Erklärbare Künstliche Intelligenz

Wintersemester 2026/27

Otto-von-Guericke-Universität Magdeburg, Bereich Philosophie

Dozent: Timo Freiesleben
E-Mail: timo.freiesleben@ovgu.de

Ort: G40B-414
Zeit: Donnerstags, 15–17 Uhr c.t.
Material: Die Texte und weitere Materialien finden Sie unter: https://elearning.ovgu.de/course/view.php?id=22302
Sprechstunde: Während der Vorlesungszeit donnerstags, 14–15 Uhr, oder nach Vereinbarung per E-Mail

Hintergrundliteratur: Die folgenden Bücher und Überblicksartikel werden im Seminar immer wieder herangezogen und sind für die Vor- und Nachbereitung hilfreich. Alle sind frei zugänglich.

- **SML:** Freiesleben, T., & Molnar, C. (2024). *Supervised Machine Learning for Science: How to Stop Worrying and Love Your Black Box*. Online-Buch, https://ml-science-book.com/.
- **IML:** Molnar, C. (2025). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 3. Aufl., https://christophm.github.io/interpretable-ml-book/.
- **PSML:** Durán, J.M., & Pozzi, G. (Hg.) (2026). *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*. Synthese Library 527. Springer (Open Access), https://doi.org/10.1007/978-3-032-03083-2.
- **SEP-SE:** Woodward, J., & Ross, L. „Scientific Explanation“. In: E.N. Zalta & U. Nodelman (Hg.): *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/scientific-explanation/.

Überblick: Hinter neuesten wissenschaftlichen Durchbrüchen wie AlphaFold oder GraphCast steht maschinelles Lernen. Die dabei eingesetzten Modelle leiden jedoch häufig unter einem grundlegenden Problem: Ihre Entscheidungen und die internen Prozesse, die zu diesen Entscheidungen führen, sind für Menschen weitgehend intransparent. In den vergangenen zehn Jahren hat das Forschungsfeld der Erklärbaren Künstlichen Intelligenz (XAI) versucht, diese Intransparenz mithilfe verschiedener Techniken zu überwinden und das Verhalten von KI-Systemen verständlich zu machen.
Dieses Seminar betrachtet XAI aus einer wissenschaftstheoretischen Perspektive. Inwiefern erzeugen XAI-Techniken tatsächlich Erklärungen? Wird durch sie lediglich das Verhalten eines Modells erklärt, oder ermöglichen sie auch einen Zugang zu den zugrunde liegenden Phänomenen? Wie lässt sich die Korrektheit von XAI-Erklärungen evaluieren? Und inwiefern ist es Menschen überhaupt möglich, die internen Prozesse und das „Denken“ von KI-Systemen zu verstehen?
Um diese Fragen fundiert zu diskutieren, werden wir verschiedene wissenschaftliche Methodologien und Ansätze der XAI untersuchen, darunter modellagnostische Verfahren und mechanistische Interpretierbarkeit. In den abschließenden Sitzungen wenden wir uns den weiterreichenden Konsequenzen zu: Was bedeutet es für unser Verständnis und unseren Umgang mit neuen wissenschaftlichen Durchbrüchen, wenn XAI an der Aufgabe scheitert, die Funktionsweise komplexer KI-Systeme verständlich und wissenschaftlich belastbar zu erklären?

