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

1 - Background: Epistemology, Models in Science

Robert Michels

26 November 2024

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

Contact

- If you want to discuss something about the module with me, please write of contact me to make an appointment by e-mail
- E-mail address: robert.michels@edu.ulisboa.pt

Formalities

Overview of the sessions

- The module consists of four sessions, plus an exam (all on Tuesday, 14:00-17:00 in Sala Matos Romao):
  1. 26 November: Epistemological background, models in science; Reading: Frigg and Hartmann (2024), intro + §3
  2. 3 December: Computer models, simulation, AI models in science; do they pose a special epistemic problem? – Reading: Humphreys (2009), Wood (2022)
  3. 10 December: to be decided
  4. 17 December: to be decided
  5. 7 January: Written exam

Formalities

Possible readings/topics for sessions 3 and 4:

- *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

- Should we try to move the date of the exam?

Overview

Epistemological Preliminaries

Epistemology of science

Models in Science

Epistemology of Models

Epistemological Preliminaries

Epistemological preliminaries

- We can think of epistemology as the systematic study of cognitive success and failure (Steup and Neta (2024))
- Two core questions given this approach:
  - Who or what can be cognitively successful (*objects* of cognitive success)?
  - What kinds of cognitive success are there (*types* of cognitive success)?

Epistemological preliminaries

Who or what can be cognitively successful (*objects* of cognitive success)?

- Mental states (e.g. having a true belief)
- Acts (e.g. learning by reading a book)
- Persons (e.g. someone know how to cook spaghetti)
- Group of persons (e.g. participants of a seminar manage to understand the epistemology of science)
- Theories (e.g. the theory of evolution explains why certain animal species went extinct)
- Methods (e.g. statistics-based opinion surveys correctly predicts consumer behaviour)
- Instruments (e.g. a telescope may allow us to gather information about distant galaxies)
- Models (e.g. the Lotka Volterra model of predator-prey population dynamics explains population fluctuations in the hare-lynx population in Canada)

Epistemological preliminaries

What kinds of cognitive success can be had (*types* of cognitive success)? – some candidate notions:

- Belief
- Opinion
- Knowledge
- Explanation
- Understanding
- ...

Let's look at each concept in a bit more detail to see whether it describes a state of cognitive success!

Epistemological preliminaries

Belief as a cognitive success-term?

- Simply *having a belief* is not a cognitive success – one reason: our direct control over which beliefs we have is limited (cf. Boespflug and Jackson (2024))
- However, having *a belief of the right kind*, can be a cognitive success – more on the next slide
- Also, if there are degrees of belief (i.e. if one can believe more or less strongly or with more or less confidence; see Schwitzgebel (2024), §2.3), then one may be cognitively success by believing something to the right degree

Epistemological preliminaries

Belief as a cognitive success-term – success- and failure-conditions

- Having what kind of belief can be a cognitive success? – one answer: those which meet certain *epistemic norms*
- E.g. the norm that the beliefs we hold are *true*
- Norm for degrees of belief: Lewis's Principal Principle (Lewis (1986)): The degree to of one's belief in a certain outcome of a chancy process (e.g. a coin toss) should (absent other evidence) equal the objective chance of that outcome

Opinion as a cognitive success term?

- 'Opinion' – ambiguous term – three definitions proposed by Merriam Webster:
  - a view, judgment, or appraisal formed in the mind about a particular matter – not a cognitive success term
  - belief stronger than impression and less strong than positive knowledge – cognitive success term?
  - a formal expression of judgment or advice by an expert – to be able to give an opinion in this third sense is a cognitive success

Opinion as a cognitive success term – success conditions

- 'Opinion' – ambiguous term – three definitions proposed by Merriam Webster:
  - a view, judgment, or appraisal formed in the mind about a particular matter – not a cognitive success term
  - belief stronger than impression and less strong than positive knowledge – cognitive success term?
  - a formal expression of judgment or advice by an expert – to be able to give an opinion in this third sense is a cognitive success

Epistemological preliminaries

Knowledge as a cognitive success-term?

- Having knowledge is clearly a cognitive success
- Success and failure conditions given by philosophical account of knowledge

Epistemological preliminaries

Knowledge as a cognitive success-term?

