Philosophie der Künstlichen Intelligenz

Wintersemester 2026/27

Otto-von-Guericke-Universität Magdeburg, Bereich Philosophie

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

Ort: G40B-238 (Ausnahme: am 11.11.2026 findet die Sitzung in G40B-326 statt)
Zeit: Mittwochs, 15–17 Uhr c.t.
Material: Die Texte und weitere Materialien finden Sie unter: https://elearning.ovgu.de/course/view.php?id=22301
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.

- **AIMA:** Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*. 4. Aufl. Pearson. Dt. Übersetzung der 3. Aufl.: *Künstliche Intelligenz: Ein moderner Ansatz*. Pearson Studium, 2012.
- **SEP-KI:** Bringsjord, S., & Govindarajulu, N.S. „Artificial Intelligence". In: E.N. Zalta & U. Nodelman (Hg.): *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/artificial-intelligence/.
- **DLRM:** Buckner, C. (2024). *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence*. Oxford University Press.

Überblick: Dieses Seminar betrachtet künstliche Intelligenz aus einer erkenntnistheoretischen und wissenschaftstheoretischen Perspektive. Wir werden Fragen stellen wie: Sollte das Forschungsfeld der KI durch seine Ziele, Methoden oder die Probleme charakterisiert werden, die es zu lösen versucht? Sollten intelligente Maschinen lediglich korrekt *handeln*, oder müssen sie auch korrekt *schlussfolgern*? Und was bedeutet „Korrektheit" in diesem Zusammenhang: Sollten Maschinen an formalen Theorien der Rationalität gemessen werden, oder sind menschliche Fähigkeiten und menschliches Verhalten der relevante Maßstab?
Wir werden diskutieren, inwiefern unterschiedliche Antworten auf diese Fragen die moderne KI-Forschung beeinflusst haben und noch immer beeinflussen, von logikbasierter KI über den Konnektionismus bis hin zu modernem Reinforcement Learning durch menschliches Feedback. Abschließend werden wir untersuchen, wie diese unterschiedlichen Perspektiven zu verschiedenen Maßstäben für die Bewertung des Fortschritts hin zu einer künstlichen allgemeinen Intelligenz (AGI) führen, und betrachten, was die jüngsten Erfolge großer Sprachmodelle darüber (nicht) aussagen können, wie nah wir einer AGI sind.

Aufbau: Das Seminar orientiert sich grob an der bekannten Vierteilung von Russell & Norvig: KI als *menschlich handelnde*, *menschlich denkende*, *rational denkende* und *rational handelnde* Systeme. Der erste Block (Sitzungen 1–3) fragt, was KI ist und ob menschenähnliches Verhalten als Kriterium ausreicht. Der zweite Block (Sitzungen 4–6) behandelt die Ideale formaler Rationalität und ihre Grenzen. Der dritte Block (Sitzungen 7–9) verfolgt den Weg vom Konnektionismus über Lernfähigkeit als Kern von Intelligenz bis zur Frage, woher die Ziele und Werte lernender Systeme stammen. In der letzten Sitzung wenden wir die erarbeiteten Maßstäbe auf große Sprachmodelle und die Frage nach AGI an.

Lernziele: Am Ende des Seminars sollen die Teilnehmenden

- zentrale Positionen zur Bestimmung von „Intelligenz" und „KI" kennen und gegeneinander abwägen können,

- die philosophischen Annahmen hinter symbolischer KI, Konnektionismus und Reinforcement Learning benennen können und
- verschiedene Maßstäbe (Verhalten, innere Prozesse, formale Rationalität, Lernfähigkeit) für die Bewertung von KI-Systemen sowie aktuelle Behauptungen über allgemeine KI kritisch beurteilen können.

Voraussetzungen: Technische Vorkenntnisse sind nicht erforderlich; die nötigen Grundlagen zu Logik, neuronalen Netzen und Reinforcement Learning werden im Seminar eingeführt. Grundkenntnisse in theoretischer Philosophie (Erkenntnistheorie, Wissenschaftstheorie oder Philosophie des Geistes) sind hilfreich. Die meisten Texte sind auf Englisch; die Diskussion findet auf Deutsch statt.

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. Bei längeren Texten wird zu Beginn des Semesters angegeben, welche Abschnitte im Fokus stehen.

Präsentation: Jede*r Studierende hält 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: Wer sich die Grundlagen von KI intuitiv und anschaulich erarbeiten möchte, findet in der Videoreihe *Crash Course Artificial Intelligence* (https://tinyurl.com/AIcrash-course) kurze Einführungen zu Themen wie symbolischer KI, neuronalen Netzen und Reinforcement Learning. 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/). Für KI im Allgemeinen ist AIMA (siehe oben; https://aima.cs.berkeley.edu/) das Standardwerk.

Sitzungsplan

Block I: Was ist KI und woran misst man sie?

