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

2 - Computer models, simulation, AI models in science

Robert Michels

26 November 2024

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

Formalities

Reading for the next two sessions?
We have to decide the reading for next week and the week after:

- *Deep Learning: Philosophical Issues* (Buckner (2019)) – philosophical introduction to the technical background of deep neural networks
- *Explaining Machine Learning Decisions* (Zerilli (2022)) – discussion of XAI (explainable AI), technical methods to make opaque ML models epistemically accessible to us
- *Instruments, Agents, and Artificial Intelligence: Novel Epistemic Categories of Reliability* (Duede (2022)) – what is the epistemic role of AI, is it an instrument to gain knowledge, does it play the role of an expert, or does it play a different epistemic role?
- *Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI* (Durán and Jongsma (2021)) – disscussion of epistemic and ethical issues about trust in AI in medicine

Formalities

Moving the date of the exam – approved!

- Fourth and last session before the exam moved from Tuesday 17 December 14:00 to **Friday 13 December 10:00-13:00 – Sala Matos Romao**
- Exam on **Tuesday 17 December 14-16:00**! (instead of the regular fourth session of the module)

Overview

Computer models

AI – some basic background

Simulating simulations

Questions for the discussion

Computer models

Reminder: "Regular" models in science

- We have discussed models in science and seen and played around with one simple model, Schelling's model of segregation
- Models are used in science to *explain* and *gain understanding* of and *knowledge about* particular complex, dynamic phenomena and to *make predictions* about unobserved occurrences of these phenomena
- They can be used to do this since they are *analogous* to the phenomena they model; they capture certain aspects of the phenomena (positive analogy; e.g. basic dynamics of segregation in Schelling's model), but not others (negative analogy; e.g. legal or other social factors of segregation not represented in Schelling's model)
- Models are *less general than scientific theories* – Newtonian mechanics (theory) vs application of Newtonian mechanics to a particular kind of physical system such as a pendulum (model)

Moving closer towards our subject: Computer simulation models

A first, narrow characterization of computer simulation models

> *In its narrowest sense, a computer simulation is a program that is run on a computer and that uses step-by-step methods to explore the approximate behavior of a mathematical model. Usually this is a model of a real-world system (although the system in question might be an imaginary or hypothetical one). Such a computer program is a computer simulation model. (Winsberg (2022), §1.1)*

What the narrow characterization leaves out

- This is a narrow characterization, since it ignores any factor which goes into designing the model and into interpreting the model
- These factors are however relevant to make the program a model

A broader characterization

A second, broader characterization of computer simulation models (Winsberg)

> *Successful simulation studies do more than compute numbers. They make use of a variety of techniques to draw inferences from these numbers. Simulations make creative use of calculational techniques that can only be motivated extra-mathematically and extra-theoretically. As such, unlike simple computations that can be carried out on a computer, the results of simulations are not automatically reliable. Much effort and expertise goes into deciding which simulation results are reliable and which are not. (Winsberg (2022), §1.2)*

Note how this emphasizes the need of a human expert both in design, interpretation, and evaluation of the model

What the broader characterization leaves out

- The broader definition still leaves out an important characteristic of computer simulations:
- They are very often used to model *dynamic processes*

A dynamic characterization

A characterization emphasizing the dynamic nature of the simulations (Winsberg (2022), §1.3)

> *Simulations are closely related to dynamic models. More concretely, a simulation results when the equations of the underlying dynamic model are solved. This model is designed to imitate the time-evolution of a real system. To put it another way, a simulation imitates one process by another process. In this definition, the term "process" refers solely to some object or system whose state changes in time. 18 If the simulation is run on a computer, it is called a computer simulation. (Hartmann (1996), 83.)*

Note that, as Humphreys points out, Hartmann's definition leaves out modelling of static aspects (Winsberg (2022), §1.3)

Summary: characteristics of computer simulation models

Characterizing computer simulation models: summary of what matters for us in the context of this seminar

- Computer simulation models are based on mathematical models consisting of differential equations which are solved by an algorithm running on a computer
- They involve substantial human interpretation in design (which equations are used to describe a particular phenomenon), interpretation (what does the model tell us?), evaluation (is the model reliable? Does it verify or falsify a hypothesis?)
- Many computer simulation models capture dynamic processes (e.g. orbit of a planet, growth of a population, etc.)

