Epistemology and Philosophy of Science, Module 3: Epistemic Opacity in Applications of Machine Learning in Science

3 - Zerilli: Explaining Machine Learning Decisions

Robert Michels

26 November 2024

LanCog, Centre of Philosophy, University of Lisbon
robert.michels@edu.ulisboa.pt

Overview

Reminder

Simulating simulations

Questions for the discussion

Reminder

Reminder: Computer simulation models

Different kinds of computer simulation models

- *Equation based model*: approximates solutions of differential equations which capture changes of physical quantities over time computationally
- *Agent-based simulation*: represents behaviour of agents which follow certain local rules (e.g. rule to move if more than the desired degree of neighbours are different in the Schelling segregation model)
- *Monte Carlo simulation*: equation-based model which relies on a Monte Carlo algorithm based on randomness.

Reminder: Computer simulation models

Different purposes for computer simulation models:

- *Predictive purposes*: Used to e.g. predict future evolution of a dynamical process
- *Heuristic purposes*: Used to e.g. illustrate or teach a certain
- *Understanding purposes*: Used to gain understanding of a regular model or of the target phenomenon.

Reminder: Strong vs weak AI

- *Strong AI* – artificial intelligence which equals or surpasses general human cognition across a wide range of different cognitive tasks, including e.g. reasoning, recognition, speech production, etc.
- *Weak AI* – artificial intelligence which equals or surpasses general human cognition with respect to a specific, often very narrow cognitive task

Reminder: Artificial neural networks

- Consist of neurons and connections between them arranged in layers (input layer, hidden layer(s), output layer)
- *Deep neural networks* have multiple hidden layers
- Information processing in artifical neural networks: each neuron receives inputs from neurons on previous layers, this information is assigned weights (tracking importance of the information), combines this information and passes it on

Reminder: Epistemic opacity in deep neural networks

- Regular computer simulations not involving artificial neural networks may already involve calculations which are too many or too complex for a human
- Due to their at least partly autonomous learning process, humans are not able to grasp the internal mechanisms (aka the reasoning process and decisions) which a deep neural network may involve – they are *epistemically opaque*, i.e. not transparent to our recognition, a *black box*
- We may know that a DNN performs its task very well (i.e. reliably gives us the right outputs for the inputs we feed into it), but can in many cases not know why or how it does that

Simulating simulations

An example from applications of AI in cosmology (Meskhidze (2023))

- Physicists investigate the large-scale structure (clusterings of planets, stars, galaxies) in the universe by modelling dark matter as a fluid which is initially homogenous
- The creation of sctructures in the universe is captured by deviations from homogeneity
- For small deviations, this can be computed 'by hand', but large deviations which occur later in the evolution of the model can only be handled by computer simulation

An example from applications of AI in cosmology (Meskhidze (2023))

- The deviations from homogeneity of a fluid are modelled using a large number of particles which interact via Newtonian gravitational forces
- Such simulations are very computation intensive
- Two factors:
  - Extremely large number of particles (in one instance 16 million) have to be simulated
  - A high number of simulations is needed to infer from the model (using a Monte Carlo algorithm) which values the cosmological parameters (expansion rate of the universe, cruvature, ...) have in observed states of the universe
- Running the simulation as often as needed is practically impossible

An example from applications of AI in cosmology (Meskhidze (2023))

- To solve this problem, cosmologists have begun to rely on machine learning to train an AI on a small number of simulation runs and then use it to interpolate from this small basis the results of a large number of simulation runs
- 'PkANN' uses neural network
- A model of this kind is not a model of a physical target phenomenon, but rather a model of models, a simulation of simulations

Questions for the discussion

Zerilli (2022)

General questions

- What is Zerilli's main thesis?
- How does he argue for it?
- Are the arguments good? Valid reasoning, true or at least plausible premises?
- How does Zerilli's argument relate to topics which we have previously discussed?

References

Meskhidze, H. (2023). Can machine learning provide understanding? how cosmologists use machine learning to understand observations of the universe. Erkenntnis, 88(5):1895–1909.

Zerilli, J. (2022). Explaining machine learning decisions. Philosophy of Science, 89(1):1–19.