Review of “Explain the Black Box for the Sake of Science: The Scientific Method in the Era of Generative Artificial Intelligence” by Gianmarco Mengaldo

The paper discusses how the scientific method changes in the era of AI and explores opportunities offered by Explainable AI (XAI) for discovering new knowledge. It defines criteria for “XAI for science”. It also points out some shortcomings of current XAI methods and mentions a few promising future developments.

I find the general idea of the paper, the re-interpretation of the scientific method in the era of AI, interesting and intriguing. The paper is well written, clear, it is fun to read. However, I do think that there are a couple of significant shortcomings of the paper – (1) the paper is too optimistic of what AI tools can deliver, (2) important aspects are not discussed in depth or are left out. Thus the discussion feels limited in depth. I expect more in-depth discussions from a paper having the impressive title “Explain the Black Box for the Sake of Science: The Scientific Method in the Era of Generative Artificial Intelligence” and coming up with the new concept of “XAI for science”. This summarises basically my main criticism, I provide more details below.

General comments

1. I find the general message of the paper too optimistic considering the current abilities of the XAI tools and their pitfalls. This current pitfalls, however, do no mean that XAI cannot provide viable solutions in future scientific research. But then this is more a future goal instead of an assessment of the current state. The paper should place greater emphasis on the future development of XAI by offering recommendations and outlining possible trajectories to guide its role in science. What must be achieved to ensure a reliable 'XAI for science'?

2. The paper promises a lot, but fails to convince in case of certain crucial aspects. It says that it discusses the “what” and “why” questions (“how” seems to be out of reach because it is not discussed in the paper). However, the “what” question that can be answered currently (and that is answered in the example discussed) is only a “glimpse” (terminology used in the paper) of the actual “what” question, i.e. what principles has the machine learned and used to obtain certain results. The “why” question as it is now in the paper is misleading and should be rephrased. By comparing the machine view with the human / GenAI view (as illustrated in Fig 1) one cannot answer “why the machine deemed those data important”, as phrased in the paper, but instead only “why **we** (or GenAI) **think** that the machine deemed those data important”. This difference is crucial. The first version suggest that we can indeed find out why the machine deemed those data important, even when we use only post-hoc interpretability methods, whereas the latter version expresses clearly that we can only come up with **new hypothesis** based on XAI methods. These hypothesis have to be then verified and proven, as it is done in the classical scientific process. As it is formulated currently in the paper, the why question is actually equivalent with the “how” question, which is not discussed in the paper.

3. The paper states that by comparing interpretability-guided explanation with existing human knowledge, it may: (i) generate nothing new or (ii) yield new knowledge. There is however a third option: **generate false "knowledge"**. Although mentioned in the last section, this third option is not stated here (L 166-167). It should be stated here as well for the sake of completeness.

4. Divergence between the machine view and the human / gen AI view can lead to new scientific knowledge. It can lead however also to false "knowledge". How to decide whether we should trust the new "knowledge" or not? What strategies, techniques could we use? This is not discussed in the paper, although it is a crucial aspect.

5. Also related to the point above, the author writes that the AI results "should present viable features that could connect to existing knowledge". I think this should be explained better. The new knowledge might not connect well to existing knowledge. How to differentiate between right and false new knowledge especially when the knowledge found by the machine does not connect well with the existing knowledge?

6. The author states that "XAI may also alleviate some of the risks that we may face when using AI for scientific discovery, that we share with Messeri and Crockett (Messeri and Crockett, 2024)." I would be interested in how exactly XAI could alleviate the illusions of "explanatory depth", "exploratory breadth" and "objectivity"? I can even think of ways XAI strengthening these illusions: XAI offers an explanation, which might suggest there is no reason for more in-depth analysis (illusions of "explanatory depth"), XAI is an AI method, thus it rules out hypotheses not testable with AI (illusion of "exploratory breadth"), it suggests objectivity, but it is data specific (illusion of "objectivity"). While the author states "we share the concerns" of Messeri and Crockett these are not further examined, and XAI is instead suggested as a remedy, which may seem like an overly simplistic response to a complex challenge. It would provide more depth to the paper if it would discuss this issue in more detail.

7. The paper defines foundational **pillars for scientific XAI**: accuracy, reproducibility, understandability. While I agree with the importance of these concepts, the author does not discuss how these criteria could be assessed and quantified. Related to the described example, the author writes that after applying post-hoc interpretability, one can "assess that the relevance maps produced satisfy the ARU requirements", but then the discussion of the example stops with just providing these maps. I think the most important step is missing from this example:

    a. What decision do we take based on the provided relevance map and how? Is this finding "nothing new", "new knowledge" or a spurious result?

    b. Are the ARU requirements satisfied and how to check that?

    Going deeper into these questions, also using the example provided in the paper, would make the paper more useful for the Earth Science community, from a practical perspective.

Specific comments

1. Related to the translation to actionable knowledge: from current AI methods we mainly obtain “hints” for hypotheses regarding principles, and then humans or generative AI can generate possible interpretations. Some comments on possible errors in this last step could be useful.
2. **Generalisability** of data-driven findings is different from generalisability in sense of the classical scientific method, i.e. mathematical and physical principles. AI models are trained using a subset of the data, and if the model learns well, its skills are generalisable to another part of the data on which the model was not trained on. However, this data has still very similar characteristics with the training data. If the model is applied then to a different dataset the “principles” might not be valid anymore, at least fine-tuning is needed. Thus, this generalisability is at a lower level compared with what is possible in the classical scientific method, when we find general mathematical and physical principles which we can apply to all possible datasets and unsees. I would appreciate it if the paper would at least shortly discuss this difference.
3. L. 27 “Identification of these principles” suggest that the principles are identified by humans, which is not the case. It should be pointed out that the identification is meant from the perspective of the machines.