Aufbau: Das Seminar gliedert sich in fünf Blöcke. Der erste Block (Sitzungen 1–3) führt in das maschinelle Lernen in der Wissenschaft ein und fragt, ob und wie opake Modelle wissenschaftliches Verstehen ermöglichen können. Der zweite Block (Sitzungen 4–6) klärt die Begriffe Interpretierbarkeit und Erklärung, bringt die

Debatten der Informatik mit Theorien wissenschaftlicher Erklärung zusammen und fragt, was eine gute Erklärung ausmacht und wie sich XAI-Erklärungen evaluieren lassen. Der dritte Block (Sitzungen 7–8) untersucht, wann modellagnostische und kausale Verfahren über das Modell hinaus Rückschlüsse auf die zugrunde liegenden Phänomene erlauben. Der vierte Block (Sitzungen 9–10) widmet sich der konzeptbasierten und mechanistischen Interpretierbarkeit, also dem Versuch, die internen Repräsentationen und Mechanismen neuronaler Netze offenzulegen. In der letzten Sitzung fragen wir, was es für die Wissenschaft bedeutet, wenn sich opake Modelle nicht vollständig erklären lassen.

Lernziele: Am Ende des Seminars sollen die Teilnehmenden

- zentrale Begriffe wie Opazität, Interpretierbarkeit, Erklärung und Repräsentation unterscheiden und auf maschinelles Lernen anwenden können,
- die Funktionsweise, Annahmen und Grenzen wichtiger XAI-Ansätze (modellagnostische Verfahren, konzeptbasierte Ansätze, mechanistische Interpretierbarkeit) erläutern können und
- kritisch beurteilen können, unter welchen Bedingungen XAI-Ergebnisse Rückschlüsse auf ein Modell oder auf reale Phänomene erlauben.

Voraussetzungen: Im Seminar werden wir verschiedene Forschungspapiere lesen, die in englischer Sprache verfasst sind. Gute Englischkenntnisse sind daher von Vorteil; das Seminar selbst findet jedoch auf Deutsch statt. Grundlagen der Wissenschaftstheorie, insbesondere zu Fragen des wissenschaftlichen Erklärens und Verstehens, sind hilfreich, werden im Seminar aber auch noch einmal eingeführt. Da wir uns auch mit der technischen Seite von XAI beschäftigen werden, ist eine gewisse Affinität zur Mathematik von Vorteil.

Erwartungen: Von allen Teilnehmenden wird erwartet, dass sie die Pflichtlektüre vor jeder Sitzung sorgfältig lesen, sich aktiv an der Diskussion beteiligen und zu jeder Sitzung zwei Fragen zur Pflichtlektüre vorbereiten. Das können kritische Einwände ebenso wie Verständnisfragen sein.

Präsentation: Alle Studierenden halten in einer der Seminarsitzungen eine 15-minütige Präsentation. Die Präsentation soll in die zentralen Ideen des Textes einführen, der in der jeweiligen Sitzung besprochen wird. Je nach Teilnehmendenzahl können die Präsentationen auch in Gruppen gehalten werden.

Essay: Wer mehr Leistungspunkte erwerben möchte, schreibt zusätzlich einen Essay. Das Thema wählen Sie selbst; es muss aber an die Inhalte des Seminars anknüpfen und vorab kurz mit mir abgesprochen werden. Ein guter Essay vertritt eine klare These und verteidigt sie gegen naheliegende Einwände. Der Essay ist als Einzelarbeit zu verfassen; der Umfang richtet sich nach der Zahl der Leistungspunkte:

- 4 ECTS: 2.500 Wörter (±5 %)
- 5 ECTS: 3.000 Wörter (±5 %)
- 6 ECTS: 3.500 Wörter (±5 %)

Abgabefrist ist der **15.03.2027, 12 Uhr**, per E-Mail als PDF an timo.freiesleben@ovgu.de. Bitte denken Sie daran, sich fristgerecht im Prüfungsportal der OVGU für die Prüfungsleistung anzumelden.

Gewichtung:

- Für Teilnehmende, die den Kurs für 2 ECTS belegen, ergibt sich das Bestehen bzw. die Note zu 100 % aus der Präsentation.
- Für Teilnehmende, die den Kurs für 4, 5 oder 6 ECTS belegen, ergibt sich das Bestehen bzw. die Note zu 30 % aus der Präsentation und zu 70 % aus dem Essay.

