Philosophy of AI

Winter 2020/21

LMU MUNICH, FACULTY OF PHILOSOPHY, PHILOSOPHY OF SCIENCE AND THE STUDY OF RELIGION, MCMP

Instructors:	Stephan Hartmann	Timo Freiesleben
Email:	S.Hartmann@lmu.de	Timo.Freiesleben@campus.lmu.de

Location: Due to COVID-19, this course will take place online via Zoom.

- **Link:** https://lmu-munich.zoom.us/j/94508603112?pwd=bEF3d1hvMEpBY2pnYkFRSDBTUUQ1UT09
- **Meeting ID:** 945 0860 3112, **Passcode:** 439008

Time: Mondays 2:15-3:45 pm
Material: The readings and further material can be found in the dropbox:

- https://www.dropbox.com/sh/ybvl4jjr9ugtj7a/AABLFa9m-V88bE8x2-UjP1cGa?dl=0

Office hours: By appointment

Background literature: Here are some interesting books and an article which will be discussed in the course. You need to consult them occasionally.

- **AIMA:** Russell, S., & Norvig, P. (^{3}2016). Artificial Intelligence: A Modern Approach. Addison Wesley.
- **BOW:** Pearl, J., & Mackenzie, D. (2018). The Book of Why: The New Science of Cause and Effect. Basic Books.
- **SEP-AI:** Bringsjord, S. & Govindarajulu, N.S. (2020). “Artificial Intelligence”, The Stanford Encyclopedia of Philosophy (Summer 2020 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2020/entries/artificial-intelligence/.
- **PAI:** Smith B.C. (2019). The Promise of Artificial Intelligence: Reckoning and Judgment. MIT Press.
- **IML:** Molnar, C. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/.

Overview: This course is neither about the danger of robots taking over world domination, nor about trolley problems. Instead, we will look at questions currently being discussed in artificial intelligence research from a Philosophy of Science perspective. Among the topics discussed are the concept of intelligence, explainable AI, the epistemology of machine learning, the ethics of AI, and the relationship between AI and other sciences. To allow for informed discussions, we will give short introductions to key elements of modern AI such as (deep) neural networks, reinforcement learning and causal graphs.

Objectives: This course introduces students to the philosophical topics discussed and the methods used in the field of artificial intelligence. By the end of the course, students should have an understanding of the philosophical foundations of AI and its underlying assumptions. Moreover, students should be able to outline conceptual problems of the field and discuss them critically and in-depth.

Prerequisites: It will be helpful, esp. in the later sessions, to have some familiarity with the problems and questions of (general) philosophy of science and epistemology. A high-school understanding of probability, linear algebra, and calculus helps to wrap your head around the methods. Programming in Matlab/Octave comes in handy in the optional exercises. It is, however, no prerequisite for successful participation.

Coursework: Students are requested to attend all sessions, carefully study the reading assignments, and participate in the discussions. Students are also strongly encouraged to solve four practical exercises which illustrate how the methods we discuss do their magic. Solving these exercises is not a necessary condition for passing the course. It may, however, give you a small bonus on the grade of the final essay. If you want to get the bonus, then send in your solutions in one single ZIP-file with the naming convention lastname.firstname.zip (example: freiesleben.timo.zip) to Timo before the Christmas holidays.

Final Essay: Students are allowed to write the final essay in groups of up to four people. The expected word requirements depend the number of collaborating students:

- 1 Person: 4,500 words (±5%)
- 2 Persons: 6,500 words (±5%)
- 3 Persons: 8,000 words (±5%)
- 4 Persons: 9,000 words (±5%)

Due to the current situation (COVID-19), the new deadline is **09.04.2021** (Before: **20.03.2021**) noon. Do not forget to register in the LSF for the exam (In our case there is no exam but only the essay) between **18.01–29.01.2021**. Students choose the topic of their essay themselves but should discuss it in advance with one of the instructors. In the final session, students may get a time slot of 5 minutes to pitch their essay ideas and to get feedback from the whole group. This opportunity is optional and not obligatory.

Assessment: The final mark will be determined by the mark of the final essay. Correct solutions in at least two of the four exercises will give a bonus of 0.3 marks to the final mark. (Note: There will be no bonus if the final essay gets a 1.0 mark.) In group essays, only those students who handed in the exercises get the bonus on the final grade.