Three kinds of computer simulation models

Equation based simulations (Winsberg (2022), §2)

- Computer simulation which relies on laws describing a particular phenomenon from a particular scientific theory or formulated ad hoc (e.g. to test a candidate law)
- Usual form of the laws: differential equations which describe change of certain physical quantities over time
- In equation based simulations these equations are applied to a particular dataset consisting of e.g. observations about a physical system, such as our solar system in cosmology
- Often used to test (or, as a heuristic mode, develop) a theory, or to make predictions about the target phenomenon

Three kinds of computer simulation models

Agent-based simulations (Winsberg (2022), §2)

- Usually not based on differential equations/mathematically formulated laws which determine the evolution of the complete target system
- Rather, based on local rules followed by individual agents
- E.g. in Schelling's segregation model the rule that an individual moves to a free spot if its threshold for similarity of neighbours is not met – this is a local rule, which every 'agent' follows independently of what the other agents do

Moving closer towards our subject: Computer simulation models

Monte Carlo simulations (Winsberg (2022), §2)

- Simulations which involve algorithms which rely on randomness to deliver approximate results and hence involve a small probability that their outcome is wrong
- Monte Carlo computer simulations are an example of a kind of computer simulation which may surpass human capabilities (amount of computation may not be feasible for humans in a comparably short time), without relying on AI

Moving closer towards our subject: Computer simulation models

Three purposes of computer simulation models (Winsberg (2022), §3)

1. **Heuristic purposes**

- Computer simulation models can be used to convey certain information to the researcher or to an audience for a certain purpose
- The model need not be 'perfect', in the sense of very accurately representing the target phenomenon which it models, to be used as a heuristic
- Example: Simulations of e.g. a kinetic model of gas, using a Monte Carlo algorithm (an algorithm which relies on randomness and can be processed faster than a 'regular' algorithm, but has a certain error-probability)

Moving closer towards our subject: Computer simulation models

Three purposes of computer simulation models (Winsberg (2022), §3)

2. **Predictive purposes**:

- Computer simulations may be used to predict, or retrodict, the evolution of a particular phenomenon
- Example: computer simulation of climate change

Moving closer towards our subject: Computer simulation models

Three purposes of computer simulation models (Winsberg (2022), §3)

3. **Understanding purposes**:

- Scientists may have collected a large amount of data which would be impossible or difficult to process by a human, but a computer simulation model may help extracting certain information from the model and gain understanding of the target phenomenon from it
- One may add: they may not only gain understanding, but also arrive at explanations of aspects of the phenomenon and gain knowledge about it
- Example: discovering patterns in orbital motion of planets based on observation data gathered by telescopes

Moving closer towards our subject: Computer simulation models

Why not a "regular" model?
Computer simulation needed for a number of reasons:

- Solving the equations on which the model is based may take a human too long
- The amount of data collected may be too large to process for a human
- The equations on which the model is based may be not analytically solvable (the exact result can in principle be worked by logical and mathematical reasoning/processing) in principle or in a reasonable time, but can be numerically approximated (an approximate result can be generated by trying a range of different values for the variables in the equation)

AI – some basic background

Strong vs weak AI

- *Strong AI* (sometimes also called *Artificial General Intelligence (AGI)*) – 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. – so far not achieved
- *Weak AI* – artificial intelligence which equals or surpasses general human cognition with respect to a specific, often very narrow cognitive task – widely used in science, economy, by society at large

$\Rightarrow$ Focus on weak AI!

