Philosophy and Theory of Artificial Intelligence

A Reader

Draft from October 30, 2024.
Please don't cite or distribute without permission.
Comments welcome!

Levin Hornischer
Levin.Hornischer@lmu.de
https://levinhornischer.github.io/PhilTheoAI/

Contents

Preface

This is the reader for the course "Philosophy and Theory of Artificial Intelligence" given during the winter semester 2024/25 at *LMU Munich* as part of the *Master in Logic and Philosophy of Science*. The reader is updated as the course progresses. A website with all the course material is found at

`https://levinhornischer.github.io/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.

After an introduction to modern 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

Further study material In addition to the provided papers, some helpful short explainer videos on AI are found here. And on philosophy of neuroscience 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

This chapter's big question

What actually is an AI system and, in particular, a neural network?

Key concepts

- 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 subsymbolic/neural networks/connectionism.
- Definitions of AI: Turing test (more on this in chapter 2), 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 (as your way of telling the neural network what to optimize for), backpropagation, learning rate, local/global minima (equilibrium), regularization, overfitting/underfitting.

Before one can do *philosophy of* X, one needs a good understanding of X. So we start with an introduction to AI, both practically and theoretically.

For a practical introduction, we see, in the very first lecture, how an AI system is actually built in practice. We consider the standard example of training a neural network to classify hand-written digits (on the MNIST dataset). Thus, we get a concrete idea of what an 'AI system' really is and this does not remain an abstract term in future discussion. We build the system in the form of a coding exercise, which is purpose-built for

this course and available on the course website. But—fear not—you do not need any coding experience for this! In class, we go through the parts of the coding exercise. At home, you are then asked to change the parameters and see how the performance of the neural network changes. Your challenge will be to find some parameters with which the networks reaches an accuracy of 98%. In the next lecture, we will discuss your observations.

For a theoretical introduction, the readings below and also the coding exercise introduce the central concepts of modern AI, which are summarized in this chapters list of 'key concepts'. Make sure that, by doing the readings, you know what these concepts mean. We discuss them in the second lecture of the course. The readings also mentioned the following further advanced concepts. You can skip them on a first reading and come back to them at a later stage:

- 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)

Readings

- 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.
- Short explainer videos of central concepts in AI are found here. An excellent detailed mini-series explaining neural networks is found here.
- The coding exercise on the course website.

Further material

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

- An introduction to AI from the standard textbook: Russell and Norvig (2021, ch. 1). They have a fourfold definition of AI: acting humanly (Turing test), thinking humanly (cognitive modeling), thinking rationally (logic, probability), acting rationally (rational agent; perfect vs limited rationality).
- 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

2 AI and philosophy of mind

This chapter's big question

Can AI systems think? Like humans?

Key concepts

- Turing test
- 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)

A classic text on the question whether an AI system—or, simply, a 'machine'—can think is Turing's 1950 paper, in which he introduces what is now known as the *Turing test*. Some even take this as the very definition of artificial intelligence, i.e., when a machine should be considered intelligent. (Three other definitions are mentioned by Russell and Norvig (2021, ch. 1), see 'further reading' in chapter 1.) The Turing test aims to make precise the intuitive question 'Can machines think?' by operationalizing it in a behavioristic way: can you distinguish whether answers to your questions where produced by the 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).*

Fast-forward 70+ years, where we have 'machines' like ChatGPT, what about Turing's 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.

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 extent "neural representations are meaning too", the next paper discusses how to relate classical symbolic approach to cognitive modeling with the connectionist approach.

Readings

- 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.
- 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.
- 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.

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.

This is quite a dense text: on a first reading focus on understanding the key concepts mentioned above. This is a text worth coming back to over and over again.

Further material

- For an overview of the discussion around the Turing test, see: G. Oppy and D. Dowe (2021). "The Turing Test." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Winter 2021. Metaphysics Research Lab, Stanford University.
- 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

This chapter's big question

How do AI systems gain knowledge, if any?

