Philosophy and Theory of Artificial Intelligence

A Reader

Draft from November 20, 2023.
Please don't cite or distribute without permission.
Comments welcome!

Levin Hornischer
Levin.Hornischer@lmu.de
https://github.com/LevinHornischer/PhilTheoAI

Contents

Preface

This is the reader for the course "Philosophy and Theory of Artificial Intelligence" given during the winter semester 2023/24 at *LMU Munich* as part of the *Master in Logic and Philosophy of Science*. The reader is written as the course progresses. A website (or rather git repository) with all the course material is found at

`https://github.com/LevinHornischer/PhilTheoAI`.

Comments I'm happy about any comments: spotting typos, finding mistakes, pointing out confusing parts, or simply questions triggered by the material. Just send an informal email to `Levin.Hornischer@lmu.de`.

Content This course provides, as its title suggests, an introduction to both the *philosophy* and the *theory of artificial intelligence*. This field researches the philosophical foundations of artificial intelligence (AI). It recently gained much prominence because of its urgent relevance. AI made astonishing but also disconcerting technological progress. For a recent example, just think of *ChatGPT*. However, we are lacking a theoretical understanding of AI. We would like to answer questions like the following. Why are neural networks—that underlie modern AI—so good at learning from data? And what kind of knowledge do they have? How do they compare to the human-interpretable symbolic AI models that have been used previously? What is even a good language to talk about AI models and the computation that they perform? What are the possibilities and limitations of AI models? We will in particular also investigate the problems of modern AI: How to do deal with its ethical issues like bias or fairness. We look at the black-box problem of neural networks: that they are difficult to interpret and hard to explain. And we consider their lack of robustness: that in similar situations they unexpectedly might behave incorrectly. Answering these questions is not just an engineering task: it crucially also is a philosophical task—which we undertake in this course.

The course title was inspired by the conference series of the same name.

Objectives In terms of content, the course aims to convey an overview of the questions, methodology, and results of the philosophy and theory of

AI. We cover both classic material and cutting-edge research. In terms of skills, the course aims to teach: (1) the basic ability to program an AI model, (2) the ability to critically reflect on the many issues of AI by relating it to established theory in philosophy, and (3) the ability to apply results from the theory of AI to assess its power and limits.

Prerequisites The course does not assume any programming knowledge. It assumes basic familiarity with philosophy (first-year university level), logic (e.g., an introductory course) and mathematics (though not really beyond high-school level). None of these are strictly necessary: by far most of the reader can be understood also without, it will mostly be helpful to appreciate, e.g., remarks about connected and more advanced topics.

Schedule and organization The course is organized as a seminar. Hence, for each session, we have assigned readings. During the session, we first make sure that we all have understood the provided key AI concepts relevant for the session (by arriving at an explanation in the group), and then we critically discuss the readings. The schedule for the readings is found on the course's website. In sum, the readings aim to provide an overview of the field of philosophy and theory of AI.

The organizational principle for selecting the readings was 'question-based'. Each chapter concerns one 'big question' about the philosophy of AI. See the table of contents for a list of those chapters. As usual, there is much more possible content than time, and during the course we can still decide on which of the readings we will focus on.

Other organizational principles would be possible, too; e.g., 'method-based'.

Layout These notes are informal and partially still under construction. For example, there are margin notes to convey more casual comments that you'd rather find in a lecture but usually not in a book. Todo notes indicate, well, that something needs to be done. References are found at the end.

This is a margin note.

This is a todo note

Furth study material In addition to the provided papers, some helpful short explainer videos are found here. References for 'classical' philosophy of AI (i.e., up to the early 1990s) are, e.g., Boden (1990) and Copeland (1993).

Notation Throughout, 'iff' abbreviates 'if and only if'.

Acknowledgement I have taken great inspiration in designing this course from other courses on this topic both by Stephan Hartmann and Timo Freiesleben and by Cameron Buckner.

1 Introduction to AI

Key concepts:

Make sure that you know what these mean, by using the references provided below.