What distinguishes AI from ‘regular’ computing

Two distinguishing factors

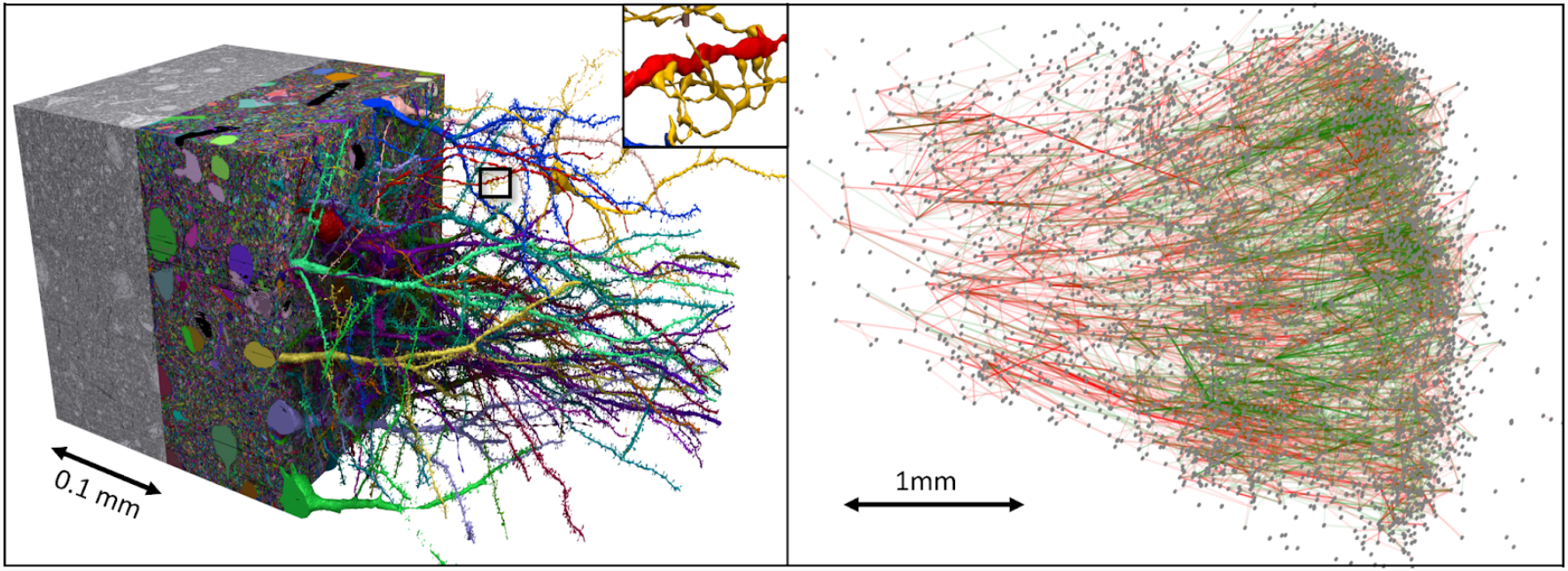
1. Reliance on artificial neural networks
2. Reliance on machine learning

Artificial neural networks

Background

- Artificial neural networks (ANN) were originally designed to mimic the structure and (to a degree) the working of the human brain
- Basic structural elements: neurons and connections between them arranged in layers

Neural network in the brain



Reconstruction of 5000 neurons and connections from a volume of human neocortex; source `https://research.google/blog/a-browsable-petascale-reconstruction-of-the-human-cortex/?hl=hr`