1. (04.11.2026) Einführung: Ideengeschichte der KI

- Themen: *Vorstellungsrunde; Ziele und Anforderungen des Seminars; Ideengeschichte der KI; Wie lässt sich KI definieren?*
- Hintergrund: AIMA, Kap. 1.1–1.5; SEP-KI, Abschn. 1

2. (11.11.2026) Menschlich handeln: Der Turing-Test

- Themen: *Das Imitationsspiel; operationale Definitionen von Intelligenz; Turings Antworten auf Einwände; Turing-Tests heute*
- Pflichtlektüre:
  - Turing, A.M. (1950). „Computing machinery and intelligence". *Mind*, 59(236), 433–460
- Hintergrund: Oppy, G., & Dowe, D. „The Turing Test". In: *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/turing-test/;
  Sterrett, S.G. (2000). „Turing's two tests for intelligence". *Minds and Machines*, 10(4), 541–559.
  French, R.M. (1990). „Subcognition and the limits of the Turing test". *Mind*, 99(393), 53–65;
  Jones, C.R., & Bergen, B.K. (2026). „Large language models pass a standard three-party Turing test". *Proceedings of the National Academy of Sciences*, 123(21)

3. (18.11.2026) Reicht menschliches Verhalten? Behaviorismus und seine Kritiker

- Themen: *Psychologismus vs. Behaviorismus; das „Blockhead"-Argument; starke vs. schwache KI; das Chinesische Zimmer*
- Pflichtlektüre:
  - Block, N. (1981). „Psychologism and behaviorism". *The Philosophical Review*, 90(1), 5–43
- Hintergrund: Searle, J.R. (1980). „Minds, brains, and programs". *Behavioral and Brain Sciences*, 3(3), 417–424 (nur Hauptartikel);
  Boden, M.A. (1988). „Escaping from the Chinese Room". In: *Computer Models of Mind: Computational Approaches in Theoretical Psychology*. Cambridge University Press.
  Cole, D. „The Chinese Room Argument". In: *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/chinese-room/;
  AIMA, Kap. 27

Block II: Rational denken, rational handeln

4. (25.11.2026) Rational denken: Logik und physische Symbolsysteme

- Themen: *Die Hypothese physischer Symbolsysteme; Wissensrepräsentation und logisches Schließen; Bayesianismus; Bayessche Netze; „Good Old-Fashioned AI"*
- Pflichtlektüre:
  - McCarthy, J., & Hayes, P.J. (1969). „Some philosophical problems from the standpoint of artificial intelligence". In: B. Meltzer & D. Michie (Hg.): *Machine Intelligence 4*. Edinburgh University Press, 463–502 (nur Abschn. 1 und 2)
- Hintergrund: Newell, A., & Simon, H.A. (1976). „Computer science as empirical inquiry: Symbols and search". *Communications of the ACM*, 19(3), 113–126;
  Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. MIT Press;
  Hartmann, S., & Sprenger, J. (2010). „Bayesian epistemology". In: S. Bernecker & D. Pritchard (Hg.): *Routledge Companion to Epistemology*. Routledge, 609–620;
  AIMA, Kap. 7–9; SEP-KI, Abschn. 3

5. (02.12.2026) Grenzen der symbolischen KI: Heideggers Erbe

- Themen: *Relevanz und Common Sense; implizites Wissen und Verkörperung; die phänomenologische Kritik an der KI; Reinforcement Learning*
- Pflichtlektüre:
  - Dreyfus, H.L. (2007). „Why Heideggerian AI failed and how fixing it would require making it more Heideggerian". *Artificial Intelligence*, 171(18), 1137–1160
- Hintergrund: Suchman, L.A. (1987). *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press (in Auszügen).
  Dennett, D.C. (1984). „Cognitive wheels: The frame problem of AI". In: C. Hookway (Hg.): *Minds, Machines and Evolution*. Cambridge University Press, 129–151;
  Shanahan, M. „The Frame Problem". In: *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/frame-problem/;
  Dreyfus, H.L. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press

6. (09.12.2026) Rational handeln: Ideale und beschränkte Rationalität

- Themen: *Entscheidungstheorie; rationale Agenten; perfekte, kalkulative und beschränkte Rationalität; die Rationalität von Heuristiken*
- Pflichtlektüre:
  - Russell, S.J. (1997). „Rationality and intelligence". *Artificial Intelligence*, 94(1–2), 57–77
- Hintergrund: Gigerenzer, G., & Brighton, H. (2009). „Homo heuristicus: Why biased minds make better inferences". *Topics in Cognitive Science*, 1(1), 107–143;
  Russell, S.J., & Subramanian, D. (1994). „Provably bounded-optimal agents". *Journal of Artificial Intelligence Research*, 2, 575–609;
  AIMA, Kap. 2

