epistemological claims that have been made on behalf of deep neural networks. To what extent can deep learning be represented within the framework of existing theories of statistical and causal inference, and to what extent does it represent a new epistemological paradigm? Are deep neural networks genuinely theory-neutral, as it is sometimes claimed, or does the underlying architecture of these systems encode substantive theoretical assumptions and biases? Without the aid of a background theory or statistical model, how can we, the users of a deep neural network, be in a position to trust the reliability of its predictions? In principle, are there any cognitive tasks with respect to which deep neural networks are incapable of outperforming human expertise? Do recent developments in artificial intelligence shed any new light on traditional philosophical questions about the capacity of machines to act intelligently, or the computational and mechanistic bases of human cognition?

Readings

All course readings will be available through the course's Canvas website.

Course Requirements

You are required to write three papers (4–6 pages). All papers are due by 5pm on the assigned day.

First paper	due April 20$^{\text{th}}$	worth 25%
Second paper	due May 11$^{\text{th}}$	worth 35%
Final Paper	due June 1$^{\text{st}}$	worth 40%

Paper topics will be uploaded to the Canvas site in advance of the due dates. Students may, if they wish, design their own paper topics after consultation with the instructor. Late papers will be docked a grade per day (e.g., B+ to B) unless you have received approval ahead of time. There will be no final exam.

Graduate students are required to write a substantial term paper (15+ pages, due June 1st). Topics must be approved ahead of time, concern an issue discussed in the class, and make significant use of course readings as well as additional recommended material.

Attendance of seminars is expected. Class participation will play a role in determining the final grade of borderline cases.

Format

All class meetings will be synchronous over Zoom. The default expectation for Zoom meetings is that you should be actively engaged, with camera on. If you require an exception, reach out to us and let us know.

Roadmap

The following schedule provides an overview over the topics that we will address during this semester as well as the assigned readings. Additional recommended readings will be announced in class. Readings may change as the semester goes on. Updated versions of this syllabus will be posted on Canvas as changes are made.

Date	**Topic**	**Readings**
Week 1	From GOFAI to NFAI	Turing, "Computing Machinery and Intelligence" Searle, "Is the Brain's Mind a Computer Program?" Churchland and Churchland, "Could a Machine Think?"
Week 2	Introduction to Neural Nets and Machine Learning	LeCun et al., "Deep Learning" Warner and Misra, "Understanding Neural Networks as Statistical Tools"
Week 3	Regularization and Overfitting	Darwiche, "Human-Level Intelligence or Animal-Like Abilities?" Pearl, *The Book of Why* (selections)
Week 4	NLP and Beyond	Otter et al., "A Survey of the Usages of Deep Learning for Natural Language Processing" Slonim et al., "An Autonomous Debating System"
Week 5	Connectionism	Fodor and Pylyshyn, "Connectionism and Cognitive Architecture: A Critical Analysis" Ramsey et al., "Connectionism, Eliminativism´and the Future of Folk Psychology"
Week 6	Trusting Expert Systems	Burge, "Computer Proof, Apriori Knowledge, and Other Minds" Wong et al., "How Much Can We Really Trust You? Towards Simple, Interpretable Trust Quantification Metrics for Deep Neural Networks"
Week 7	The Curve Fitting Problem	Forster and Sober, "How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions" Hitchcock and Sober, "Prediction versus Accommodation and the Risk of Overfitting"
Week 8	Causation vs Correlation	Cartwright, "Causal Laws and Effective Strategies" Pearl, *The Book of Why* (selections)
Week 9	Conclusion & Outlook	(no reading)