- History of AI: Ada Lovelace, Alan Turing, McCulloch & Pitts, Logic Theorist, Dartmouth workshop, division of AI into life (cybernetics, connectionism, differential equations) and mind (symbolic computing, logic), big data, deep learning revolution.
- Types of AI: classical/symbolic vs neural networks/connectionism, symbolic vs subsymbolic computation.
- Definitions of AI: Turing test, paradigm examples, technological vs scientific aim of AI, virtual vs physical machines
- Types of learning tasks: Supervised learning, unsupervised learning, reinforcement learning. Machine learning pipeline (conceptualization, data, model, deployment).
- Key concepts of artificial neural networks: neurons, layers, feedforward/recurrent, weights, activation function, loss function, backpropagation, learning rate, local/global minima (equilibrium), regularization, overfitting/underfitting.
- Further/advanced concepts: biological plausibility (backprop too global, Boden's 'too neat, too simple, too few, too dry', neuromodulation like GasNet), Hebbian learning (fire together, wire together), predictive coding (Helmholtz's 'unconscious inference, the 'Bayesian brain', cognition as predicting incoming low-level sense information from higher-level neural layers), perceptron (XOR problem), localist (concepts represented by single neurons) vs distributed networks (concepts stored across the whole system)

That's not needed on a first reading, but you can come back to them at a later stage.

Literature:

- A very accessible overview, written at the beginning of the deep learning revolution: M. A. Boden (2016). *AI: Its nature and future*. Oxford: Oxford University Press. Chapters 1 and 4.

- A concise introduction to deep learning and its philosophical aspects: C. Buckner (2019). "Deep learning: A philosophical introduction." In: *Philosophy Compass* 14.10, e12625. DOI: https://doi.org/10.1111/phc3.12625.

- Short explainer videos of central concepts in AI are found here. An excellent detailed mini-series explaining neural networks is found here.

- An encyclopedia entry on AI: S. Bringsjord and N. S. Govindarajulu (2022). "Artificial Intelligence." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta and U. Nodelman. Fall 2022. Metaphysics Research Lab, Stanford University

- A classic on how to define artificial intelligence (now known as the *Turing test*):

  A. M. Turing (1950). "Computing Machinery and Intelligence." In: *Mind* 59.236, pp. 433–460. DOI: https://doi.org/10.1093/mind/LIX.236.433.

  This aims to make precise/operationalize the vague/intuitive question 'Can machines think?' by asking, roughly, whether you cannot distinguish whether answers to your questions where produced by a machine or by a human.

  Study question: As with any conceptual analysis, ask yourself if this one is convincing? Can you think of aspects of intelligence that cannot be tested for in this 'purely behavioristic' way? How else could you test for those?

Turing is a giant of computer science. There is even a Hollywood movie (The Imitation Game*, 2014) portraying Turing's eventful and also tragic life (e.g., he was prosecuted in 1952 for his sexual orientation).*

This text contains comments about race (e.g., p. 448) and gender stereotypes (e.g., p. 434) which should be reflected on critically. Given the classic status of the text, it is included in the syllabus here, but this is to flag that these comments are not silently endorsed.

Building an AI model:

- A great interactive visualization of neural networks is found here.

- A coding exercise (built for this course) is found on the course website here. Its purpose is to build an AI model—step by step and without any coding experience required—in order to get an idea what this process looks like.

2 AI and philosophy of mind

Key concepts:

- Symbol grounding problem
- Octopus test/Chinese room though experiment
- Classicist vs connectionist theories of mind
- Levels of analysis: neural, subsymbolic, symbolic; Marr's levels
- Physical symbol system hypothesis (Newell–Simon) vs Connectionist dynamical system hypothesis (Smolensky)

Literature:

- After having read about the Turing test, the next paper discusses (and denies) whether large language models (LLMs) can, apart from producing the sensible text response, also be said to understand the meaning of this text and in this sense be intelligent.

  E. M. Bender and A. Koller (07/2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463. URL: https://aclanthology.org/2020.acl-main.463.

  The first author, Emily M. Bender, is an influential researcher in Natural Language Processing and AI Ethics, who is also well-known for the stochastic parrot *paper (Bender, Gebru, et al. 2021).*

- Picking up on the question to what extend "neural representations are meaning too", the next paper discusses how to relate classical symbolic approach to cognitive modeling with the connectionist approach:

  P. Smolensky (1988). "On the proper treatment of connectionism." In: *Behavioral and brain sciences* 11.1, pp. 1–74. DOI: https://doi.org/10.1017/S0140525X00052791.