Nutzung von KI-Werkzeugen: Ich bin der Überzeugung, dass Studierende alle verfügbaren Ressourcen nutzen sollten, um einen guten Essay zu schreiben und das Seminar erfolgreich abzuschließen. Dazu gehört auch der verantwortungsvolle Einsatz von KI-Werkzeugen. Sie können KI zum Beispiel gerne nutzen, um

(i) sich technische Konzepte aus dem Seminar zu erschließen, (ii) kritisches Feedback zu Ihrem Manuskript zu erhalten, (iii) Ihre Texte klarer und flüssiger zu formulieren und (iv) nach relevanter wissenschaftlicher Literatur zu suchen, mit der Sie sich auseinandersetzen können. Behalten Sie bei der Nutzung von KI-Werkzeugen bitte deren Grenzen im Blick. KI-generierte Informationen können unvollständig, veraltet oder falsch sein. Es liegt daher in Ihrer Verantwortung, alle Behauptungen, Literaturangaben und Erklärungen, die Sie erhalten, sorgfältig zu überprüfen.
Ich ermutige Sie zwar, KI überlegt zur Unterstützung Ihres Schreibprozesses einzusetzen, Ihr Essay muss letztlich aber Ihre eigene Arbeit sein. Von KI generierten Text einzureichen, der in die Arbeit kopiert wurde, ist nicht zulässig! Am Ende des Semesters nehme ich mir für alle Teilnehmenden, die einen Essay geschrieben haben, 15 Minuten Zeit, um den Essay gemeinsam zu besprechen und Ihnen Feedback zu geben. Sollte sich dabei zeigen, dass Sie die Argumentation Ihres eigenen Essays nicht verstehen, wird das Seminar als nicht bestanden bewertet.

Weiterführende Ressourcen: Eine zugängliche Einführung in maschinelles Lernen in der Wissenschaft bietet unser Buch SML (siehe oben); eine philosophische Einführung in maschinelles Lernen gibt Buckner, C. (2019). „Deep learning: A philosophical introduction“. *Philosophy Compass*, 14(10), e12625 (https://doi.org/10.1111/phc3.12625). Für eine etwas technischere Einführung empfehle ich den Coursera-Kurs von Andrew Ng (https://www.coursera.org/learn/machine-learning). Eine fundiertere theoretische Grundlage im maschinellen Lernen bietet das frei verfügbare Buch von Hardt, M., & Recht, B. (2022). *Patterns, Predictions, and Actions: Foundations of Machine Learning*. Princeton University Press (https://mlstory.org/). Einen umfassenden Überblick über modellagnostische Interpretationsverfahren gibt IML von Christoph Molnar (siehe oben). Wer tiefer in die mechanistische Interpretierbarkeit einsteigen möchte, findet im *Circuits*-Thread auf Distill (https://distill.pub/2020/circuits/) und im *Transformer Circuits Thread* (https://transformer-circuits.pub/) zahlreiche frei zugängliche, anschaulich aufbereitete Arbeiten. Einen philosophischen Einstieg in die mechanistische Interpretierbarkeit bieten Herrmann, D.A., & Levinstein, B.A. (2026). *Radical AI Interpretability*. arXiv:2606.26523 (https://arxiv.org/abs/2606.26523) sowie Millière, R., & Buckner, C. (2024). „A philosophical introduction to language models – Part II: The way forward“. arXiv:2405.03207 (https://arxiv.org/abs/2405.03207).

Sitzungsplan

Block I: Maschinelles Lernen in der Wissenschaft

1. (05.11.2026) Einführung: Maschinelles Lernen in der Wissenschaft

- Themen: *Vorstellungsrunde; Ziele und Anforderungen des Seminars; Was ist (überwachtes) maschinelles Lernen?; Fallbeispiele AlphaFold und GraphCast; Warum Erklärbarkeit?*
- Hintergrund: SML, Kap. 1–3;
  Jumper, J., et al. (2021). „Highly accurate protein structure prediction with AlphaFold". *Nature*, 596(7873), 583–589;
  Lam, R., et al. (2023). „Learning skillful medium-range global weather forecasting". *Science*, 382(6677), 1416–1421;
  Wang, H., et al. (2023). „Scientific discovery in the age of artificial intelligence". *Nature*, 620(7972), 47–60

2. (12.11.2026) Verstehen mit opaken Modellen?