- Two traditions in epistemology: knowledge as a defined term vs knowledge as a primitive (i.e. undefined) term
- Primitive term (see e.g. Williamson (2000)): success conditions given by relevant external factors, e.g. being in the right mental state and having the right kind of evidence
- Defined term: success conditions: given by correct definition of knowledge – see next slide

Epistemological preliminaries

Knowledge as a defined term?

- There have been many different attempts at defining knowledge by philosophers (cf. module 1)
- Many of these attempts depart from the basic idea of the 'justified true belief'-picture of knowledge criticised by Gettier (1963)
- They provide different suggestions for a third required ingredient for knowledge besides having a belief and that belief being true, called 'justification' by Gettier and others

Knowledge as a defined term? – internalism vs externalism

- An important distinction in this context: *internalism* vs *externalism* about justification
- Roughly, internalists think that justification concerns only mental states and externalists deny this, allowing also mind-external factors to provide justification
- This distinction will be useful for our discussion of epistemic problems with AI models in science, so it is worth going into a bit more detail

Epistemological preliminaries

Knowledge as a defined term? – internalism vs externalism

- A classic internalist view: *Evidentialism* (Feldman and Conee (1985))
- A simple Evidentialist definition of knowledge (not Feldman and Conee's actual definition!):
  - *S* knows that *p* if, and only if, i) S believes that *p*, ii) *p* is true, and iii) that *S* believes that *p* fits *S*'s evidence.
- What makes this an internalist view is the (independent) assumption that one's evidence is always fixed by one's mental state

Epistemological preliminaries

Knowledge as a defined term? – internalism vs externalism

- A classic externalist view: *Reliabilism* (see e.g. Goldman (1979))
- A Reliabilist definition of knowledge (not Goldman's own) based on his main idea:
  - *S* knows that *p*, if, and only if, i) S believes that *p*, ii) *p* is true, and iii) '*S*'s believing *p* results from a reliable cognitive belief-forming process (or set of processes).' (Goldman (1979), 13)
- Example of a reliable belief forming process: *seeing* – I know that there is a painting on the wall, since my belief that it is is true, I acquired that belief by seeing the painting on the wall and seeing an object in close proximity is a reliable process to form a belief

Epistemological preliminaries

Explanation as a cognitive success-term?

- Having an explanation is clearly a cognitive success
- Again, success conditions depend on what one takes explanation to be

Explanation as a success-term – success/failure?

- Different definitions of what explanation is
  - If having an explanation amounts to having a deductive argument for a sentence stating the fact (classical Deductive-nomological/Hempel Oppenheim view of explanation, cf. module 2), then this is the success condition
  - If having an explanation of a fact amounts to being able to identify the kind of causal mechanism which brought it about (see e.g. Machamer et al. (2000)), then this is the success condition

Epistemological preliminaries

Understanding as a success-term – success/failure?

- Success conditions: depends on how one understands 'understanding'
- One prominent proposal: understanding a subject mean grasping systematic relations within it and to other subjects (cf. Zagzebski (2001))
- Understanding is arguably a graded notion, i.e. one can understand something *more or less well* – success conditions here are conditions for having a sufficiently high degree of understanding

Understanding as a success-term – success/failure?

- Also, one can distinguish between at least three different kinds of understanding:
  - *understanding that*: Understanding that something is the case as – I understand that the person depicted by the painting in this room is Matos Romão.'
  - *understanding why*: Grasping reasons or an explanation for – I understand that the painting is in this room, because Matos Romão gifted the university his library.
  - *objectual understanding*: Understanding a subject, topic, phenomenon, person, etc. – I understand propositional modal logic
- How these notions are connected is subject to debate (see Baumberger et al. (2017), §5.2)

Epistemological preliminaries

Understanding and knowledge (Baumberger et al. (2017), §5.1)

- How are understanding and knowledge related?
- Some authors argue that understanding is a special kind of knowledge – especially plausible for *understanding that*, which may just be knowledge of the understood fact; *understanding why* may be taken to be knowledge of an explanation
- However, some argue that at least *objectual understanding* is not reducible to knowledge
- E.g. Kvanvig (2003) argues that one may objectually understand a topic without having knowledge, because we can understand, but not know something which is not true (e.g. the phlogiston theory of heat); in order to know the corresponding propositions (e.g. the propositions describing phlogiston), they would have to be true – because knowledge in general requires truth

Epistemology of science

Science as an epistemic enterprise

What we mean by 'science'

- By 'science' philosophers of science usually mean the natural sciences, including in particular physics, biology, chemistry, psychology, ...
- Social sciences and economics also discussed, but to a lesser degree
- Focus is usually not on the humanities

Science as an epistemic enterprise

- Science is clearly an epistemic enterprise
- One of the aims of science is to achieve cognitive success of different forms
- However...

