Philosophy of Science for Machine Learning

Seminar ML4520a, Winter 2024/25, Uni Tübingen

Instructors

Dr. Konstantin Genin [konstantin.genin@uni-tuebingen.de]
Sebastian Zezulka [sebastian.zezulka@uni-tuebingen.de]

Research Group “Ethics and Epistemology of Machine Learning”
ethics.epistemology.ai

[Alma | Moodle]

Office hours: by appointment, please get in contact if you have any questions.

Seminar Time and Location

Tuesdays, 2-4 pm c.t.

Ground Floor Lecture Hall
Maria-von-Linden-Straße 6,
72076 Tübingen

First meeting: 15.10.2024

Course Description

For most of the twentieth century, philosophers of science and researchers in artificial intelligence worked on similar problems, kept up with each other’s work, and took frequent inspiration from each other. The founding generations of AI and machine learning all had a working familiarity with core issues in philosophy of science. Although those days are behind us, the philosophical problems have not gone away. Talk of probabilities is everywhere, but their interpretation is often unclear. Appeals to simplicity are commonplace, but a clear justification is absent. The importance of values in data-scientific practice is broadly acknowledged, but their scope and bearing remain controversial. These are all perennial issues in the philosophy of science, and philosophers have developed a lot of resources for dealing with them. The premise of this course is that these resources can be fruitfully imported into machine learning. This course aims to give the student a familiarity with core issues in the philosophy of science, with an emphasis on their relevance to machine learning. It will be organized largely around what might be called the *ur*-problem for both fields: the problem of induction or, what, if not deductive validity, justifies inferences that go beyond the data that we have collected?

Course Requirements

tl;dr To earn 3 ECTS (graded/ungraded) you have to

(a) regularly attend the seminar,
(b) submit one question before each session (not handing in a question is allowed for two sessions),
(c) give one (group) presentation (of about 25 min),
(d) submit a one-page essay proposal by **January 10th, 2025**, and
(e) submit a 1,500-word essay by **31.03.2025**.

To earn 6/8/12 ECTS, you have to write an essay of 3/4/5 thousand words.

Readings

There will be two readings for every meeting. **Everyone should make an effort to read these two papers.** We will provide you with the materials, but if there is some difficulty please make an effort to find the material yourself. Class time will be divided roughly evenly between lectures, student presentations, and discussions.

Questions

You have to submit **one question** about the papers before each session. It is allowed not to submit a question for two sessions. You are free to develop questions in teams, but everyone must submit a unique question.

Please upload your questions here:

Additionally, we will provide you with some questions to guide your reading.

Presentations

A group of students will be experts for every session. This responsibility includes

- presenting the **core arguments** of both required readings(*) in about 25 min,
- preparing **2-3 questions for discussion**, and
- **fielding questions** from the rest of the class.

Readings will be assigned with regard to some degree for student preference. You will have to coordinate with your group on how to present the readings.

You must sign up for one presentation and, in doing so, for the course by 21 October 2024:

Presenters should make an effort to present the material in the readings as charitably, clearly, and succinctly as possible. Presenters may take on the extra responsibility of background reading for the material they are presenting. The presentations should last about 25 minutes, allowing for about 30 minutes of discussion. **We will make ourselves available beforehand to discuss the**

material for the presentation, please feel free to contact us before your presentation.

(*) You only have to present one of the core readings in case you're solely responsible for a session. The presentation can also be slightly shorter. Please talk to us beforehand.

Essay

There will be a 1,500-word essay (Hausarbeit) that is due at the end of the term. The exact deadline will be announced later. The subject matter is flexible and intended to answer to individual interests, but students must submit a 1-page proposal for approval and feedback by **January 10th, 2025**.

Grades

Grading is determined as follows:

- Class participation and questions: 10%
- Presentation: 45%
- Final essay: 45%

Missing class and late assignments

We recognize that occasional problems associated with illness, family emergencies, job interviews, other professors, etc. will inevitably lead to legitimate conflicts over your time. If you expect that you will be unable to turn in an assignment on time, or must be absent from a class meeting, please notify us (via email) in advance and we can agree on a reasonable accommodation. Otherwise, your grade will be penalized.

Academic Integrity

Each student is responsible for being aware of the university policies on academic integrity, including the policies on cheating and plagiarism. If you use AI assistance for more than copy-editing, you must report this with your assignment.