- Themen: *Epistemische Opazität; Erklären vs. Verstehen; „Link Uncertainty"; Unter welchen Bedingungen ermöglichen Black-Box-Modelle ein Verstehen von Phänomenen?*
- Pflichtlektüre:
  - Sullivan, E. (2022). „Understanding from machine learning models". *The British Journal for the Philosophy of Science*, 73(1), 109–133
- Hintergrund: Räz, T., & Beisbart, C. (2024). „The importance of understanding deep learning". *Erkenntnis*, 89(5), 1823–1840;
  Creel, K.A. (2020). „Transparency in complex computational systems". *Philosophy of Science*, 87(4), 568–589;
  Boge, F.J. (2022). „Two dimensions of opacity and the deep learning predicament". *Minds and Machines*, 32(1), 43–75;
  Beisbart, C. (2026). „In which ways is machine learning opaque?". In: PSML, Kap. 1

3. (19.11.2026) Tiefe neuronale Netze als wissenschaftliche Modelle

- Themen: *Vorhersagen, Erklären und Explorieren mit neuronalen Netzen; künstliche neuronale Netze als wissenschaftliche Modelle; Gibt es einen Zielkonflikt zwischen Vorhersage und Verstehen?*
- Pflichtlektüre:
  - Cichy, R.M., & Kaiser, D. (2019). „Deep neural networks as scientific models". *Trends in Cognitive Sciences*, 23(4), 305–317
- Hintergrund: Zednik, C., & Boelsen, H. (2022). „Scientific exploration and explainable artificial intelligence". *Minds and Machines*, 32(1), 219–239;
  Kriegeskorte, N. (2015). „Deep neural networks: A new framework for modeling biological vision and brain information processing". *Annual Review of Vision Science*, 1, 417–446;
  Crook, B., & Kästner, L. (2026). „Don't fear the bogeyman: On why there is no prediction-understanding trade-off for deep learning in neuroscience". In: PSML, Kap. 20

Block II: Interpretierbarkeit, wissenschaftliche Erklärung und Evaluation von Erklärungen

4. (26.11.2026) Was heißt Interpretierbarkeit?

- Themen: *Wozu Interpretierbarkeit?; Transparenz vs. post-hoc Erklärungen; Interpretierbare Modelle vs. Black Boxes; Ist Interpretierbarkeit ein einheitlicher Begriff?*
- Pflichtlektüre:

  - Lipton, Z.C. (2018). „The mythos of model interpretability“. *Queue*, 16(3), 31–57

- Hintergrund: 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;
  Krishnan, M. (2020). „Against interpretability: A critical examination of the interpretability problem in machine learning“. *Philosophy & Technology*, 33(3), 487–502;
  Freiesleben, T., & König, G. (2023). „Dear XAI community, we need to talk! Fundamental misconceptions in current XAI research“. In: L. Longo (Hg.): *Explainable Artificial Intelligence. xAI 2023.* Springer (CCIS 1901), 48–65;
  IML, Kap. 2–3

5. (03.12.2026) Interpretierbarkeit und Theorien wissenschaftlicher Erklärung

- Themen: *Was soll interpretiert werden, für wen und wozu?; Interpretierbarkeit und Erklärbarkeit; Modelle von wissenschaftlichem Erklären; Verstehen des Modells vs. Verstehen mit dem Modell; kontrastive und kontextabhängige Erklärungen; Pluralismus der Interpretierbarkeit*

- Pflichtlektüre:

  - Räz, T. (2026). „From explanations to interpretability and back“. In: PSML, Kap. 9