Videos: There are many instructive videos on AI, machine learning related topics on the internet. The playlist of Crash Course `https://tinyurl.com/AIcrash-course`, for example, offers a number of entertaining little introductions to topics which are also covered in this course such as 'what is AI?', 'symbolic AI', 'neural networks', or 'ethical AI'.
We especially recommend Stanford University's free online course on machine learning `https://www.coursera.org/learn/machine-learning?#syllabus` taught by Andrew Ng. Amongst other things, this course introduces regression methods, neural networks, and unsupervised techniques such as k-means in a highly accessible way. In week 2, it also provides a great introduction to all you need to know about Octave to manage the optional exercises taught in only a few hours. Watching this video is recommended to all students with no programming basics in Octave.

Topics

1. **Week: (02.11.2020) Introduction**
    - Topics: *General Introduction, History of AI, the State of the Art*
    - Required Reading: None
    - Background Readings: AIMA, ch. 1.2–1.4; SEP-AI, sec. 1

2. **Week: (09.11.2020) Intelligence**
    - Topics: *What Means AI?, How to Test for AI?, Strong & Weak AI*
    - Required Readings:
        - Turing, A.M. (1950). "Computing machinery and intelligence". Repr. in: R. Epstein, G. Roberts and G. Beber (eds.): Parsing the Turing Test. Springer, Dordrecht 2009, pp. 23–65
        - AIMA, ch. 1.1
    - Background Readings: SEP-AI; AIMA, ch. 26; BOW, ch. 10
3. **Week: (16.11.2020) Methods 1: Logic and Symbolic AI**
    - Topics: *FOL, Decision Trees, Assumptions, Problems and Advantages of Logical AI*
    - Required Readings:
        - PAI 2+3
        - Dreyfus, H.L. (2007). "Why Heideggerian AI failed and how fixing it would require making it more Heideggerian". Artificial Intelligence, 171(18), 1137-1160, esp. pp. 1137–1139
    - Background Readings: AIMA, chs. 7–12; SEP-AI, secs. 3.2+8.3; IML, ch. 4
4. **Week: (23.11.2020) Methods 2: Probability, Statistics, and Causality**
    - Topics: *Bayesianism, Regression Models, Inference, Rationality Assumptions, Problems and Advantages of Statistics*
    - Required Readings:
        - BOW, chs. 1 + 2
        - Hartmann, S., & Sprenger, J. (2010). "Bayesian epistemology". In: S. Bernecker & D. Pritchard (eds.): Routledge Companion to Epistemology, pp. 609–620
    - Background Readings: SEP-AI 4.3; IML 4;
      Spirtes, P. (2010). "Introduction to causal inference". Journal of Machine Learning Research, 11(5);
      McCarthy, J., & Hayes, P.J. (1981). "Some philosophical problems from the standpoint of artificial intelligence". In: Readings in Artificial Intelligence. Morgan Kaufmann, pp. 431–450
5. **Week: (30.11.2020) Methods 3: Neural Networks and Connectionism**
    - Topics: *Biological Motivation, Types of Artificial Neural Networks, the Strengths and Weaknesses of Artificial Neural Networks, Supervised Learning, Conncetionism*
    - Required Readings:
        - Bzdok, D., Altman, N., & Krzywinski, M. (2018). "Points of significance: statistics versus machine learning"
        - Buckner, C. (2018). "Empiricism without magic: Transformational abstraction in deep convolutional neural networks". Synthese, 195(12), 5339–5372
    - Background Readings: POI, chs. 5+6; SEP-AI, sec. 4.2;
      Buckner, C. & Garson, J. (2019). "Connectionism", , secs. 4+5. In: The Stanford Encyclopedia of Philosophy (Fall 2019 Edition), Edward N. Zalta (ed.),
      https://plato.stanford.edu/archives/fall2019/entries/connectionism/;
      Stinson, C. (2020). "From implausible artificial neurons to idealized cognitive models: Rebooting philosophy of artificial intelligence". Philosophy of Science, 87(4), 590–611;
      McCulloch, W.S., & Pitts, W. (1943). "A logical calculus of the ideas immanent in nervous activity". The Bulletin of Mathematical Biophysics, 5(4), 115–133;
      Siegelmann, H.T., & Sontag, E.D. (1995). "On the computational power of neural nets". Journal

of Computer and System Sciences, 50(1), 132–150;
Wolpert, D.H., & Macready, W.G. (1997). “No free lunch theorems for optimization”. IEEE Transactions on Evolutionary Computation, 1(1), 67–82

6. **Week: (07.12.2020) Methods 4: Unsupervised and Reinforcement Learning**

    - Topics: *Learning Without Labels, Agent Architectures, Goal Alignment, Concept Learning*
    - Required Reading:
        - Dreyfus, H.L. (2007). “Why Heideggerian AI failed and how fixing it would require making it more Heideggerian”. Artificial Intelligence, 171(18), 1137–1160
    - Background Readings: Sutton, R.S. & Barto, A.G. (2018). Reinforcement Learning: An Introduction. MIT Press;
      Hinton, G.E., Sejnowski, T.J., & Poggio, T.A. (Eds.). (1999). Unsupervised Learning: Foundations of Neural Computation. MIT Press

