Explainable Artificial Intelligence

Winter 2021/22

LMU Munich, Faculty of Philosophy, Philosophy of Science and the Study of Religion, MCMP
&
Faculty of Mathematics, Informatics and Statistics, Statistical Learning and Data Science

Instructors:	Timo Freiesleben	Gunnar König
Email:	Timo.Freiesleben@campus.lmu.de	gunnar.koenig@stat.uni-muenchen.de

Location: Room 021 in Ludwigsstrasse 31
Time: Fridays 12:15-13:45 pm
Material: The readings and further material can be found in the dropbox:

- https://www.dropbox.com/sh/j64v2jc3rttb0cc/AACmTVi_snZ3vX7mNnjaKkmfa?dl=0

Office hours: By appointment

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

- **IML:** Molnar, C. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/.
- **XAI-Social:** Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. arXiv preprint arXiv:1706.07269.
- **CAUS:** Pearl, J. (2009). Causality. Cambridge university press.
- **SEP-SE:** Woodward, James and Lauren Ross (2021), "Scientific Explanation", The Stanford Encyclopedia of Philosophy (Summer 2021 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/.

Overview: Modern Machine Learning (ML) algorithms are considered to be black-boxes we have no epistemic access to. XAI tackles this assumption by providing methods that allow gaining insights into the behavior of ML algorithms. This course introduces the central philosophical concepts and challenges in XAI such as explanation, interpretability, and opacity. Moreover, we discuss state-of-the-art XAI methods and their strengths and weaknesses. We focus particularly on causal explanations, the role of XAI for Science, and model-agnostic interpretation techniques.

Objectives: This course introduces students to the philosophical topics discussed and the methods used in the field of explainable artificial intelligence. By the end of the course, students should have an understanding of the philosophical problems around XAI and technical attempts to overcome them. Moreover, students should be able to outline conceptual problems of the field and discuss them critically and in-depth.

Prerequisites: Basic knowledge about concepts from Philosophy of Science (e.g. explanation and causality) is desirable. Moreover, an affinity to mathematics is necessary to understand all concepts discussed. Ideally, participants have also some basic understanding of what machine learning is and does. The seminar will be held in English.

Coursework: All students are requested to attend the seminar sessions, carefully study the reading assignments, and participate in the discussions. Dependent on the subject and the ECTS points, students must fulfill additional requirements:

- *Philosophy 9 ECTS:* Final Essay of 4500 words (± 250 words), presentation 15 minutes, topic: come up with topic/argument yourself. Final grade (80% essay, 20% presentation).
- *Statistics 9 ECTS:* Final Essay of 15 pages (à 1.500 letters), presentation 35 minutes, topic: paper selected by us. Final grade (60% essay, 40% presentation).
- *Data Science 6 ECTS:* Final Essay of 15 pages (à 1.500 letters), topic: paper selected by us. Final grade (100% essay).
- *Statistics 3 ECTS:* Video 15-20 minutes, topic: present a currently used XAI package in a short video. Final grade (100% video).

Final Essay: The deadline for submitting the essay is **21.03.2022** noon. For Philosophy students: do not forget to register in the LSF for the exam (In our case there is no exam but only the essay and the presentation) between **17.01.-28.01.2022**.

Presentation: The last three sessions of the seminar are devoted to the presentation of your own project ideas. The topic of the presentation is the same as the topic of your final essay. The length and the topic of the presentation depend on your subject as specified above. The presentation is not aimed to show off your great elaborate skills and sophistication, instead, try to present your work in a way that all people in the audience can follow your thoughts and get the upshot of your argument/topic. It is better to focus on the main points than getting lost in the details. If possible, illustrate your ideas with examples.

Cool Additional Resources: If you are interested in knowing more about ML, we recommend the free online introduction to machine learning (i2ml) course `https://introduction-to-machine-learning.netlify.app/` from our very own statistics department at the LMU, particularly by the chair of Prof. Bernd Bischl. We can also recommend Stanford University's free online course on machine learning `https://www.coursera.org/learn/machine-learning?#syllabus` taught by Andrew Ng. Both courses provide an intro to regression methods, neural networks, and unsupervised techniques such as k-means in a highly accessible way.
If you are new to scientific writing, we highly recommend the free course `https://www.coursera.org/learn/sciwrite?` taught by Kristin Sainani. Amongst other things, it shows how to make your English writing more fluent, accessible, and efficient.
You are free to write your final essay in word or any other word processing software. However, some of you might be interested in trying out latex. Latex eases writing formulas, graphs, making citations, and general formatting - you will love it! In the dropbox folder, you find a basic example that contains all the crucial elements and makes starting latex a lot easier.
How to work with latex?: It is easiest to work on `overleaf.com`. Register there and start a new project. There is a button to upload files. Choose the three files from dropbox and click on compile. The syntax of latex is very intuitive, but you have to play around with it to see how it works. Concerning the bib file: You can copy these directly from google scholar. Open google scholar, search for the paper you are interested in and click on the quotation marks beneath the paper. Choose bibtex and copy the item into your bib file.