- Hintergrund: SEP-SE;
  Erasmus, A., Brunet, T.D.P., & Fisher, E. (2021). „What is interpretability?“. *Philosophy & Technology*, 34(4), 833–862;
  Beisbart, C., & Räz, T. (2022). „Philosophy of science at sea: Clarifying the interpretability of machine learning“. *Philosophy Compass*, 17(6), e12830;
  Páez, A. (2019). „The pragmatic turn in explainable artificial intelligence (XAI)“. *Minds and Machines*, 29(3), 441–459;
  Buijsman, S. (2022). „Defining explanation and explanatory depth in XAI“. *Minds and Machines*, 32(3), 563–584;
  Räz, T. (2024). „ML interpretability: Simple isn't easy“. *Studies in History and Philosophy of Science*, 103, 159–167

6. (10.12.2026) Was macht eine gute Erklärung aus? Die Evaluation von XAI-Erklärungen

- Themen: *Was ist eine gute Erklärung?; anwendungs-, menschen- und funktionsbasierte Evaluation; Erklärungen als kontrastiv, selektiv und sozial; Treue zum Modell (Faithfulness) vs. Plausibilität für Menschen; Stakeholder und ihre Erklärungsbedürfnisse; Gibt es eine „Ground Truth“ für XAI-Erklärungen?*

- Pflichtlektüre:

  - Doshi-Velez, F., & Kim, B. (2017). „Towards a rigorous science of interpretable machine learning“. arXiv:1702.08608

- Hintergrund: Miller, T. (2019). „Explanation in artificial intelligence: Insights from the social sciences“. *Artificial Intelligence*, 267, 1–38;
  Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). „What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research“. *Artificial Intelligence*, 296, 103473;
  Nauta, M., et al. (2023). „From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI“. *ACM Computing Surveys*, 55(13s), 1–42;
  Chromik, M., Eiband, M., Buchner, F., Krüger, A., & Butz, A. (2021). „I think I get your point, AI! The illusion of explanatory depth in explainable AI“. In: *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 307–317;
  IML, Kap. 32

Block III: Vom Modell zum Phänomen

7. (17.12.2026) Modellagnostische Verfahren und wissenschaftliche Inferenz

- Themen: *Globale und lokale modellagnostische Erklärmethoden; Modell-Audit vs. wissenschaftliche Inferenz; marginales vs. bedingtes Sampling; epistemische Unsicherheit*
- Pflichtlektüre:
  - Freiesleben, T., König, G., Molnar, C., & Tejero-Cantero, Á. (2024). „Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena". *Minds and Machines*, 34(3), 32
- Hintergrund: SML, Kap. 9; IML, Kap. 4, 14–15, 17–19 und 23;
  Chen, H., Janizek, J.D., Lundberg, S., & Lee, S.-I. (2020). „True to the model or true to the data?". arXiv:2006.16234;
  Watson, D.S. (2022). „Conceptual challenges for interpretable machine learning". *Synthese*, 200(2), 65;
  Boge, F.J., & Mosig, A. (2025). „Put it to the test: Getting serious about explanation in explainable artificial intelligence". *Minds and Machines*, 35, 26

8. (07.01.2027) Kausalität: Von Vorhersagen zu kausalen Erklärungen?

- Themen: *Kausale Erklärungen und Interventionismus; Strukturelle kausale Modelle, Interventionen und Kontrafaktische; Kausalentdeckung und Effektschätzung mit ML; Wann erlauben XAI-Verfahren kausale Schlüsse?*
- Pflichtlektüre:
  - Buijsman, S. (2023). „Causal scientific explanations from machine learning". *Synthese*, 202(6), 202
- Hintergrund: SML, Kap. 10;
  Baron, S. (2023). „Explainable AI and causal understanding: Counterfactual approaches considered". *Minds and Machines*, 33(2), 347–377;
  Beckers, S. (2022). „Causal explanations and XAI". In: *Proceedings of the First Conference on Causal Learning and Reasoning*. PMLR 177, 90–109;
  Boge, F.J., & Mosig, A. (2025). „Causality and scientific explanation of artificial intelligence systems in biomedicine". *Pflügers Archiv – European Journal of Physiology*, 477(4), 543–554;
  Zhao, Q., & Hastie, T. (2021). „Causal interpretations of black-box models". *Journal of Business & Economic Statistics*, 39(1), 272–281