Key concepts

- Empiricists/nurture vs rationalists/nature
- Domain-general vs domain-specific cognitive systems
- Moderate empiricism: allow 'innate' general inductive biases to learn domain-specific knowledge.
- Control Problem: If several general modules (representing various cognitive faculties) are posited, how do they fruitfully interact?
- Neurosymbolic computation: "combine robust learning in neural networks with reasoning and explainability by offering symbolic representations for neural models".
- Neural-symbolic cycle: compile a neural network from symbolic knowledge (in weights or semantic loss function) and decompile the neural network into symbolic knowledge
- Relational embedding
- Kahneman "Thinking, Fast and Slow" (2011): System 1 (implicit, fast, parallel, instinctive, emotional) vs system 2 (explicit, slow, sequential, deliberative, logical).

Literature:

- Another 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 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. It endorses the "new empiricist DoGMA [that a] (Do)main General Modular Architecture is the best hope for modeling rational cognition in AI" (p. 26).

As remarked in footnote 28 of the text, this is in reference to the famous paper of Quine (1951) "Two Dogmas of Empiricism" (here's a lecture on it).

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".

- Given this distinction between rationalist/symbolic and empiricist/subsymbolic approaches to AI, wouldn't it make sense to combine the two (as Kant attempted)? Especially in light of the fact that they have complementary benefits? To get the best of both worlds, rather than sharp opposition? There is a movement aiming to do precisely this: known as *neurosymbolic computation*. (Buckner cites 'cognitive' versions of this—namely ACT-R, SOAR, Sigma, and CMC—on page 39.) A recent overview is the following.

  A. d. Garcez and L. C. Lamb (2023). "Neurosymbolic AI: the 3rd wave." In: *Artificial Intelligence Review*. DOI: https://doi.org/10.1007/s10462-023-10448-w

4 AI and philosophy of language

This chapter's big question

Do Large Language Models have linguistic and cognitive competence or are they just stochastic parrots?

The term 'stochastic parrot' is from the title of the paper by Bender, Gebru, et al. (2021).

Key concepts

-

Literature

- R. Millière and C. Buckner (2024a). "A Philosophical Introduction to Language Models – Part I: Continuity With Classic Debates." In: arXiv: 2401.03910 [cs.CL]. URL: https://arxiv.org/abs/2401.03910
- R. Millière and C. Buckner (2024b). "A Philosophical Introduction to Language Models – Part II: The Way Forward." In: arXiv: 2405.03207 [cs.CL]. URL: https://arxiv.org/abs/2405.03207
- Great introductory videos to LLMs: part 1, part 2, and part 3.

5 AI and philosophy of science

This chapter's big question

Is machine learning the 'end of theory'?

Referring to a 2008 Wired article by Chris Anderson with the title 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete'.

Key concepts

- Scientific method/induction
- Dynamic interaction
- Opacity problem
- elementwise/localist vs holistic/distributed representation
- Model audit vs scientific inference

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.

- A recent text going into more detail and modern applications is this one:

  T. Freiesleben, G. König, et al. (2022). *Scientific Inference With Interpretable Machine Learning: Analyzing Models to Learn About Real-World Phenomena*. arXiv: 2206.05487 [stat.ML]

  This recently turned into an online book.

  They highlight a problem of modern AI models: even if they learn well from the data and predict the scientific phenomenon near perfectly, they are still not the *kind of* models that scientists favor. This is because these AI models are very complex (so it is hard to 'understand' the model) and it is unclear how the parts of the model relate to the parts of the phenomenon. This is known as the *opacity problem* (references: Sullivan 2020 and Boge 2022). Interpretable machine learning and explainable artificial intelligence (see chapter 8) aim to make AI models more 'understandable'. But it is not clear whether these methods can be used to draw scientific inference about the real phenomenon from the AI model. The paper discusses why and how this problem can be addressed.

6 AI and ethics

This chapter's big question

AI systems are just objective computer models, so they must be fair, right?

There already are dedicated courses on the ethics of AI at LMU, so we focus here on some specific aspect: algorithmic fairness. We do still provide references to general overviews on ethics of AI.