Topics

1. **Week: (22.10.2021) A (very) short intro to ML**
   - Topics: *Algorithms, Models, Learning-Paradigms, Supervised-Learning, State-of-the-art*
   - Required Reading: None
   - Background Readings:
     - Buckner, C. (2019) Deep learning: A philosophical introduction. Philosophy Compass.; 14:e12625. https://doi.org/10.1111/phc3.12625
     - Alpaydin, E. (2020). Introduction to machine learning. MIT press.
     - Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
     - Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.
2. **Week: (29.10.2021) What is XAI about?**
   - Topics: *In what sense is ML opaque and for who?, What do we explain and to whom?, Why do we need explanations?*
   - Required Readings:
     - Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3), 31-57.
   - Background Readings:
     - IML, ch. 1+2
     - Boge, F.J. (2021) Two Dimensions of Opacity and the Deep Learning Predicament. Minds & Machines. https://doi.org/10.1007/s11023-021-09569-4
     - Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
     - Molnar, C., Casalicchio, G., & Bischl, B. (2019). Quantifying model complexity via functional decomposition for better post-hoc interpretability. arXiv preprint arXiv:1904.03867.
3. **Week: (05.11.2021) Philosophical and Psychological Accounts of Explanation**
   - Topics: *What means A explains B (Deductive-Nomological, Statistical Relevance, Models, Counterfactual, and Pragmatic), How do people explain behavior?, How do people select and evaluate explanations?*
   - Required Readings:
     - Mittelstadt, B., Russell, C., & Wachter, S. (2019, January). Explaining explanations in AI. In Proceedings of the conference on fairness, accountability, and transparency (pp. 279-288).
   - Background Readings:
     - SEP-SE
     - XAI-Social
     - Bokulich, A. (2011). How scientific models can explain. Synthese, 180(1), 33-45.
     - Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). Minds and Machines, 29(3), 441-459.
     - Bokulich, A. (2014). How the tiger bush got its stripes:'how possibly'vs.'how actually'model explanations. The Monist, 97(3), 321-338.
     - Woodward, J. (2004). Counterfactuals and causal explanation. International Studies in the Philosophy of Science, 18(1), 41-72.
     - Byrne, R. M. (2019, August). Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In IJCAI (pp. 6276-6282).

4. **Week: (12.11.2021) Causal Explanations**

   - Topics: *Causal explanations as a gold-standard, Short history, ladder of causation, Bayesian Networks, Structural Causal Models, Interventions and Counterfactuals, Actual Causation*
   - Required Readings:
     - Reutlinger, A., & Saatsi, J. (Eds.). (2018). Explanation beyond causation: philosophical perspectives on non-causal explanations. Oxford University Press.
   - Background Readings:
     - CAUS
     - Woodward, J. (2005). Making things happen: A theory of causal explanation. Oxford university press.
     - Pearl, J., & Mackenzie, D. (2018). The book of why: the new science of cause and effect. Basic books.
     - Spirtes, P. (2010). "Introduction to causal inference". Journal of Machine Learning Research, 11(5);
     - Lagnado, D. A., Gerstenberg, T., & Zultan, R. I. (2013). Causal responsibility and counterfactuals. Cognitive science, 37(6), 1036-1073.

5. **Week: (19.11.2021) Representation and Interpretable Models**

   - Topics: *Representation, Interpretability of (statistical) models, Linear Models, Decision Trees, Statistical Inference*
   - Required Readings:
     - Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3), 199-231.
   - Background Readings:
     - IML ch. 5
     - Shmueli, G. (2010). To explain or to predict?. Statistical science, 25(3), 289-310.
     - Bailer-Jones, D. M. (2003). When scientific models represent. International studies in the philosophy of science, 17(1), 59-74.
     - Frigg R., Nguyen J. (2017) Models and Representation. In: Magnani L., Bertolotti T. (eds) Springer Handbook of Model-Based Science. Springer Handbooks. Springer, Cham. https://doi.org/10.1007/978-3-319-30526-4-3;
     - Romeijn, Jan-Willem, "Philosophy of Statistics", The Stanford Encyclopedia of Philosophy (Spring 2017 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2017/entries/statistics/.
     - Contessa, G. (2007). Scientific representation, interpretation, and surrogative reasoning. Philosophy of science, 74(1), 48-68.