7. **Week: (14.12.2020) Explainable AI**

    - Topics: *Explanation, Interpretation, Causality, Explainability/Accuracy Trade-Off, Model-Agnostic/Specific*
    - Required Readings:
        - Doshi-Velez, F., & Kim, B. (2017). “Towards a rigorous science of interpretable machine learning”. arXiv preprint arXiv:1702.08608
        - Rudin, C. (2019). “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. Nature Machine Intelligence, 1(5), 206–215
    - Background Readings: IML, esp. ch. 2;
      Sullivan, E. (2019). “Understanding from machine learning models”. The British Journal for the Philosophy of Science;
      Roscher, R., Bohn, B., Duarte, M.F., & Garcke, J. (2020). “Explainable machine learning for scientific insights and discoveries”. IEEE Access, 8, 42200–42216

8. **Week: (11.01.2021) Machine Epistemology**

    - Topics: *The Problem of Induction, No Free Lunch Theorems, Bias-Variance Trade-Off, Why does Deep Learning Work so Well?*
    - Required Readings:
        - Sterkenburg, T. and Grünwald, P. (2020). “The No-Free-Lunch Theorems of Supervised Learning”. [Preprint] URL: http://philsci-archive.pitt.edu/id/eprint/18505 (accessed 2020-12-18).
    - Background Readings: POI, ch. 7;
      Geman, S., Bienenstock, E., & Doursat, R. (1992). “Neural networks and the bias/variance dilemma”. Neural Computation, 4(1), 1–58
      Wheeler, G. (2016). “Machine epistemology and big data”
      Williamson, J. (2004). “A dynamic interaction between machine learning and the philosophy of science”. Minds and Machines, 14(4), 539–549

9. **Week: (18.01.2021) Ethics of AI**

   - Topics: *Fairness, Biased Data, Transparency, Adversarial AI, "IT takes our jobs"*
   - Required Readings:
     - Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). "The ethics of algorithms: Mapping the debate". Big Data & Society, 3(2), 2053951716679679
     - Zou, J., & Schiebinger, L. (2018). "AI can be sexist and racist—it's time to make it fair"
   - Background Readings: Müller, V. (2020). "Ethics of Artificial Intelligence and Robotics", The Stanford Encyclopedia of Philosophy (Winter 2020 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/win2020/entries/ethics-ai/;
     Danks, D., & London, A.J. (2017, August). "Algorithmic Bias in Autonomous Systems". In: IJCAI, pp. 4691–4697;
     Hardt, M., Price, E., & Srebro, N. (2016). "Equality of opportunity in supervised learning". In: Advances in neural information processing systems, pp. 3315-3323;
     Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., & Vertesi, J. (2019, January). "Fairness and abstraction in sociotechnical systems". In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 59–68

10. **Week: (25.01.2021) Decision Theory and AI**

    - Topics: *Perfectly/Bounded Rational Agents, Assemble Learning, Performance Measures, Exploration & Exploitation*
    - Required Readings:
      - Russell, S.J. (1997). "Rationality and intelligence". Artificial Intelligence, 94(1–2), 57–77
      - Christiano, P. (2016). "The reward engineering problem", https://ai-alignment.com/the-reward-engineering-problem-30285c779450
    - Background: AIMA, chs. 2.1–2.4;
      Simon, H. A. (1990). "Bounded rationality". In: Utility and Probability. Palgrave Macmillan, London, pp. 15–18;
      Russell, S.J., & Subramanian, D. (1994). "Provably bounded-optimal agents". Journal of Artificial Intelligence Research, 2, 575–609;
      S. Russell on the value alignment problem, https://www.youtube.com/watch?v=WvmeTaFc_Qw

11. **Week: (01.02.2021) Philosophy of Science and AI**

    - Topics: *Is AI a Science?, Relation AI and Philosophy of Science, Limits of AI*
    - Required Readings:
      - Simon, H.A. (1995). "Artificial intelligence: an empirical science". Artificial Intelligence, 77(1), 95–127
      - Korb, K.B. (2001). "Machine learning as philosophy of science". In: Proceedings of the ECML-PKDD-01 Workshop on Machine Learning as Experimental Philosophy of Science, Freiburg
    - Background Reading: SEP-AI, sec. 9

12. **Week: (08.02.2021) 5' Pitches of Essay Ideas and Feedback on Essay Topics**

    - Topics: *Your Ideas*
    - Required Readings: None