Block III: Vom Konnektionismus zum (eigenständigen) Lernen

7. (16.12.2026) Konnektionismus und Kognition

- Themen: *Neuronale Netze als idealisierte Modelle des Gehirns?; Empirismus vs. Nativismus; Abstraktion aus Erfahrung; verteilte vs. symbolische Repräsentationen; Kompositionalität*
- Pflichtlektüre:
  - Buckner, C. (2018). „Empiricism without magic: Transformational abstraction in deep convolutional neural networks". *Synthese*, 195(12), 5339–5372.
- Hintergrund: Stinson, C. (2020). „From implausible artificial neurons to idealized cognitive models: Rebooting philosophy of artificial intelligence". *Philosophy of Science*, 87(4), 590–611.
  Buckner, C., & Garson, J. (2025). „Connectionism". In: *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/connectionism/;
  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

8. (13.01.2027) Intelligenz als Lernfähigkeit?

- Themen: *Fertigkeit vs. Intelligenz; Intelligenz als Effizienz des Fertigkeitserwerbs; die Rolle von Vorwissen und Erfahrung; Grade der Generalisierung; Abstraktion und Analogiebildung*
- Pflichtlektüre:
  - Chollet, F. (2019). „On the measure of intelligence". arXiv:1911.01547 (Teil I und II)

- Hintergrund: Mitchell, M. (2021). „Abstraction and analogy-making in artificial intelligence“. *Annals of the New York Academy of Sciences*, 1505(1), 79–101.
  Lake, B.M., Ullman, T.D., Tenenbaum, J.B., & Gershman, S.J. (2017). „Building machines that learn and think like people“. *Behavioral and Brain Sciences*, 40, e253 (nur Hauptartikel, in Auszügen);
  DLRM, Kap. 1–2

9. (20.01.2027) Kann man intelligent sein, ohne seine Ziele selbst zu wählen? Ziele, Werte und Belohnung

- Themen: *Woher kommen Belohnungen?; extern vorgegebene Belohnung vs. intern konstruierter Wert; Ziele setzen vs. Ziele verfolgen; die „Reward-is-enough“-Hypothese; menschliches Feedback (RLHF) als Ersatz für eine Belohnungsfunktion; Intelligenz als Maximierung?*
- Pflichtlektüre:
  - Molinaro, G., & Collins, A.G.E. (2023). „A goal-centric outlook on learning“. *Trends in Cognitive Sciences*, 27(12), 1150–1164.
- Hintergrund: Juechems, K., & Summerfield, C. (2019). „Where does value come from?“. *Trends in Cognitive Sciences*, 23(10), 836–850
  Silver, D., Singh, S., Precup, D., & Sutton, R.S. (2021). „Reward is enough“. *Artificial Intelligence*, 299, 103535;
  Gabriel, I. (2020). „Artificial intelligence, values, and alignment“. *Minds and Machines*, 30(3), 411–437;
  Chu, J., Tenenbaum, J.B., & Schulz, L.E. (2024). „In praise of folly: Flexible goals and human cognition“. *Trends in Cognitive Sciences*, 28(7), 628–642
  Christiano, P. (2016). „The reward engineering problem“, https://ai-alignment.com/the-reward-engineering-problem-30285c779450

Block IV: Klassische Debatten im neuen Gewand

10. (27.01.2027) Sind wir schon da? Große Sprachmodelle und allgemeine KI

- Themen: *Was zeigen die Erfolge großer Sprachmodelle (nicht)?; Verstehen vs. statistische Mustererkennung; Weltmodelle; Benchmarks und ihre Validität; Rückblick auf das Semester*
- Ablauf: Alle Teilnehmenden werden einer von vier Gruppen zugeordnet und lesen den Text ihrer Gruppe. In der Sitzung werden die „Expert*innen“ in neue, gemischte Runden aufgeteilt und vertreten dort die Position ihres Textes. Abschließend führen wir die stärksten Argumente im Plenum zusammen.
- Pflichtlektüre:
  - (Gruppe 1) Wang, P. (2019). „On defining artificial intelligence“. *Journal of Artificial General Intelligence*, 10(2), 1–37
  - (Gruppe 2) Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (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*, 610–623
  - (Gruppe 3) Mitchell, M., & Krakauer, D.C. (2023). „The debate over understanding in AI's large language models“. *Proceedings of the National Academy of Sciences*, 120(13), e2215907120
  - (Gruppe 4) Chen, E.K., Belkin, M., Bergen, L., & Danks, D. (2026). „Does AI already have human-level intelligence? The evidence is clear“. *Nature*, 650, 36–40
- Hintergrund: Hendrycks, D., et al. (2025). „A definition of AGI“. arXiv:2510.18212
  Mitchell, M. (2024). „Debates on the nature of artificial general intelligence“. *Science*, 383(6689), eado7069
  Lyre, H. (2024). ünderstanding ai: Semantic grounding in large language models. arXiv preprint arXiv:2402.10992.