Science as an epistemic enterprise

The non-epistemic dimensions of science

- Science is, but is not merely an epistemic enterprise!
- Pragmatic dimension: science helps us construct useful things, provides us with knowledge which improves our lives, ...
- Political dimension: science may e.g. informs politic decisions
- Social dimension: science may e.g. contribute to addressing social problems

Science as an epistemic enterprise

- Who or what are the objects of cognitive success **in Science**?
- What kinds of cognitive success is attained **in Science**?

Science as an epistemic enterprise

Who or what are the objects of cognitive success in Science?

- Plausibly: persons, groups of persons, theories, methods, instruments, models
- Plausibly not: mental states?, acts?

Science as an epistemic enterprise

What kinds of cognitive success is attained in Science?

- Belief, high degree of certainty/belief, knowledge, explanation, understanding – yes!
- Opinion? – in the sense of 'expert opinion' – yes, but this amounts to beliefs, high degree of belief, knowledge of experts

Science as an epistemic enterprise

Who or what are the objects of cognitive success in Science?

- Subject to discussion!
- A seemingly plausible proposal:
  - *Theories*, *models*, *methods*, . . . can all be cognitively successful, but this success is merely instrumental; we all measure their success by their contribution to our cognitive success, i.e. the cognitive success of humanity as a whole.

Models in Science

Models in Science

What is a model?

A first approximation:

> *I consider the following as the core idea of what constitutes a scientific model: A model is an interpretative description of a phenomenon that facilitates access to that phenomenon. ("Phenomenon" refers to "things happening"[...].) This access can be perceptual as well as intellectual. If access is not perceptual, it is often facilitated by visualization, although this need not be the case. Interpretative descriptions may rely, for instance, on idealizations or simplifications or on analogies to interpretative descriptions of other phenomena. Facilitating access usually involves focusing on specific aspects of a phenomenon, sometimes deliberately disregarding others. As a result, models tend to be partial descriptions only. Models can range from being objects, such as a toy airplane, to being theoretical, abstract entities, such as the Standard Model of the structure of matter and its fundamental particles. (Bailer-Jones (2009), 1–2.)*

Models in Science

An example: Schelling's model of segregation

`http://nifty.stanford.edu/2014/mccown-schelling-model-segregation/`

Models in Science

What is a model?

- Important distinction: theory vs model
- Scientific theories are more general than models; a theory describes a phenomenon in very general terms, a model in contrast captures a more specific instance of the phenomenon
- Example: Newtonian mechanics is a general physical theory which describes the motion of objects and the forces which act on them. In order to apply this theory to a particular physical system, for example to describe the motion of a simple pendulum, the theory has to be combined with specific assumptions about the construction of the pendulum, giving us a model of the theory for this particular physical system

Models in Science

What is a model?

- Are models always instances of more general theories?
- Perhaps not; at least there are models for phenomena for which there is not general theory – Schelling's segregation model may be an example, since there is no general theory of segregation, or the Lotka-Volterra model of dynamics of predator-prey populations
- There are different views concerning the relation between theories and models (Frigg and Hartmann (2024), §4.2)

Models in Science

What sort of thing is a model? (Frigg and Hartmann (2024), §2)

- Some models are themselves *physical objects* – e.g. small model of a bridge, or an atom, or a cell
- There is controversy about the nature of more abstract models which are not (such as e.g. the Newtonian model of the simple pendulum)
- Some philosophers argue that such models are *fictions*, i.e. the same sort of imaginary thing as Sherlock Holmes or Harry Potter
- Others take them to be *abstract objects*
- or *set-theoretic objects* which represent a certain structure (cf. models in logic)
- or *descriptions* or *equations*

Models as analogous (Hesse (1967))

