Bixin Guo

Philosophy of AI

Outline

1. Introduction to the Course

Melanie Mitchell 2019, *Artificial Intelligence: A Guide for Thinking Humans,* Part I. Background.

2. The "Ancient" History

Buchanan 2005 "A (very) brief history of artificial intelligence"

3. Classical AI

Allen Newell & Herbert Simon, *Computer Science as Empirical Inquiry: Symbols and Search*

Supplementary Reading: John Haugeland, *Semantic Engines* (Introduction to *Mind Design,* First Edition)

4. Connectionism and Machine Learning

Hinton, G.E. (1992). "How neural networks learn from experience." Scientific American 267, no. 3: 144- 151.

Supplementary Reading: Buckner, C. and Garson, J., "Connectionism", The Stanford Encyclopedia of Philosophy.

5. Deep Learning

Buckner 2019 "Deep learning: A philosophical introduction"

Optional: play with https://playground.tensorflow.org/

Supplementary Video: Watch lectures 1,2, and 7 from Deep Mind/UCL lectures on Deep Learning: https://www.youtube.com/playlist?list=PLqYmG7hTraZCDxZ44o4p3N5Anz3lLRVZF

6. Large Language Models

Floridi, L. (2023). AI as agency without intelligence: on ChatGPT, large language models, and other generative models. Philosophy & Technology, 36(1), 15.

Supplementary Reading: Vaswani et al. "Attention is all you need" (Google translate)

Supplementary Reading: Brown et al. "Language Models are Few-Shot Learners" (GPT-3) [sections 1 and 6]

7. The Turing Test

Alan Turing, “Computing g Machinery and Intelligence”, *Mind*.

Supplementary Reading: Susan Sterrett, Turing’s Two Tests for Intelligence

8. Consciousness

Chalmers, D. (2023). Could a large language model be conscious?. arXiv preprint arXiv:2303.07103.

McDermott, D. (2007). Artificial intelligence and consciousness. The Cambridge handbook of consciousness, 117-150.

Supplementary Reading: Chalmers, D. (1995). Facing up to the problem of consciousness. Journal of consciousness studies, 2(3), 200-219

Supplementary Reading: Dennett, *Consciousness Explained*

9. AI for Science

Bradshaw, Langley, & Simon 1983 "Studying Scientific Discovery by Computer Simulation”

Korb 2004 Machine Learning as Philosophy of Science

Supplementary Reading: Collins 1989 "Computers and the Sociology of Scientific Knowledge”

Supplementary Reading: Iten et al. 2020 "Discovering physical concepts with neural networks"

10. Transparency and Interpretability

Creel, K. A. (2020). Transparency in complex computational systems. Philosophy of Science, 87(4), 568- 589.

Zednik 2019 "Solving the Black Box Problem"

Supplementary Reading: Lipton 2019 "The Mythos of Model Interpretability”

Supplementary Reading: Nguyen, “Transparency is Surveillance”

11. Bias in AI

Will Knight, The Dark Secret at the Heart of AI

Gabbrielle Johnson, “Algorithmic Bias: On the Implicit Biases of Social Technology.” Synthese (June 20, 2020, online first)

Supplementary Reading: Fazelpour & Danks (2021) “Algorithmic Bias: Senses, Sources, Solutions”

Supplementary Video: Gender Shades (Video of Joy Buolamwini discussing the *Gender Shades Project*)

12. Algorithmic Fairness

Grant (2023) "Equalized Odds is a Requirement of Algorithmic Fairness"

Supplementary Reading: Creel & Hellman (2021) The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems

13. Healthcare

Bjerring, J. C., & Busch, J. (2021). Artificial intelligence and patient-centered decision-making. Philosophy & Technology, 34, 349-371.

Supplementary Reading: Schönberger, D. (2019). Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. International Journal of Law and Information Technology, 27(2), 171-203.

Supplementary Reading: Kasirzadeh Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy

14. Privacy in the age of Big Data

Matzner (2013). Why Privacy is Not Enough Privacy in the Context of "Ubiquitous Computing" and "Big Data".

Supplementary Reading: Nissenbaum, "Privacy as Contextual Integrity"

15. Moral Machines

Luke Muehlhauser & Nick Bostrom, *Why we need friendly AI*

Supplementary Reading: Stephen Cave, Rune Nyrup, Karina Vold, & Adrian Weller, *Motivations and Risks of Machine Ethics*

Supplementary Reading: Colin Allen , Gary Varner, & Jason Zinser, Prolegomena to any future artificial moral agent