Block IV: Konzeptbasierte und mechanistische Interpretierbarkeit

9. (14.01.2027) Ins Innere neuronaler Netze: Konzepte, Schaltkreise und Repräsentationen

- Themen: *Konzepte, Schaltkreise und Universalität; Polysemantizität und Superposition; Probing; Wann repräsentiert eine Komponente eines Netzes ein Konzept?*
- Pflichtlektüre:
  - Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). „Zoom in: An introduction to circuits". *Distill*, 5(3), e00024.001, https://distill.pub/2020/circuits/zoom-in/
- Hintergrund: Harding, J. (2023). „Operationalising representation in natural language processing". *The British Journal for the Philosophy of Science*, https://doi.org/10.1086/728685;
  Stecher, P.M., Sikimic, V., & Kahle, R. (2026). „Do different artificial neural networks think alike? On the feasibility of cataloging representations". *Minds and Machines*, 36(1), 17;

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). „The platonic representation hypothesis". arXiv:2405.07987;
Elhage, N., et al. (2022). „Toy models of superposition". *Transformer Circuits Thread*;
Freiesleben, T. (2026). „Artificial neural nets and the representation of human concepts". In: PSML, Kap. 13;
IML, Kap. 27 und 29

10. (21.01.2027) Mechanistische Interpretierbarkeit als mechanistische Erklärung?

- Themen: *Mechanismen und mechanistische Erklärung in den Lebenswissenschaften; „Teile und herrsche" vs. koordinierte Entdeckungsstrategien; Interventionen auf Modellkomponenten; Lässt sich das „Denken" großer KI-Modelle verstehen?*
- Pflichtlektüre:
  - Kästner, L., & Crook, B. (2024). „Explaining AI through mechanistic interpretability". *European Journal for Philosophy of Science*, 14(4), 52
- Hintergrund: Craver, C., & Tabery, J. „Mechanisms in Science". In: *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/science-mechanisms/;
  Millière, R., & Buckner, C. (2024). „A philosophical introduction to language models – Part II: The way forward". arXiv:2405.03207;
  Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). „Causal abstractions of neural networks". *Advances in Neural Information Processing Systems*, 34;
  Lindsey, J., et al. (2025). „On the biology of a large language model". *Transformer Circuits Thread*;
  Sharkey, L., et al. (2025). „Open problems in mechanistic interpretability". arXiv:2501.16496

Block V: Ausblick

11. (28.01.2027) Wissenschaft ohne Verstehen? Opazität und wissenschaftliche Entdeckung

- Themen: *Entdeckungs- vs. Begründungszusammenhang; Brauchen KI-gestützte Durchbrüche XAI?; Illusionen des Verstehens; Verstehen als Ziel der Wissenschaft; Rückblick auf das Semester*
- Pflichtlektüre:
  - Duede, E. (2023). „Deep learning opacity in scientific discovery". *Philosophy of Science*, 90(5), 1089–1099
- Hintergrund: Andrews, M. (2026). „The immortal science of ML: Machine learning and the theory-free ideal". *Erkenntnis*, 91, 2685–2707;
  Krenn, M., et al. (2022). „On scientific understanding with artificial intelligence". *Nature Reviews Physics*, 4(12), 761–769;
  Messeri, L., & Crockett, M.J. (2024). „Artificial intelligence and illusions of understanding in scientific research". *Nature*, 627(8002), 49–58;
  SML, Kap. 15