Key concepts

- Policy goals as prediction tasks, and the potentially problematic modeling assumptions behind this (overarching goal, population choice, decision space).
- Bias in data: statistical and societal
- Model architecture: interpretability, perturbation, choice of features.
- Model evaluation assumptions (no interference, uniform, simultaneous)
- Identifying advantaged and disadvantaged groups: intersectionality (Crenshaw)
- Oblivious (purely probabilistic) vs non-oblivious (also including similarity metric between individuals and causality) fairness definitions.
- Oblivious: With D = decision, S = predicted score, A = sensitive variable, X = insensitive variable, Y = true outcome, (1) $D \perp A|Y = 0$, (2) $D \perp A|Y = 1$, (3) $Y \perp A|D = 0$, (4) $Y \perp A|D = 1$, (5) $D \perp A$, (6) $D \perp A|X$.
- Impossibility results: The seminal one that equalized odds and predictive parity are jointly impossible (Chouldechova 2017;

Cf. the machine learning pipeline: here we might call it the 'prediction-model construction pipeline'.

Kleinberg et al. 2016). Other ones to include a notion of counterfactual fairness using causal models (Beigang 2023).

Literature:

- An encyclopedia overview on Ethics of AI:

  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).

- A famous paper on the ethics of Large Language Models.

  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.

- An overview of the field of algorithmic fairness.

  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.

- A discussion of the interplay between formalized technical discourse of fairness and informal ethical discourse of fairness.

  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.

- A recent impossibility result showing that three plausible requirements for a predictive algorithm to be fair (equalized odds, predictive parity, and counterfactual fairness) are in fact jointly inconsistent. It can be seen as a continuation of the quite short section 5 of Mitchell et al. (2021) discussing causal reasoning in fairness definitions.

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

7 Theory of AI: Power and Limits

This chapter's big question

Is there anything AI models cannot do?

Key concepts

- Stochastic learning theory (PAC learnability, No Free Lunch theorems)
- Universal approximation theorems

We start this part of the course with a short background lecture on classic computability theory (Church–Turing thesis, Halting problem, Gödel Incompleteness, complexity theory) and a bit on extension to continuous computation (computable analysis, analog computation, Shannon–Pour-El thesis).

Maybe the two most central classical theorems from the theory of AI are the *No Free Lunch theorem* and the *Universal Approximation theorem*. Their meaning and consequences are discussed in the next two readings, respectively.

add brief take-home message of the two papers

- 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. 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

  The Universal Approximation theorem shows that for every (continuous) function describing the true input-output relation in the world, there is a neural network with a weight setting that approximates this function to any desired accuracy. However, the paper asks

whether a learning algorithm (like backpropagation) can actually find this weight setting via learning from finite data about the true function. They provide a negative results: Despite the existence of a neural network

"We show that there are problems where stable and accurate NNs exist, yet no algorithm can produce such a network. Regardless of how many computational resources or data one throws at the problem, this impossibility result holds." Colbrook 2022

Just some further readings are as follows. We will go into much more detail in the course *Advanced Topics in the Foundations of AI* that I will offer next semester.

- For some further computability-theoretic analysis of the inverse problems studied in the preceding paper, see

  H. Boche et al. (2022). *Inverse Problems Are Solvable on Real Number Signal Processing Hardware*. arXiv: 2204.02066 [eess.SP].

- For a general analysis of the capabilities of analog computation, see

  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

  (For an overview of analog computation, see Bournez and Pouly 2021.)

- For the use of computability theory in the study of human intelligence (aka cognitive science :-)), see

  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

8 Reliable AI

This chapter's big question

What can we do about the black box nature of neural networks?

Key concepts

- Interpretable AI: Reasons for requiring interpretability, interpretability as transparency, interpretatability as post-hoc explanation.
- 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

  Reasons for requiring interpretability: *trust* (that model will perform well, though hard to make precise), *causality* (infer true causal relation in the world from the correlations detected by the model), *transferability* (to new domains of application, still knowing that the model will work).