- Further reading on the idea that cognitive agents are dynamical systems:

  T. van Gelder (1998). "The dynamical hypothesis in cognitive science." In: *Behavioral and Brain Sciences* 21.5, pp. 615–628. DOI: 10.1017/S0140525X98001733.

3 AI and epistemology/philosophy of science

Key concepts:

- Scientific method/induction
- Dynamic interaction
- Empiricists/nurture vs rationalists/nature

Literature:

- There is an intriguing similarity between the fields of machine learning and philosophy of science. Philosophy of science investigates the best way of doing scientific induction: building a theory or model from observed data. And machine learning does something very similar: training a model (e.g., neural network) on collected data. So can we use the rich knowledge of philosophy of science to build a theory of machine learning? Or are the fields further apart after all?

  The following two short papers discuss this question. The first essentially identifies the two fields, and the second describes their relation in a more nuanced way as a *dynamic interaction*.

  K. B. Korb (2004). "Introduction: Machine Learning as Philosophy of Science." In: *Minds and Machines* 14, pp. 433–440. DOI: https://doi.org/10.1023/B:MIND.0000045986.90956.7f.

  J. Williamson (2004). "A Dynamic Interaction Between Machine Learning and the Philosophy of Science." In: *Minds and Machines* 14, pp. 539–549. DOI: https://doi.org/10.1023/B:MIND.0000045990.57744.2b.

- Another such dynamic interaction might be between computer science (specifically deep learning) and philosophy (specifically epistemology), as explored in the following book, of which we read the first chapter. Specifically, it discusses the old philosophical question of how we gain knowledge: Empiricists say we get all knowledge comes from sensory experience, while rationalists say that in getting knowledge we rely on our innate concepts about the basic structure of the world. Deep learning is often associated with empiricists

and symbolic AI with rationalist, but the chapter argues for a more nuanced moderate position.

C. Buckner (Forthcoming). *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence*. Oxford University Press. Chapter 1 "Moderate Empiricism and Machine Learning".

4 AI and ethics

Note: There already are dedicated courses on the ethics of AI at LMU. So, after a general overview, we focus here on some specific aspects: algorithmic fairness.
Key concepts:

-

Literature:

- J.-S. Gordon and S. Nyholm (n.d.). "Ethics of Artificial Intelligence." In: *The Internet Encyclopedia of Philosophy*. Available at: https://iep.utm.edu/ethic-ai/ (accessed: 6 Mar 2022).
- E. M. Bender, T. Gebru, et al. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623.
- S. Mitchell et al. (2021). "Algorithmic Fairness: Choices, Assumptions, and Definitions." In: *Annual Review of Statistics and Its Application* 8.1, pp. 141–163. DOI: 10.1146/annurev-statistics-042720-125902. eprint: https://doi.org/10.1146/annurev-statistics-042720-125902. URL: https://doi.org/10.1146/annurev-statistics-042720-125902.
- P. Schwöbel and P. Remmers (2022). "The Long Arc of Fairness: Formalisations and Ethical Discourse." In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, pp. 2179–2188. DOI: 10.1145/3531146.3534635. URL: https://doi.org/10.1145/3531146.3534635.
- S. Verma and J. Rubin (2018). "Fairness Definitions Explained." In: *Proceedings of the International Workshop on Software Fairness*. FairWare '18. Gothenburg, Sweden: Association for Computing Machinery, pp. 1–7. DOI: 10.1145/3194770.3194776. URL: https://doi.org/10.1145/3194770.3194776

- F. Beigang (2023). "Yet Another Impossibility Theorem in Algorithmic Fairness." In: *Minds and Machines*. DOI: https://doi.org/10.1007/s11023-023-09645-x

5 Theory of AI: Power and Limits

Key concepts:

- Classic computability theory (Church–Turing thesis, Halting problem, Gödel Incompleteness, complexity theory)
- Stochastic learning theory (PAC learnability, No Free Lunch theorems) and universal approximation theorems
- Analog computation (Shannon–Pour-El thesis)
- Classic philosophy of AI: Gödel impossibility for (symbolic) AI? Modern impossiblity result for machine learning?

