I commend Gianmarco Mengaldo's efforts in addressing what is a rapidly developing issue within (Earth) science – the role of artificial intelligence. "Explain the Black Box for the Sake of Science: The Scientific Method in the Era of Generative Artificial Intelligence" is a reasoned and serious attempt at addressing some of the opportunities and risks arising from the rapid ascent of a new generation of AI tools.

Given the subject matter, the paper could be deeply philosophical. e.g What differentiates 'the scientific method' from other epistemic frameworks? What is the scientific method? What is science? Depending on discipline, school of thought, or dogma, very different answer could be supplied. Mengaldo's flavour of Popperianism would, arguably, be widely shared within the Earth science community. Latourian's may well disagree. For the purposes of this paper I would say let them. But it needs to be acknowledged that there is not a single, agreed understanding of science and this may matter given the subject matter. Indeed, it is my opinion that it does matter for this manuscript as I argue below.

More generally, and philosophical issues aside, great care needs to be taken in defining terms. How, what, why are used in the manuscript as centrally important concepts. How and why could become muddled. To help avoid that, I would suggest a (clearer) treatment of teleology is provided. Or at least, let us hear what are Mengaldo's assumptions are on the matters of intentionality when it comes to both the scientific method and the operation of particular AI algorithms.

It would also be useful to clearly state the particular types of AI that are exclusively discussed. Perhaps in an alternative universe in which AI research was not entirely dominated by large language models and other connectionist algorithms and (often very large) data sets, the very basis of Mengaldo thesis would be moot. We would not need some sort of intermediary scientific process in order to 'understand' AI algorithms, because symbolic approaches would be – in some respects by design – much more transparent. But we do not live in that universe.

I found it odd that there was not more of a discussion about the nature of the AI algorithms and some of the mathematical basis for connectionist approaches. I do not think it is safe to assume that Earth system scientists have such knowledge. Indeed, unless I have misunderstood, the problem that Mengaldo seeks to address is that connectionist approaches

produce black box algorithms because the approximation of functions via connection interactions/weights means that not only can we not use them to render a model that 'is as simple as possible, but no simpler', but the way the algorithm operates in a very high dimensional space means it can be utterly incomprehensible to humans. The latest generations of large transformer models have billions of parameters. The structures in data these algorithms are finding or creating doesn't 'mean' anything to us. Hence the need for XAI.

Mengaldo's ambition is to use insights from *how* algorithms are using *what* data to produce specific outputs. This isn't just a sensitivity analysis. I understand the motivation here to address that deep epistemic issue outlined immediately above. If one adopts an instrumentalist stance to science then perhaps' Mengaldo's task becomes a tiny bit less Herculean. AI algorithms are models, all models are wrong, some are more useful. We can define utility in a number of ways. This could short-circuit any convoluted discussions around causation. I was not very convinced by how that concept was treated in this manuscript. How is (non)linear regression associated with causation? Does it matter? Mengaldo does indeed refer to some relevant literature, but I think this raises more questions than it answers.

Mengaldo discusses accuracy, reproducibility, and *understandability*. It is understandability that I think Mengaldo wishes to address with XAI. Connectionist models are sometimes already doing a better job than process-based models with respect to some Earth science with regards accuracy. Reproducibility is certainly an issue when it comes to non-analogue conditions, when (empirical) data begins to significantly move away from the training data (e.g. non-analogue models of climate). But there is then the issue of how do they do what they do? That may really be incomprehensible, but could we still glean important *information* from these algorithms? Important information in this respect would be how this information could be used to inform *process-based* models of the phenomena of interest. The example Mengaldo uses discusses temperature and/or precipitation.

This understanding matters because comprehension is an important element of trust. Yes, an AI may reliably produce highly accurate output, but if we cannot understand how it has used certain data and how resilient that algorithm is to different data, then we may feel limited in our abilities to trust it. In my more pessimistic moments I wonder if at some point in the not

too distant future, such concerns will be considered as quaint. The raw computational power and size of connectionist models will means highly accurate and reliable outputs will be produced. A constant supply of miracles before breakfast. As to the question of the how, why bother? In that respect, induction would have won. These models outputs are true because they are always right. Black swans be damned.

Perhaps underneath these discussions there is an issue of comprehension or in some sense tractability. Mengaldo begins the discussion with Newton's famous law of gravitation. One way of telling the history of science is that we started with the easy problems – the low hanging and then falling fruit when it comes to gravity – and have more recently been struggling with complexity. It may be the case that much of the natural phenomena that we are interested in simply cannot be described using such elegant formalism. Algorithms like backpropagation can be shown to – given certain conditions – approximate polynomial functions. Understanding in terms of formalism can be preserved. But what these models are effectively doing is producing polynomials that have so many terms, are so complex that we cannot relate them to any processes – we cannot understand them. XAI is motivated to bridge that gap, but it may prove too profound a chasm to scale. In effect, we would have replaced one complex difficult to understand system (a natural phenomenon) with an engineered system that ultimately proves as *effectively* as complex and difficult to understand. What of scientific progress then? Does it stop?