Artificial neural networks

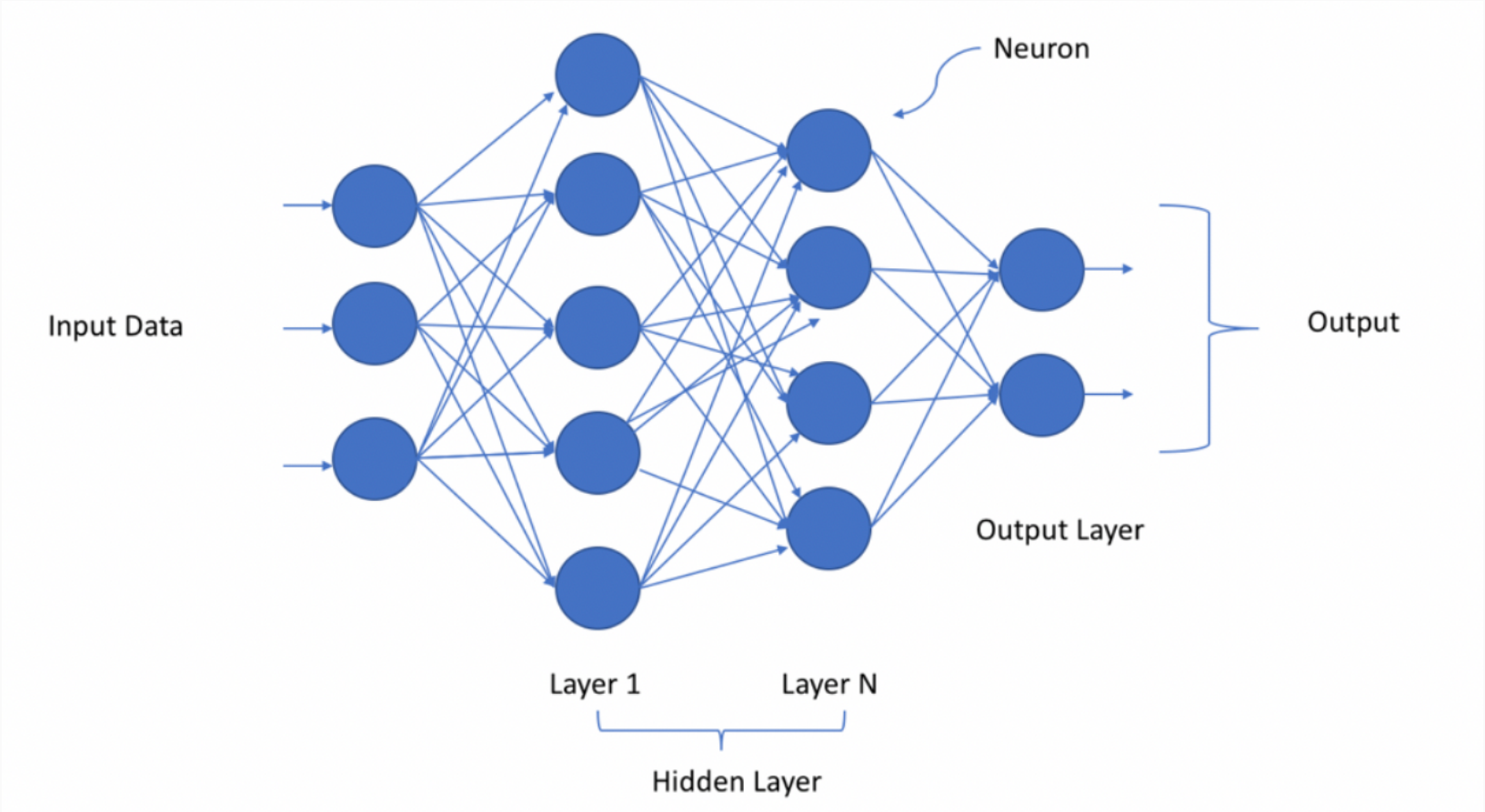
Deep neural networks (DNNs)

- *Deep neural networks* are the basic components of most current machine learning applications (including transformer, the currently most advanced AI models such as ChatGPT, Dall-E, etc.)
- An ANN is deep if it has multiple hidden layers

More on the basic structure of ANNs

- In any ANN, neurons are arranged in a sequence of layers, where each neuron on a layer takes inputs from (usually multiple) neurons the previous layer and passes on outputs to neurons on the next layer
- Three different kinds of layers in DNNs:
  - *Input layer*: neurons on this layer receive 'raw' data, date which has not been processed by neurons
  - *Output layer*: neurons on this layer output the data of the network to us or for further processing
  - *Hidden layers*: a layer which is between the input and output layers

Arn artificial neural network



Schema of an artificial neural network with two hidden layers; source `https://towardsdatascience.com/a-laymans-guide-to-deep-neural-networks-ddcea24847fb`