- Importantly, no matter what we take models to be, a crucial aspect of what makes them models is that they are *not identical* to the phenomena they are supposed to represent
- Instead, they are *analogous* to them
- If something is in analogy to another thing, this always implies that the former shares some of the properties of the thing it is analogous to (*positive analogy*), but also lacks some (*negative analogy*)
- Example: Schelling's segregation model is positively analogous to real racial segregation in the US in that it captures that the preference for having neighbours of the same race is relevant to satisfaction with place of living, but it is negatively analogous to the real phenomenon in that it ignores other relevant factors, such as economic factors

Models in Science

Some kinds of models (Frigg and Hartmann (2024), §1)

- *Scale models*: smaller or larger versions/replicas of the modelled object/phenomenon
- *Analogical models*: are in some sense analogous to the modelled object/phenomenon, i.e. share certain significant properties with it, but not others
- *Idealized models*: involve simplifications or distortions in order to allow us to – extreme case: *toy models* – involving extreme simplifications or distortions
- *Exploratory models*: provide the starting point for development of a theory or more adequate model

Epistemology of Models

Epistemology of Models

Models and cognitive achievements

- Which cognitive achievements can be gained based on scientific models?

Epistemology of Models

Which cognitive achievements can be gained based on scientific models?

- Belief, high degree of belief, knowledge, explanation, understanding – yes!
- Same as science in general?!? – also mediate through cognitive success of humanity as a whole?

Epistemology of Models

The epistemology of models (Frigg and Hartmann (2024), §3)
What epistemic roles do models play?

- They are means to reach cognitive success states like knowledge, but to use them to reach these states, we have to learn about them – *models themselves are targets of our cognitive states*
- But not just about them, we also want to *learn about the target system/phenomenon via the model*!
- Models are used to *explain the target system/phenomenon*
- Models are used to gain or improve our *understanding of the target system/phenomenon*

For us, the last three will be more central! – General question

Epistemology of Models

The epistemology of models (Frigg and Hartmann (2024), §3)
What epistemic roles do models play?

- For us, the last three will be more central!
- General question: Are there differences between models which in some sense rely on machine learning/AI and regular models? – Read Humphreys (2009) for an influential answer to this question!

References

Bailer-Jones, D. M. (2009). Scientific Models in Philosophy of Science. University of Pittsburgh Press.

Baumberger, C., Beisbart, C., and Brun, G. (2017). What is understanding? an overview of recent debates in epistemology and philosophy of science. In Baumberger, S. G. C. and Ammon, S., editors, Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science, pages 1–34. Routledge.

Boespflug, M. and Jackson, E. (2024). Doxastic Voluntarism. In Zalta, E. N. and Nodelman, U., editors, The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Winter 2024 edition.

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.

Bibliography ii

Feldman, R. and Conee, E. (1985). Evidentialism. Philosophical Studies, 48(1):15–34.

Frigg, R. and Hartmann, S. (2024). Models in Science. In Zalta, E. N. and Nodelman, U., editors, The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Fall 2024 edition.

Gettier, E. L. (1963). Is justified true belief knowledge? Analysis, 23:121–3.

Goldman, A. I. (1979). What is justified belief? In Pappas, G., editor, Justification and Knowledge: New Studies in Epistemology, pages 1–25. D. Reidel.

Hesse, M. (1967). Models and analogies in science. In Edwards, P., editor, Encyclopedia of Philosophy, volume 5, pages 354–359. Macmillan.

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

Kvanvig, J. L. (2003). The Value of Knowledge and the Pursuit of Understanding. Cambridge University Press.

Lewis, D. (1986). A subjectivist guide to objective chance. In Philosophical Papers, volume 2, pages 83–132. Oxford University Press.

Machamer, P., Darden, L., and Craver, C. F. (2000). Thinking about mechanisms. Philosophy of Science, 67(1):1–25.

Schwitzgebel, E. (2024). Belief. In Zalta, E. N. and Nodelman, U., editors, The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Spring 2024 edition.

Bibliography iii

Steup, M. and Neta, R. (2024). Epistemology. In Zalta, E. N. and Nodelman, U., editors, The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Winter 2024 edition.

Williamson, T. (2000). Knowledge and its Limits. Oxford University Press.

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

Zagzebski, L. (2001). Recovering understanding. In Steup, M., editor, Knowledge, Truth, and Duty: Essays on Epistemic Justification, Responsibility, and Virtue. Oxford University Press.

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