Literature:

- T. F. Sterkenburg and P. D. Grünwald (2021). "The no-free-lunch theorems of supervised learning." In: *Synthese* 199.3-4, pp. 9979–10015. DOI: 10.1007/s11229-021-03233-1
- M. B. Pour-El (1974). "Abstract computability and its relation to the general purpose analog computer (some connections between logic, differential equations and analog computers)." In: *Transactions of the American Mathematical Society* 199, pp. 1–28. DOI: https://doi.org/10.2307/1996870
- I. Van Rooij (2008). "The Tractable Cognition Thesis." In: *Cognitive Science* 32.6, pp. 939–984. DOI: 10.1080/03640210801897856. URL: https://onlinelibrary.wiley.com/doi/abs/10.1080/03640210801897856
- M. J. Colbrook et al. (2022). "The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem." In: *Proceedings of the National Academy of Sciences* 119.12, e2107151119. DOI: 10.1073/pnas.2107151119. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2107151119. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2107151119

6 Reliable AI

Key concepts:

- Interpretable AI
- Explainable AI
- Robustness in AI

Literature: Interpretable AI

- Z. C. Lipton (09/2018). "The Mythos of Model Interpretability." In: *Commun. ACM* 61.10, pp. 36–43. DOI: 10.1145/3233231. URL: https://doi.org/10.1145/3233231
- F. Doshi-Velez and B. Kim (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. arXiv: 1702.08608 [stat.ML]
- C. Rudin (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." In: *Nature Machine Intelligence* 1, pp. 206–215. DOI: https://doi.org/10.1038/s42256-019-0048-x

Literature: Explainable AI

- T. Miller (2019). "Explanation in artificial intelligence: Insights from the social sciences." In: *Artificial Intelligence* 267. DOI: https://doi.org/10.1016/j.artint.2018.07.007
- J. Woodward and L. Ross (2021). "Scientific Explanation." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University

Literature: Robust AI

- T. Freiesleben and T. Grote (2023). "Beyond generalization: a theory of robustness in machine learning." In: *Synthese* 202.109. DOI: https://doi.org/10.1007/s11229-023-04334-9

Bibliography

Beigang, F. (2023). "Yet Another Impossibility Theorem in Algorithmic Fairness." In: *Minds and Machines*. DOI: https://doi.org/10.1007/s11023-023-09645-x (cit. on p. 10).

Bender, E. M., T. Gebru, et al. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623 (cit. on pp. 6, 9).

Bender, E. M. and A. Koller (07/2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463. URL: https://aclanthology.org/2020.acl-main.463 (cit. on p. 6).

Boden, M. A., ed. (1990). *The Philosophy of Artificial Intelligence*. Oxford Readings in Philosophy. New York: Oxford University Press (cit. on p. 2).

— (2016). *AI: Its nature and future*. Oxford: Oxford University Press (cit. on p. 4).

Bringsjord, S. and N. S. Govindarajulu (2022). "Artificial Intelligence." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta and U. Nodelman. Fall 2022. Metaphysics Research Lab, Stanford University (cit. on p. 5).

Buckner, C. (2019). "Deep learning: A philosophical introduction." In: *Philosophy Compass* 14.10, e12625. DOI: https://doi.org/10.1111/phc3.12625 (cit. on p. 5).

— (Forthcoming). *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence*. Oxford University Press (cit. on p. 8).

Colbrook, M. J., V. Antun, and A. C. Hansen (2022). "The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem." In: *Proceedings of the National Academy of Sciences* 119.12, e2107151119. DOI: 10.1073/pnas.2107151119. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2107151119. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2107151119 (cit. on p. 11).

Copeland, J. (1993). *Artificial Intelligence: A Philosophical Introduction*. Wiley-Blackwell (cit. on p. 2).

Doshi-Velez, F. and B. Kim (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. arXiv: 1702.08608 [stat.ML] (cit. on p. 12).

Freiesleben, T. and T. Grote (2023). "Beyond generalization: a theory of robustness in machine learning." In: *Synthese* 202.109. DOI: https://doi.org/10.1007/s11229-023-04334-9 (cit. on p. 12).