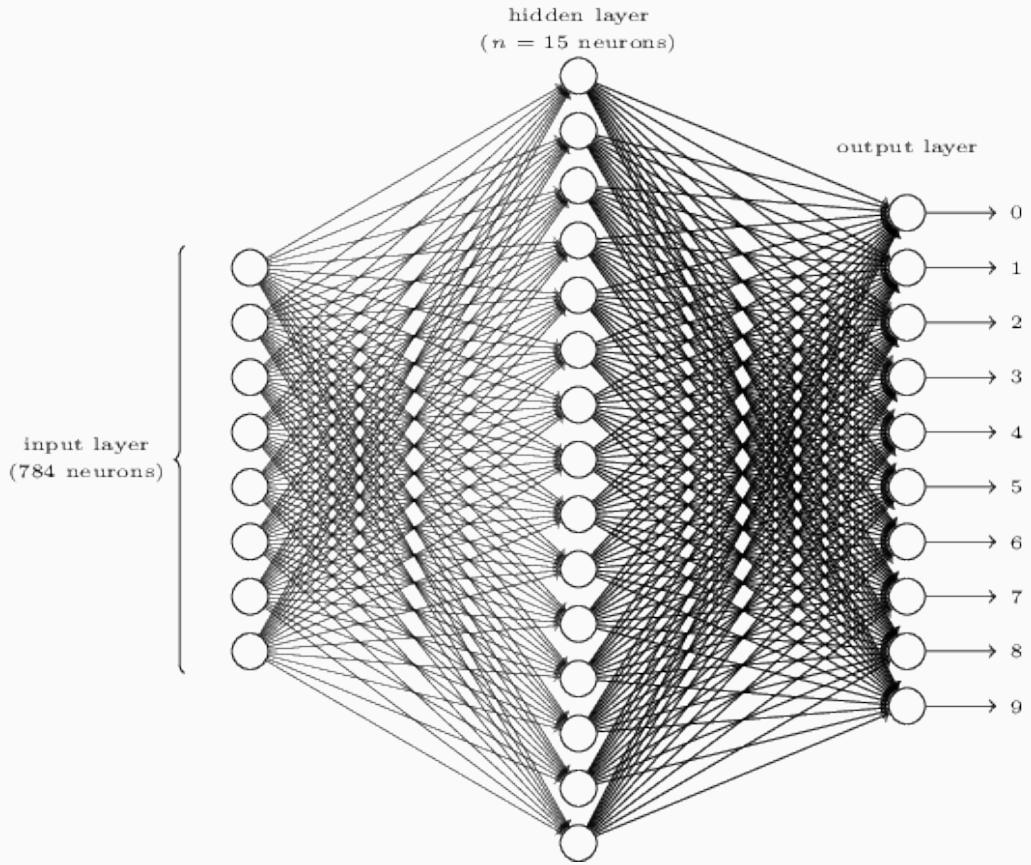
Artificial neural networks

More on the basic structure of ANNs

- Each connection between two neurons is assigned a weight, a number which represents how important the input of a neuron is for its output
- The neuron mathematically combines all the inputs and the corresponding weights and then passes on an output, which is then assigned a weight and feeds into another neuron on the next layer as one of its inputs, and so on
- How many neurons there are in each layer and how many layers there are depends on the task the ANN/DNN is supposed to perform

Artificial neural network

Example: A simple DNN designed to learn to recognise a hand-written number



DNN with one hidden layer designed to recognize the hand-written number 9; source `http://neuralnetworksanddeeplearning.com/chap1.html`

Artificial neural networks

Example: A simple DNN designed to learn to recognise hand-written numbers
From: `http://neuralnetworksanddeeplearning.com/chap1.html`

- Designed to process a 28 x 28 pixel image of a scanned hand written numbers and correctly identify them
- It has 28 x 28 = 784 input neurons, one for each pixel of the imput image
- The image is grey-scale and the shade of each pixel is encoded by a number between 0.0 (white) and 1.0 (black)
- The output layer has 10 neurons, each standing for a number from 0-9 – desired result: if the DNN gets the pixel pattern of an image of e.g. a 9 as input, the 9th output neuron is the one which passes on the highest output number, indicating that the DNN has recognized it

Artificial neural networks

How this DNN processes data

- Each input neuron passes on the grey-scale value for its pixel to all neurons on the hidden layer, each together with a different weight
- Each hidden-layer neuron hence gets one complete numerical representation of the whole image with colour and importance information for each pixel
- Each output layer neuron gets this information condensed into a number from each hidden layer neuron as input, and only the neuron corresponding to the recognized number passes on the highest output value

How this DNN processes data

- Based on the different patterns of weight assigned to the pixels in the image, each hidden layer neuron may recognize a different part of the shape of a number
- Each output neurons then checks whether the combination of these partial number shapes forms its number and outputs a low value if not and a high value if it does
- Note that the details of what is recognized depends on the example, how the DNN is designed and *what it learns*

Machine Learning with DNNs

How does a DNN learn?