  Interpretability as *transparency*: *simulatability* (a human can simulate the model at once), *decomposability* (each part of the model admits intuitive interpretation; cf. elementwise/localist representation), *algorithmic transparency* (the learning algorithm with which the model is built is understood well: provably converges to best solution, etc.). Humans aren't transparent in any of these senses. Linear models are in general only interpretable in the last sense, and the input features that they used might be processed while they are readily interpretable for neural networks.

  Interpretability as *post-hoc interpretation/explanation*: Provide additional information to elucidate why the model provided a certain

output, without necessarily being faithful to the underlying mechanism producing the output (hence this can be misleading). If humans are interpretable, it is in this sense. Examples: *text explanation* (provide a verbal explanation of model output; no guarantee of correctness), *visualization* (e.g., visualize high-dimensional representations in 2D images), *local explanation* (e.g., saliency maps showing which parts of the input were most important to the output in the sense that changing them will most likely change the output; can be misleading), *explanation by example* (e.g., which datapoints are most similar/important for the model behavior).

Since the publication of the paper, many more local explanation methods are known: counterfactual explanation, SHAP, intergrated gradients, etc. But they also face certain impossibilities: see Bilodeau et al. (2024).

- 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

- B. Mittelstadt et al. (2019). "Explaining Explanations in AI." in: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, pp. 279–288. DOI: 10.1145/3287560.3287574. URL: https://doi.org/10.1145/3287560.3287574

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. 15).

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. 7, 11, 15).

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. 8).

Bilodeau, B. et al. (2024). "Impossibility theorems for feature attribution." In: *Proceedings of the National Academy of Sciences* 121.2, e2304406120 (cit. on p. 19).

Boche, H., A. Fono, and G. Kutyniok (2022). *Inverse Problems Are Solvable on Real Number Signal Processing Hardware*. arXiv: 2204.02066 [eess.SP] (cit. on p. 17).

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. 5).

Bournez, O. and A. Pouly (2021). "A Survey on Analog Models of Computation." In: *Handbook of Computability and Complexity in Analysis*. Ed. by V. Brattka and P. Hertling. Cham: Springer International Publishing, pp. 173–226. DOI: 10.1007/978-3-030-59234-9_6 (cit. on p. 17).

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. 6).

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. 6).

— (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. 10).

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. 16).

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. 19).

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. 19).

Freiesleben, T., G. König, et al. (2022). *Scientific Inference With Interpretable Machine Learning: Analyzing Models to Learn About Real-World Phenomena*. arXiv: 2206.05487 [stat.ML] (cit. on p. 13).

Garcez, A. d. and L. C. Lamb (2023). "Neurosymbolic AI: the 3rd wave." In: *Artificial Intelligence Review*. DOI: https://doi.org/10.1007/s10462-023-10448-w (cit. on p. 10).

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. 15).

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. 12).

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. 18).

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. 19).

Millière, R. and C. Buckner (2024a). "A Philosophical Introduction to Language Models – Part I: Continuity With Classic Debates." In: arXiv: 2401.03910 [cs.CL]. URL: https://arxiv.org/abs/2401.03910 (cit. on p. 11).

— (2024b). "A Philosophical Introduction to Language Models – Part II: The Way Forward." In: arXiv: 2405.03207 [cs.CL]. URL: https://arxiv.org/abs/2405.03207 (cit. on p. 11).

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. 15).

Mittelstadt, B., C. Russell, and S. Wachter (2019). "Explaining Explanations in AI." In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, pp. 279–288. DOI: 10.1145/3287560.3287574. URL: https://doi.org/10.1145/3287560.3287574 (cit. on p. 19).

Oppy, G. and D. Dowe (2021). "The Turing Test." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Winter 2021. Metaphysics Research Lab, Stanford University (cit. on p. 8).

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. 17).

Quine, W. V. O. (1951). "Two Dogmas of Empiricism." In: *Philosophical Review* 60.1, pp. 20–43 (cit. on p. 10).

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. 19).

Russell, S. J. and P. Norvig (2021). *Artificial Intelligence: A Modern Approach*. Pearson series in artificial intelligence. Harlow: Pearson (cit. on pp. 6 sq.).

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. 15).

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. 8).

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. 16).

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. 8).

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. 17).

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. 8).

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. 12).

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. 19).