Gordon, J.-S. and S. Nyholm (n.d.). "Ethics of Artificial Intelligence." In: *The Internet Encyclopedia of Philosophy*. Available at: https://iep.utm.edu/ethic-ai/ (accessed: 6 Mar 2022) (cit. on p. 9).

Korb, K. B. (2004). "Introduction: Machine Learning as Philosophy of Science." In: *Minds and Machines* 14, pp. 433–440. DOI: https://doi.org/10.1023/B:MIND.0000045986.90956.7f (cit. on p. 7).

Lipton, Z. C. (09/2018). "The Mythos of Model Interpretability." In: *Commun. ACM* 61.10, pp. 36–43. DOI: 10.1145/3233231. URL: https://doi.org/10.1145/3233231 (cit. on p. 12).

Miller, T. (2019). "Explanation in artificial intelligence: Insights from the social sciences." In: *Artificial Intelligence* 267. DOI: https://doi.org/10.1016/j.artint.2018.07.007 (cit. on p. 12).

Mitchell, S. et al. (2021). "Algorithmic Fairness: Choices, Assumptions, and Definitions." In: *Annual Review of Statistics and Its Application* 8.1, pp. 141–163. DOI: 10.1146/annurev-statistics-042720-125902. eprint: https://doi.org/10.1146/annurev-statistics-042720-125902. URL: https://doi.org/10.1146/annurev-statistics-042720-125902 (cit. on p. 9).

Pour-El, M. B. (1974). "Abstract computability and its relation to the general purpose analog computer (some connections between logic, differential equations and analog computers)." In: *Transactions of the American Mathematical Society* 199, pp. 1–28. DOI: https://doi.org/10.2307/1996870 (cit. on p. 11).

Rudin, C. (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." In: *Nature Machine Intelligence* 1, pp. 206–215. DOI: https://doi.org/10.1038/s42256-019-0048-x (cit. on p. 12).

Schwöbel, P. and P. Remmers (2022). "The Long Arc of Fairness: Formalisations and Ethical Discourse." In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, pp. 2179–2188. DOI: 10.1145/3531146.3534635. URL: https://doi.org/10.1145/3531146.3534635 (cit. on p. 9).

Smolensky, P. (1988). "On the proper treatment of connectionism." In: *Behavioral and brain sciences* 11.1, pp. 1–74. DOI: https://doi.org/10.1017/S0140525X00052791 (cit. on p. 6).

Sterkenburg, T. F. and P. D. Grünwald (2021). "The no-free-lunch theorems of supervised learning." In: *Synthese* 199.3-4, pp. 9979–10015. DOI: 10.1007/s11229-021-03233-1 (cit. on p. 11).

Turing, A. M. (1950). "Computing Machinery and Intelligence." In: *Mind* 59.236, pp. 433–460. DOI: https://doi.org/10.1093/mind/LIX.236.433 (cit. on p. 5).

Van Rooij, I. (2008). "The Tractable Cognition Thesis." In: *Cognitive Science* 32.6, pp. 939–984. DOI: 10.1080/03640210801897856. URL: https://onlinelibrary.wiley.com/doi/abs/10.1080/03640210801897856 (cit. on p. 11).

Van Gelder, T. (1998). "The dynamical hypothesis in cognitive science." In: *Behavioral and Brain Sciences* 21.5, pp. 615–628. DOI: 10.1017/S0140525X98001733 (cit. on p. 6).

Verma, S. and J. Rubin (2018). "Fairness Definitions Explained." In: *Proceedings of the International Workshop on Software Fairness*. FairWare '18. Gothenburg, Sweden: Association for Computing Machinery, pp. 1–7.

DOI: 10.1145/3194770.3194776. URL: https://doi.org/10.1145/3194770.3194776 (cit. on p. 9).

Williamson, J. (2004). "A Dynamic Interaction Between Machine Learning and the Philosophy of Science." In: *Minds and Machines* 14, pp. 539–549. DOI: https://doi.org/10.1023/B:MIND.0000045990.57744.2b (cit. on p. 7).

Woodward, J. and L. Ross (2021). "Scientific Explanation." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University (cit. on p. 12).