- DNNs excell at learning patterns from (usually) large sets of examples – standard example: MNST data set of 60.000 hand-written digits in the mentioned format
- To allow the DNN to learn, one uses an algorithm designed to get the DNN into an optimal state in which it reliably does what it is supposed to do (e.g. recognizes hand-written 0-9 numbers correctly)
- One commonly used algorithm is gradient descent, which works by minimizing a cost function, which returns higher values for states of the DNN which give wrong outputs
- Learning happens by applying the algorithm time after time, often starting with a random distribution of weights, until the DNN reliably gives the correct output for each input it gets – e.g. it has reached a state in which its cost function is minimized

Machine Learning with DNNs

Two machine learning paradigms

- Supervised vs unsupervised learning

Machine Learning with DNNs

Supervised learning

- The idea of supervised learning is to train the DNN with a specially prepared set of training data consisting of desired input-output pairs – in the number-recognition case, the 128 x 128 pixel grey-scale images of hand-written numbers are combined with explicit labels, i.e. the input contains the image and the correct number in a format readable by the DNN's learning algorithm
- The goal is that after the learning phase, the DNN will be able to successfully perform its task for new data (e.g. arbitrary images of hand-written numbers not in the training set)
- Different degrees of supervision – from full supervision with respect to the whole training set to only partial supervision for a small subset + autonomous learning

Epistemic opacity in supervised learning

- Humans determine the design of a DNN suitable for a certain task and supervise the learning process
- However, this does not mean that the way the DNN processes the data is transparent to them – the supervision mostly consists in making sure that the outputs match the inputs, not in how they are matched
- DNNs may involve a large number of nodes and hidden layers (often 100+), making it impossible for a human to grasp what each neuron does

Machine Learning with DNNs

Unsupervised learning

- The idea of unsupervised learning is to let a DNN learn patterns from a large unlabeled data set (e.g. text recognition and production trained on English Wikipedia)
- The DNN does this by 'mimicing' the data it processes and then self-correcting based on the errors it makes compared to the original data
- Still involves decisions of human in design of the network

Epistemic opacity in unsupervised learning

- DNNs which learn without supervision are epistemically opaque for the same reasons as those which learn with supervision
- But the absence of human supervision however adds to the opacity, since the guidance during training gives humans at least superficial indications about how the DNN works

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

Some questions for the discussion

Questions about Wood (2022)

- Why do you think could Guimerà and Sales-Pardo not just reveal to journals that their new discovery about cell division was produced by an AI-algorithm?
- What did they do to address this problem and how did what they did help?
- What is symbolic regression?
- How does it help addressing the problem of epistemic opacity?
- Can 'machine scientists' generate new knowledge? New understanding?

Some questions for the discussion

Questions about Humphreys (2009)

- What is the main point Humphreys argues for in his paper?
- What does Humphrey's mean when he says that computational science 'uses methods that push humans away from the centre of the epistemologcial enterprise'?
- What are the hybrid and the automated scenario?
- How does Humphrey's define 'epistemically opaque' and 'essentially epistemically opaque'?
- Has the distinction between internalist and externalist justification a relevance to the notion of epistemic opacity?
- How does Humphrey's discussion of epistemic opacity relate to the idea that epistemology is concerned with cognitive success?

References

Buckner, C. (2019). Deep learning: A philosophical introduction. Philosophy Compass, 14(10):1–19.

Duede, E. (2022). Instruments, agents, and artificial intelligence: Novel epistemic categories of reliability. Synthese, 200(6):1–20.

Durán, J. M. and Jongsma, K. R. (2021). Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical ai. Journal of Medical Ethics, 47(5):2020–106820.

Hartmann, S. (1996). The world as a process: Simulations in the natural and social sciences. In et al, R. H., editor, Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View. Springer.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. Synthese, 169:615–626.

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

Bibliography ii

Winsberg, E. (2022). Computer Simulations in Science. In Zalta, E. N. and Nodelman, U., editors, The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Winter 2022 edition.

Wood, C. (2022). Powerful 'machine scientists' distill the laws of physics from raw data. Quanta magazine.

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