Seminar: Philosophy of Science for Machine Learning

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

Meeting Time

Thursdays, 12-14h ct
Ground Floor Lecture Hall
Maria-von-Linden-Straße 6,
72076 Tübingen

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 remains 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

Classes will meet in person, from 12-14h ct CET, every Thursday between October 19th and February 8th, with the exception of holidays on December 28th and January 4th, for a total of 15 class meetings. There will be required readings for every meeting. Everyone should make an effort to read all the material. 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 discussion.

A group of 2-3 students will be experts for every session. This responsibility includes presenting the core arguments of the required reading in 15-20 min, preparing some questions for discussion, and fielding questions from the rest of the class. Readings will be assigned with regard to some degree for student preference.
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 10-15 minutes, allowing for 15-20 minutes of discussion. I will make myself available beforehand to discuss the material for the presentation.

There will also be a 1,500 word essay due sometime in **March**. The exact deadline will be announced early on. The subject matter is flexible and intended to answer to individual interest, but students must submit a 1-page proposal for approval by **January 19th**.

Grading is determined as follows:

Class participation: 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
It is the responsibility of each student to be aware of the university policies on academic integrity, including the policies on cheating and plagiarism.

Reading List

19.10.23	*First Class*
(Data) Science and Values	
26.10.23	Richard Rudner (1953) The Scientist Qua Scientist Makes Value Judgements.

	Heather Douglas (2000) Inductive risk and values in science.
	Corbett-Davies & Goel (2018) The Measure and Mismeasure of Fairness.
02.11.23	Liam Kofi Bright (2018) Du Bois' democratic defense of the value free ideal.
	Borsboom, Romeijn & Wicherts (2008) Measurement invariance versus selection invariance: Is fair selection possible?
	Alexander Mussgnug (2022) The predictive reframing of machine learning applications: good predictions and bad measurements.
Probability and The Problem of Induction	
09.11.23	Alan Hájek (2019), Interpretations of Probability.
	Cynthia Dwork (2022), Fairness, Randomness and the Crystal Ball.
16.11.23	David Hume (1748), An enquiry concerning human understanding. (selections)
	Leah Henderson (2022), The Problem of Induction.
23.11.23	Ulrike von Luxburg, Bernhard Schölkopf (2008) Statistical Learning Theory: Models, Concepts and Results.
	Tom Sterkenburg, Peter Grünwald (2021), The no-free-lunch theorems of supervised learning.
	Ravit Dotan (2020) Theory Choice, non-epistemic values and machine learning.
Confirmation-Theoretic Responses to Hume	
30.11.23	Carl Hempel (1945) Studies in the Logic of Confirmation.
	Rudolf Carnap (1945) On Inductive Logic.
07.12.23	Patrick Maher (2004) Probability Captures the Logic of Scientific Confirmation.

	Kevin T. Kelly and Clark Glymour (2004) Why Probability does not Capture the Logic of Scientific Justification.
Falsification-Theoretic Responses	
14.12.23	Karl Popper (1934) Logic of Scientific Discovery Part I.
	Deborah Mayo & Aris Spanos (2006) Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction.
Learning-Theoretic Responses	
21.12.23	Oliver Schulte (2022) Formal Learning Theory.
	Daniel Steel (2010) What if the Principle of Induction is Normative? FLT and Hume's Problem.
28.12.23	*No Class, Winter Break*
04.01.24	*No Class, Winter Break*
Realism/Anti-realism	
11.01.24	Clark Glymour (1992) Realism and the Nature of Theories.
	Leo Breiman (2001) Statistical Modeling: The Two Cultures
Measurement	
18.01.24	Kino Zhao (2023) Measuring the Nonexistent: Validity before Measurement.
	Ian Hacking (1995) The looping effects of human kinds.
	Eran Tal (2023) Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare.
Explanation	
25.01.24	Wesley Salmon (1992) Scientific Explanation.
	Cynthia Rudin (2019) Stop explaining black

	box machine learning models for high stakes decisions and use interpretable models instead.
Simplicity	
01.02.24	Tom Sterkenburg (2023) Statistical Learning Theory and Occam's Razor.
	Kevin T. Kelly (2005) Simplicity, Truth and the Unending Game of Science.
	Daniel Hermann (2022) PAC Learning and Occam's Razor.
Causation	
08.02.24	Angus Deaton & Nancy Cartwright (2018) Understanding and misunderstanding randomized controlled trials.
	Scheines (2004) Causation