COURSE SYLLABUS

****

YEAR COURSE OFFERED: 2020
SEMESTER COURSE OFFERED: Spring
DEPARTMENT: Philosophy
COURSE NUMBER: 3395
NAME OF COURSE: Philosophy of Deep Learning
NAME OF INSTRUCTOR: Cameron Buckner

****

The information contained in this class syllabus is subject to change without notice. Students are expected to be aware of any additional course policies presented by the instructor during the course.

****

Time and location: H 2:30-5:30 AH512
Office hours: TBD AH 509
Instructor e-mail: cjbuckner@uh.edu

Learning Objectives

Deep learning neural networks have recently blown through anticipated upper limits on Artificial Intelligence (AI) performance. Though modern deep learning technology is only a few years old, they have already worked themselves into many aspects of our daily lives. They structure and label our search results, organize our shopping lists, recognize our faces, diagnose our diseases, may soon drive our cars, and already defeat us in games as complex as chess, go, and Starcraft II. Many also regard them as the best models of human perceptual judgments. They have also, however, manifested a variety of puzzling foibles, and attracted a number of influential detractors who worry that an overabundance of naïve enthusiasm will lead to another "AI Winter" when hopes for deep learning superintelligence fail to materialize.

In this class, we will critically explore the explosion of deep learning, covering a wide spread of topics in this course so that there's something for everyone: different explanations for deep learning's success (esp. for computer vision, game playing, and medical diagnosis), standard criticisms of deep learning (esp. debates between "rationalists" in AI like Pearl and Marcus and "empiricists" like the Deep Mind group), relevant history of philosophy (esp. from Locke, Hume, Berkeley, and Kant), philosophy of science (esp. what kind of explanations deep learning can offer), adversarial examples, the black box/interpretability problem (esp. GDPR law and the XAI movement), the legal philosophy surrounding automated agents driven by deep learning, concerns about problematic bias in deep learning applications, DL as a model of cortical function in cognitive neuroscience, the use of Generative Adversarial Networks in aesthetics and art, and deep learning in scientific data analysis (to interpret fMRI data, discover new exoplanets around distant suns, or predict protein folds to discover new drugs).

All course readings will be available on the course's blackboard site or are linked in the course plan below.

Major Assignments/Exams

This will be a mostly writing-driven course. For both grads and undergrads, the normal progression will be to write a first paper by Week 11, and the second paper will be a revision of the first paper that takes into account my comments and comments from other students in the course. All students will complete peer reviews of the papers of other students (which will themselves be graded; guidelines for this will follow). Your revised Paper #2 should take into account these comments. A first draft of Paper #2 will be due week 14, after which we will complete another round of peer reviews; the final paper will be due during Finals week.

COURSE SYLLABUS

Undergrads

Paper 1 (4-5 pages)	30%
Peer reviews of Paper #1	10%
Paper 2 (10 pages)	50%
Peer Reviews of Paper #2	10%

Grads

Paper 1 (10 pages)	20%
Peer reviews of Paper #1	10%
Paper 2 (20 pages)	40%
Peer Reviews of Paper #2	10%
Course Presentations (x2)	20%

Papers

Both grads and undergrads will need to come up with their own paper topics. Everyone must get their topic approved by me two weeks before the paper is due. It is recommended that you try to tie the paper topic to prior interests or backgrounds. For general advice on how to write a philosophy paper, this is an excellent resource: http://www.jimpryor.net/teaching/guidelines/writing.html

Students with Disabilities: The College of Liberal Arts and Social Sciences, in accordance with 504/ADA guidelines, is committed to providing reasonable academic accommodations to students who request them. Students seeking accommodation must register with the Center for Students with Disabilities (CSD) 713-743-5400 and present approved documentation to me as soon as possible.

Note on "Incompletes": The College of Liberal Arts and Social Sciences has a policy on assigning the grade of Incomplete ("I"). The policy states that "The grade of I (incomplete) is a conditional and temporary grade given when students are passing a course but, for reasons beyond their control, have not completed a relatively small part of all requirements." That means there are three conditions that must be satisfied here to receive an incomplete; you have to be 1) **currently passing the course** and **2) have completed all but a small part of the course requirements**, and **3) the conditions which prevent you from completing all of the coursework have to be outside of your control.** If you do not satisfy all three of these conditions, do not ask me for an Incomplete; I cannot give you one. (If you have missed more than a small amount of coursework due to conditions outside of your control, usually a "medical withdrawal" is the right solution: https://www.uh.edu/caps/services/medical-withdrawals.html .) (An exception to this policy may be possible if it is needed to accommodate a relevant disability; if you have a relevant disability, please follow the guidance for students with disabilities above.)

