Zerilli: Explaining Machine Learning (Zerilli (2022)).

Robert Michels*

December 10, 2024

1 Summary

Some of the main arguments and claims from the paper:

- Zerilli's focus is on XAI (Explainable AI), a movement which aims to solve the problem of epistemic opacity by developing and implementing general strategies for explaining why an AI does what it does
- XAI prioritizes interpretability over comprehensiveness of an explanation – i.e. a good approach to XAI need not necessarily explain an AI's complete behaviour; the focus is rather on getting explanations which are understandable for a regular human being
- Zerilli's idea is to rely on Daniel Dennett's intentional systems theory in XAI
- Dennett distinguishes between three different stances one might take (i.e. different explanatory frameworks which one might adopt) to explain reasons for actions which justify a particular action or decision:

*robert.michels@edu.ulisboa.pt

  - *Intentional stance*: 'The intentional stance is the strategy of interpreting the behaviour of an entity (person, animal, artifact, whatever) by treating it as if it were a rational agent who governed its 'choice' of 'action' by a 'consideration' of its 'beliefs' and 'desires' (Dennett, cited from Zerilli (2022), p. 6.)'
  - *Physical stance*: 'the standard laborious method of the physical sciences in which we use whatever we know about the laws of physics and the physical constitution of the things in question to devise our prediction' (Dennett, cited from Zerilli (2022), p. 6.)
  - *Design stance*: this stance focuses on 'abstract, essentially functional, features [to] explain how the parts of a system contribute to the end in view of which the system exhibits design' (Zerilli (2022), p. 7.)

- Zerilli's main claim is that adopting the intentional stance towards AI's may be a viable XAI strategy, giving us understandable explanations of why AIs make certain decisions

- Such explanations provide us with analogies of what might be going on inside an epistemically opaque AI in terms of the folk-psychological vocabulary of beliefs, desires, choices, actions, and decisions

- Example: Intentional stance explanation of why an image recognition AI trained to distinguish between wolves and dogs misclassified some of them: attributing it the following intensional actions: 'I believe that snow in an image indicates a wolf; image X had snow in it, so I concluded it must be a wolf' (Zerilli (2022), p. 15.)

- Zerilli argues for the viability of this strategy by a) pointing out that often a folk psychological explanation of this sort of the behaviour of a human doing the same task as an AI would be satisfactory, and b) that these sorts of explanations sometimes have certain advantages (brevity, predictive accuracy), and c) by applying it to particular examples

2 Discussion

Some points raised in our discussion:

- It may be problematic to talk of AI decisions, since the ability to take decisions presupposes a lot in terms of causation (one has to be able to do otherwise for it to be a decision at all) and morality (being able to take all but the very simplest decisions ('deciding' that the picture shows a wolf vs political decisions, etc.) has a moral dimension which is not present in AIs)
- Something about intentional stance explanations which may be problematic in general is that the kind of intentional stances we have are part of our human nature (biology, evolution,...); AIs are just very different, definitely not human kinds of things – why should it be generally helpful to apply human
- Zerilli doesn't claim that the intentional stance provides us with perfect explanations, but it is important to note that the kind of explanation is can provide is limited – e.g. maybe only good if the task is very easy and success conditions are clear; not applicable for General Artificial Intelligence, which may be cognitively superior to us
- His proposal may be thought of as 'the best we can get' instead of the best we hope for
- Danger of autonomous AI 'decisions' – e.g. false positive for a missile attack by Russian military AI – maybe better to see AI as a source of information which may be used to inform a human decision, rather than as an autonomous decider

References

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