6. **Week: (26.11.2021) Model-Agnostic 1: Global Explanations**

   - Topics: *Global vs local, What do we want to explain: data, model, or process?, Effects vs Importance: PDP vs PFI, Surrogate Models, Problems of global methods*
   - Required Readings:
     - Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M. & Bischl, B. (2020). General pitfalls of model-agnostic interpretation methods for machine learning models. arXiv preprint arXiv:2007.04131.
   - Background Readings:
     - IML ch. 5.1-5.6

  - Chen, H., Janizek, J. D., Lundberg, S., & Lee, S. I. (2020). True to the Model or True to the Data?. arXiv preprint arXiv:2006.16234.

7. **Week: (03.12.2021) Model-Agnostic 2: Local Explanations**

   - Topics: *Counterfactuals, Adversarials, LIME, Shapley Values, Problems with local explanations*
   - Required Reading:
     - 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:
     - Chromik, M., Eiband, M., Buchner, F., Krüger, A., & Butz, A. (2021, April). I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In 26th International Conference on Intelligent User Interfaces (pp. 307-317).
     - Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
     - Freiesleben, T. (2020). The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples.
     - König, G., Freiesleben, T., & Grosse-Wentrup, M. (2021). A Causal Perspective on Meaningful and Robust Algorithmic Recourse. arXiv preprint arXiv:2107.07853.

8. **Week: (10.12.2021) Explaining Neural Networks**

   - Topics: *the limitations of model-agnostic methods, Feature Attributions, Conceptual Bottleneck, Network Dissection*
   - Required Readings:
     - Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. Distill, 5(3), e00024-001. Online accessible via: https://distill.pub/2020/circuits/zoom-in/
   - Background Readings:
     - IML ch. 10
     - Buckner, C. (2018). "Empiricism without magic: Transformational abstraction in deep convolutional neural networks". Synthese, 195(12), 5339–5372.
     - Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020, November). Concept bottleneck models. In International Conference on Machine Learning (pp. 5338-5348). PMLR.
     - Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6541-6549).

9. **Week: (17.12.2021) XAI in Science**

   - Topics: *How are ML models used in Science?, ML models as scientific models, inference with ML, the role of XAI in science, The significance of XAI for Philosophy of Science*
   - Required Readings:
     - Sullivan, E. (2019). "Understanding from machine learning models". The British Journal for the Philosophy of Science;
   - Background Readings:

  - Molnar, C., Freiesleben, T., König, G., Casalicchio, G., Wright, M. N., & Bischl, B. (2021). Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process. arXiv preprint arXiv:2109.01433.
  - Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. Trends in cognitive sciences, 23(4), 305-317.
  - Roscher, R., Bohn, B., Duarte, M.F., & Garcke, J. (2020). "Explainable machine learning for scientific insights and discoveries". IEEE Access, 8, 42200–42216;
  - Werner, M., Junginger, A., Hennig, P., & Martius, G. (2021). Informed Equation Learning. arXiv preprint arXiv:2105.06331.

10. **Week: (14.01.2022) XAI, Ethics of AI, and Fairness**

    - Topics: *Fairness and XAI, Ethics and XAI, Transparency, Adversarial AI, Legal aspects of XAI*
    - Required Readings:
      - Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In 2018 ieee/acm international workshop on software fairness (fairware) (pp. 1-7). IEEE.
    - Background Readings:
      - Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. arXiv preprint arXiv:1706.02744.
      - Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236.
      - Kusner, M.J., Loftus, J.R., Russell, C., & Silva, R. (2017). Counterfactual Fairness. NIPS.
      - 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
      - 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"

11. **Week: (21.01.2022) Presentation Block 1: Philosophical Foundations**

    - Your Ideas

12. **Week: (28.01.2022) Presentation Block 2: XAI Methods**

    - Your Ideas

13. **Week: (04.02.2022) Presentation Block 3: XAI, Science, and Society**

    - Your Ideas

14. **Week: (11.02.2022) XAI Special**