XAI and Philosophical Work on Explanation (and Models): A Roadmap

BEWARE Workshop, AIxIA 2022
Udine, Italy

Aleks Knoks & Thomas Raleigh
(aleks.knoks@uni.lu & thomas.raleigh@uni.lu)

December 2, 2022

Deep Neural Networks (DNNs) and other black box AI systems can do wonders, but they are also notoriously opaque

The field of XAI offers many methods aiming to explain how opaque AI systems work

Natural question(s): Which methods are better? .. How are we to evaluate candidate explanations? What makes for a good explanation? When is a candidate explanation an oversimplification? A.o.

Philosophers have long been interested in questions like these, and there's an emerging literature applying their answers to the case of XAI

Immediate goal: provide a roadmap through this literature

Broader interest: project on the rationality of trusting opaque AI systems

- When I say 'the field of XAI', I mean it in the narrow sense
- While LIME, SHAP, layerwise relevance propagation, and other XAI methods are different—see Guidotti et al. (2018)—the arguments in this literature are taken to apply to all of them
- The literature we focus on goes beyond (Miller 2019)
  - Miller: Let's supplement XAI with insights from social sciences!
  - Literature we focus on: Wait, first we need to see if XAI is at all viable and useful

  (Issues with (Miller 2019): philosophy as a social science; consensus in philosophy; cherry-picking; more to follow)
- Our roadmap / survey doesn't cover all of the literature

- Optimist end $\Rightarrow$ Pessimist end
- The usefulness XAI models in general $\Rightarrow$ the usefulness of XAI models in a particular domain (healthcare)
- Ambiguity in what's to be explained: (i) DNNs or (ii) the "target phenomenon"

- (Páez, 2019), (Fleisher, forthcoming):
- Feature 1: Shift the focus from explanations (which have to be factive) to the notion of understanding
- Feature 2: Argue that there's no reason to think that XAI methods can't provide a non-factive form of understanding (like models that use Newtonian mechanics do)
- Drawbacks: Optimists stay at an abstract level and argue for a rather weak claim

- (Durán, 2021), (Sullivan, 2022)
- Feature 1: The focus is on the use of DNNs and XAI methods in providing understanding of the target phenomenon
- Feature 2: Top-down approach (starting from bona fide explanations in sciences)
- Feature 3: Although there's no in-principle barrier to the use of XAI, the existing methods have the wrong focus
  - Sullivan: the focus should be on reducing "link uncertainty"
  - Durán: it should be on the formal connection between the DNNs and XAI models
- Durán: we shouldn't mistake the pragmatics of giving explanations for the analysis of the structure of explanation (compare to someone thinking of how to best explain the retrograde motion of planets using the Ptolemaic model to different audiences)

- (Babic et al., 2021), (Durán & Jongsma, 2021), (Krishnan, 2020)
- Feature 1: the use of XAI methods in healthcare is very limited; in particular, they can provide only "ersatz understanding" / "induce unjustified beliefs"
- Feature 2: There are better methods to ensure the safety and effectiveness of opaque AI systems (e.g., clinical trials, track records, interaction)

  Compare using DNNs in medicine to using anesthesia and drugs
- Feature 3: which method is best should be decided on a case-by-case basis

  Also, see (Meshkidze, 2021)

- Keep in mind the ambiguity in what the explanandum (DNNs vs. target phenomenon)
- Don't trust Miller (2019) when it comes to the state of the art in the philosophy of explanation
- Optimists seem to be right in holding that there's no in-principle reason for discounting XAI methods. However, we need to be clearer about the respect in which XAI models simplify
- Pessimists seem to be right in urging concreteness, case-by-case decisions, and insisting that sometimes track records are enough. Research question: Under what conditions are they enough?
- Stay tuned for an actual paper

Grazie!

B. Babic, S. Gerke, T. Evgeniou, I. Cohen (2021), "Beware explanations from AI in health care", In: *Science*, 373.

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi (2018), "A Survey of Methods for Explaining Black Box Models", In: *ACM Computing Surveys*, 51(5).

J. M. Durán (2021), "Dissection scientific explanation in AI (sXAI)", In: *Artificial Intelligence* 297.

J. M. Durán, K. R. Jongsma (2021), "Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI", In: *Journal of Medical Ethics*, 47.

W. Fleisher (forthcoming), "Understanding, idealization, and explainable AI", In: *Episteme*.

M. Krishnan (2020), "Against interpretability: A critical examination of the interpretability problem in machine learning", In: *Philosophy and Technology*, 33.

H. Meskhidze (2021), "Can machine learning provide understanding? How cosmologists use machine learning to understand observations of the universe", In: *Erkenntnis*.

T. Miller (2019), "Explanation in artificial intelligence: Insights from the social sciences", In: *Artificial Intelligence*, 267.

A. Páez (2019), "The pragmatic turn in explanainable artificial intelligence (XAI)", In: *Minds and Machines*, 29.

E. Sullivan (2022), "Understanding from machine learning models", In: *The British Journal for the Philosophy of Science*, 73.