Classes and Readings

Class	Date	Readings
0	15.10.2024	Introduction
(Data) Science and Values		
1	22.10.2024	• Rudner, Richard. "The scientist qua scientist makes value judgments." *Philosophy of science* 20.1 (1953): 1-6. • Vredenburgh, Kate. “Fairness”. In: Justin B. Bullock, and others (eds.), *The Oxford Handbook of AI Governance*, Oxford Handbooks (2024; online edn, Oxford Academic, 2022).
2	29.10.2024	• Bright, Liam Kofi. "Du Bois’ democratic defence of the value free ideal." *Synthese* 195.5 (2018): 2227-2245. • Douglas, Heather. "Inductive risk and values in science." *Philosophy of science* 67.4 (2000): 559-579
Probability and The Problem of Induction		
3	05.11.2024	• Hájek, Alan, "Interpretations of Probability"*, The Stanford Encyclopedia of Philosophy* (Winter 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.) • Cynthia Dwork (2022). Fairness, Randomness and the Crystal Ball.
4	12.11.2024	• David Hume (1748/1777). An enquiry concerning human understanding. **(Sections 1-7).** • Henderson, Leah, "The Problem of Induction", *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.)
5	19.11.2024	• Von Luxburg, Ulrike, and Bernhard Schölkopf. "Statistical learning theory: Models, concepts, and results." *Handbook of the History of Logic*. Vol. 10. North-Holland, 2011. 651-706. • Sterkenburg, Tom F., and Peter D. Grünwald. "The no-free-lunch theorems of supervised learning." *Synthese* 199.3 (2021): 9979-10015.
Confirmation-Theoretic Responses to Hume		
6	26.11.2024	• Hempel, Carl G. "Studies in the Logic of Confirmation (I.)." *Mind* 54.213 (1945): 1-26. • Carnap, Rudolf. "On inductive logic." *Philosophy of*

		Science 12.2 (1945): 72-97.
7	03.12.2024	• Sprenger, Jan, and Stephan Hartmann. “Variation 1: Confirmation and Induction”. In: *Bayesian Philosophy of Science*. Oxford University Press, (2019). • Glymour, Clark. "Instrumental probability." *The Monist* 84.2 (2001): 284-300.
Falsification-Theoretic Responses		
8	10.12.2024	• Karl Popper. Logic of Scientific Discovery. **Part I.** (1934). • Mayo, Deborah G., and Aris Spanos. "Severe testing as a basic concept in a Neyman–Pearson philosophy of induction." *The British Journal for the Philosophy of Science* (2006).
Learning-Theoretic Responses		
9	17.12.2024	• Kelly, Kevin, and Clark Glymour. "Why probability does not capture the logic of scientific justification." (2004). • Lin, Hanti. "Modes of convergence to the truth: Steps toward a better epistemology of induction." *The Review of Symbolic Logic* 15.2 (2022): 277-310.
		No Class, Winter Break
		No Class, Winter Break
Realism/Anti-realism		
10	07.01.2025	• Henderson, Leah. "Global versus local arguments for realism." *The Routledge handbook of scientific realism*. Routledge. (2017): 151-163. • Breiman, Leo. "Statistical modeling: The two cultures." *Quality control and applied statistics* 48.1 (2003): 81-82.
Measurement		
11	14.01.2025	• Chang, Hasok and Nancy Cartwright. “Measurement”. In: S. Psillos, and M. Curd (eds), *The Routledge Companion to Philosophy of Science.* New York: Routledge. (2008): 367–375. • Tal, Eran. "Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare." *Proceedings of the 2023 AAAI/ACM Conference*

		on AI, Ethics, and Society. (2023).
(Scientific) Explanation		
12	21.01.2025	● Wesley Salmon. “Scientific Explanation.” In: Merrilee H. Salmon et al. (eds), *Introduction to the Philosophy of Science*. Hacket Publishing Co mpany (1992). ● Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature machine intelligence* 1.5 (2019): 206-215.
Simplicity		
13	28.01.2025	● Belkin, Mikhail. "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation." *Acta Numerica* 30 (2021): 203-248.
Causation		
14	04.02.2025	● Richard Scheines (2004). Causation. ● Deaton, Angus, and Nancy Cartwright. "Understanding and misunderstanding randomized controlled trials." *Social science & medicine* 210 (2018): 2-21.