Counseling and Psychological Services: Counseling and Psychological Services (CAPS)--www.uh.edu/caps—are available for students having difficulties managing stress, adjusting to college, or feeling sad and hopeless. You can reach CAPS) by calling 713-743-5454 during and after business hours for routine appointments or if you or somebody you know is in crisis. The "Let's Talk" program provides a drop-in consultation service at convenient locations and hours around campus. http://www.uh.edu/caps/outreach/lets_talk.html

Student Conduct Policy: CLASS students are expected to abide by the University of Houston's Code of Student Conduct: http://www.uh.edu/dos/behavior-conduct/student-code-of-conduct/

List of discussion/lecture topics

Week 1 – Background: 1980-90s connectionism – Jan 16

- *Microcognition* – Andy Clark

Additional resources:

- Play around in the neural network playground: https://playground.tensorflow.org/

Week 2 – Background: 1980-90s connectionism – Jan 23

- Horgan & Tienson 1989 – Representation without Rules
- SEP connectionism article: https://plato.stanford.edu/entries/connectionism/

Additional resources:

- Smolensky 1988: On the Proper Interpretation of Connectionism
- Shea 2007: Content and its Vehicles in Connectionist Systems

Wednesday Jan 29: Cameron Lecturing at UT MD Anderson Cancer Center Data Science Workgroup 1:30-3:00 Seminar Students welcome, Contact Cameron for details
Lecture title: Recent issues in Transparency and Adversarial Examples in Medical and Scientific Applications of Deep Learning

Week 3 – Background: The Deep learning revolution – Jan 30

- LeCun, Bengio, & Hinton 2015: Deep Learning
- Buckner 2019: Deep learning—A Philosophical Introduction

Additional resources:

- Schmidhuber 2014: Deep Learning in Neural Networks—An Overview
- https://medium.com/intro-to-artificial-intelligence/deep-learning-series-1-intro-to-deep-learning-abb1780ee20
- https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

Week 4 – Landmark achievements – Feb 6

- Silver et al. 2016: Mastering the game of Go with deep neural networks and tree search
- Lisa Miracchi 2019: A Competence Framework for Artificial Intelligence Research

Additional Resources:

- https://medium.com/@jonathan_hui/alphago-how-it-6works-technically-26ddcc085319
- https://medium.com/applied-data-science/how-to-build-your-own-alphazero-ai-using-python-and-keras-7f664945c188
- Serre 2018: Deep Learning—The Good, the Bad, and the Ugly
- Holger Lyre 2020—does AlphaGo actually play Go?

Week 5 – Criticisms & Problems -- Feb 13

(Cameron speaking in Vienna at ESI Winter School on Machine Learning; James Garson leading discussion)

- Gary Marcus 2018: Deep Learning—a Critical Appraisal
- Lake et al. 2017: Building Machines that Learn and Think Like People

Additional resources:

- Goodfellow post on adversarial examples: https://www.kdnuggets.com/2015/07/deep-learning-adversarial-examples-misconceptions.html
- Cody Wild blog post: Know your Adversary—
    - Part 1: https://towardsdatascience.com/know-your-adversary-understanding-adversarial-examples-part-1-2-63af4c2f5830
    - Part 2: https://towardsdatascience.com/the-modeler-strikes-back-defense-strategies-against-adversarial-attacks-9aae07b93d00
- Alex Irpan blog post on reinforcement learning: https://www.alexirpan.com/2018/02/14/rl-hard.html

- Judea Pearl 2019: Theoretical impediments to machine learning and 7 sparks from the causal revolution

Week 6 – Background: The Deep learning revolution – SOTA, GANs and Episodic Buffering/value transfer– Feb 20 (Cognitive Science)

- Hassabis et al 2018: Neuroscience-Inspired Artificial Intelligence
- Gershman 2019: The generative adversarial brain https://www.frontiersin.org/articles/10.3389/frai.2019.00018/full

Additional Resources:

- https://blog.usejournal.com/the-rise-of-generative-adversarial-networks-be52d424e517
- Goodfellow 2016: NeurIPS tutorial on generative adversarial networks (read first sections and look at examples, skip math)
- Hung et al 2019: Optimizing agent behavior over long time scales by transporting value
- Wang et al 2019: Prefrontal Cortex as a meta-reinforcement learning system

Week 7 – Deep learning & neuroscience: empirical evidence – Feb 27 (Philosophy of Science)

- Yamins & DiCarlo 2016: Using goal-driven deep learning models to understand sensory cortex
- DiCarlo et al 2012 – How does the brain solve visual object recognition

Additional Resources

- Khaligh-Razavi & Kriegeskorte 2014: Deep Supervised, but not Unsupervised, Models May Explain IT Cortical Representation
- BP-STDP—Approximating Backpropagation using spike timing dependent plasticity
- Zador 2019: A Critique of Pure Learning and What Artificial Neural Networks can Learn from Animal Brains

