University of Bern
Discussion in the CAS AML M5 Philosophy and Ethics of Extended Cognition and Artificial Intelligence”
Session 3: Nov 1st, 2024

Question of the session: What is the black-box problem of AI and how can it be overcome?

Paper: Zednik, Carlos (2019), Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence, Philosophy and Technology 34 (2):265-288

Summary of the paper:

1. Introduction
    a. Machine learning is taken to be opaque
    b. There are technical approaches to explainable AI, but
    c. conceptual questions remain:
    d. The paper tries to develop a framework for addressing explainability techniques. Normative framework: possibility to evaluate approaches

2. What is the opacity of ML?
    a. General problem: in ML, rules not preprogrammed
    b. Humphreys: a process is epistemically opaque if not all epistemically relevant elements are known to a person. The definition is agent-relative; focus on relevant elements, but not spelled out what the elements are

3. Setting up the framework
    a. Stakeholders:
        i. Operators work with the system, feed it with input
        ii. Executors: make decisions based on the system
        iii. Data subjects: have given data for training and as input
        iv. Decision subjects: are treated on the basis of the output
        v. Creators: have created the ML
        vi. Examiners
    b. Different stakeholders need different kinds of knowledge; explanations give this knowledge
    c. David Marr: approach in cognitive science: levels of analysis:
        i. Computational: what and why
        ii. Algorithmic: how
        iii. Implementational: where

4. The computational level:
    a. What-questions: what the algorithm does; why-questions go for appropriateness: features represented by input are connected to features represented by output
    b. Operators mainly care about what-questions; executors and examiners are also interested in why-questions
    c. Input heatmapping as a method: those pixels that were decisive in an input to obtain a particular output are highlighted. This method helps to answer what- and why-questions

5. The algorithmic and implementational levels

   a. Cognitive science:
      i. Explanations at the algorithmic level answer how-questions, often in terms of mathematical models; often using semantics (parts of the algorithm represent something in the data/real world); important for intervention
      ii. Explanations at the physical level were often neglected, but are important, e.g. for intervention. Basic idea: find the relevant neurons/parts of the brain. Often not possible because highly distributed processes; still, for AI, it’s often a relevant question where/in which country the data are stored.
   b. Feature detection identification: the inner layers of a network become sensitive to certain features; identify these features; manipulate them to change the results
   c. Diagnostic classification: example: try to find out how complex mathematical expressions are processed.

6. Conclusions
   a. Summary of the paper
   b. Suggestion: in the future, semantic features might be less important; more can be done with models that don’t have a semantic component

Discussion points:

1. Other examples, e.g. medicine: fitness tracker
   - A fitness tracker does not make decisions, so no decision subjects
   - Executors: sometimes the machine executes the decision itself (e.g. care robots decide to serve certain people first, then do so); so the machine is the executor; same with systems in which drugs are given automatically following a certain insulin level
   - In general, the roles need not always be separated
   - There might be additional stakeholders, e.g. owners of the bank in Zednik’s example
   - When decision subjects are interested in the why, they are often interested in a justification, not just an explanation: why did my neighbor get this treatment, but me not? Etc.
2. A basic thesis of the paper is that decision subjects are not interested in the algorithmic level and related how-questions. Is this true?
   - In general, we can distinguish two kinds of explanations: answering why-questions at the same level (provide causes, typically in the past, at the same level of description in which effect is described); answering how-questions – these explanations often move to a different level, “microlevel”, describe an “underlying mechanism” or so. In science, both kinds of explanations are interesting
   - Common argument: decision subjects have to trust machines; this is only possible with some sort of understanding/explanations
      i. Knowledge of the how can build trust
      ii. However, there are other methods of validating ML methods; e.g. test the model using test data (not the same as training data)
      iii. Analogy to medicine: in this field, randomized controlled trial experiments are deemed sufficient for testing a drug; at least, for a long time, there was no knowledge of how the drugs work in the body
      iv. Still, you only run a controlled trial experiment based on prior knowledge. However, this knowledge need not be knowledge how the process works
      v. We often trust because we can defer to experts who know how things work; e.g. cars; we trust them, but not because we know in detail how they work; rather, we trust the experts.
      vi. Still, in medicine, things seems to be different; there, experts don’t know how the drugs work at the microlevel

vii. A similar point can be made regarding ML: the opacity means that even experts have problems explaining how the models work; still, the models can be validated based on their success

viii. Note also that the reasoning of medical doctors is highly opaque to others

ix. As a matter of fact, some people trust machines less than humans; but this may change if machines become more common

- In general, there is an interesting debate: How much explanation do we want to require from machines? Many people argue that responsible use of AI requires explainability; accordingly, computer scientists have inquired how to make AI more explainable. At the same time, one may argue that the request for explainability is overstated, that too much explainability is effectively an impediment for new technologies

3. Another interesting use case is science. Science has the aim to understand real-world phenomena. Can it do so although its own methods or models are highly opaque? Some people deny this. Still, with a clever use of AI and ML, one can make some progress in understanding