Week 8 – Deep learning & neuroscience: explanatory status – March 5 (Philosophy of Science)

- Catherine Stinson 2019: From artificial neurons to idealized cognitive models
- Piccinini & Craver 2011: Functional analyses as mechanism Sketches

Additional resources:

- Mazviita Chirimuuta 2018: Explanation in Cognitive Neuroscience—Causal and non-causal
- Boone & Piccinini 2016: Mechanistic Abstraction

Week 9 – Deep learning & rationalism vs. empiricism (History)

- Excerpts from Christopher Gauker 2011: *Words and Images—An Essay on the Origin of Ideas*
- Buckner 2018: Empiricism without magic—Transformational abstraction in deep convolutional neural networks

Other resources

- Laurence & Margolis 2012: Abstraction and the origin of mental ideas
- Marcus 2019: Innateness and AI

Spring break March 12

Week 10 -- Ethical concerns – Bias, Transparency – March 19 (Ethics)

- Ayanna Howard & Jason Borenstein 2017: The Ugly Truth about Ourselves and our Robot Creations—The Problem of Bias and Social Inequity
- Cynthia Rudin 2019: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Other resources:

- Kate Crawford NeurIPS keynote on Bias: https://www.facebook.com/nipsfoundation/videos/1553500344741199
- Zerilli et al 2018: Transparency in Algorithmic and Human Decision-Making: Is there a double standard?
- Example: https://www.statnews.com/2019/10/24/widely-used-algorithm-hospitals-racial-bias/

Week 11 – Deep Learning and the Law – March 26 (Legal Philosophy)

- Lu, Lee, Kim, & Danks 2019: Good Explanation for Algorithmic Transparency
- Doshi-Velez et al 2019: Accountability of AI under the Law: The Role of Explanation
- Fazelpour & Lipton 2020: Algorithmic Fairness from a Non-Ideal Perspective

Additional Resources:

- Blog post: Who's liable if AI goes wrong? https://bluenotes.anz.com/posts/2017/12/LONGREAD-whos-liable-if-AI-goes-wrong
- NTSB report on self-driving car Uber crash (on Blackboard)
- GDPR and the right to explanation: https://www.oreilly.com/radar/how-will-the-gdpr-impact-machine-learning/
- Stinson in Globe & Mail: https://www.theglobeandmail.com/opinion/article-deep-learning-why-its-time-for-ai-to-get-philosophical/

Paper #1 due March 26

Week 12 – Explainable AI methods – April 2 – (Philosophy of Science, Philosophy of Mind)

- Zachary Lipton 2016: The Mythos of Model Interpretability
- Zednik 2019: Solving the Black Box Problem: A general-purpose recipe for explainable artificial intelligence

Additional Resources

- Gunning & Aha 2019: DARPA's XAI program
- Cushman 2019: Rationalization is Rational

Peer reviews due April 2

Week 13 – Deep learning and art, creativity – April 9 (art/aesthetics)

- Owain Evans 2019: Sensory Optimization—Neural Networks as a Model for Understanding and Creating Art - https://owainevans.github.io/visual_aesthetics/sensory-optimization.html
- Lonce Wyse 2019: Mechanisms of Artistic Creativity in Deep Learning Neural Networks

Additional Resources:

- Gatys, Ecker, and Bethge 2016: Image Style Transfer using Convolutional Neural Networks
- Wang et al 2017: Image Aesthetics Assessment using Deep Chatterjee's Machine
- Chatterjee 2004: The neuropsychology of visual artistic production
- Elgammal et al 2017: CAN—Creative Adversarial Networks Generating "Art" by Learning about Styles and Deviating from Style Norms

Week 14 – Fair comparisons in evaluating deep learning systems – April 16 (Epistemology)

- Ilyas et al 2019: Are Adversarial Examples Bugs or Features?
- Zhou & Firestone 2019: Humans Can Decipher Adversarial Images

Other resources:

- A discussion of Ilyas et al 2019: https://distill.pub/2019/advex-bugs-discussion/

COURSE SYLLABUS

Paper #2 Draft due April 16

Week 15 -- Deep learning and scientific data analysis – April 23 (Philosophy of Science)

- Emily Sullivan 2019: Understanding from Deep Learning Models
- Katie Creel 2020: Transparency in Complex Computational Systems

Other Resources

- Harman & Kulkarni 2006: The Problem of Induction
- Henderson et al 2018: Deep reinforcement learning that matters
- Mohammed AlQuraishi: AlphaFold @ CASP13 "What just happened?" https://moalquraishi.wordpress.com/2018/12/09/alphafold-casp13-what-just-happened/

Peer reviews for Paper #2 draft due April 23
Final exam week: Paper #2 final draft due May 8