Penultimate draft, accepted for publication in Philosophy & Technology. Available open access at https://link.springer.com/article/10.1007/s13347-022-00558-8

# INTEGRATING ARTIFICIAL INTELLIGENCE IN SCIENTIFIC PRACTICE: EXPLICABLE AI AS AN INTERFACE

Emanuele Ratti[1,2]

**Abstract.** A recent article by Herzog provides a much-needed integration of ethical and epistemological arguments in favor of explicable AI (XAI) in medicine. In this short piece, I suggest a way in which his epistemological intuition of XAI as "explanatory interface" can be further developed to delineate the relation between AI tools and scientific research.

**Keywords**: explainable AI; explicable AI; scientific practice; diagnosis

In "On the Ethical and Epistemological Utility of Explicable AI in Medicine" (2022), Herzog provides a much-needed integration between epistemological and ethical arguments in favor of explicable AI (XAI) in medical practice. The integration starts from his terminological choices. Herzog uses the term 'explicability' rather than just 'explainability' because the former "combines the desiderata to effectively communicate information to human agents and to do so in a manner that allows accountable use" (p 5). This means that explicability captures his interest in looking at both the epistemological and ethical aspects of XAI. In Section 3, he defends XAI against well-known arguments, while Section 4 provides strong ethical and epistemological reasons for favoring XAI in the medical context. The epistemological argument is based on the idea of 'research-practice feedback': XAI can serve as a link between research and practice. One example he discusses is that XAI, *as an explanatory interface*, "can indicate the grounds on which it recommends certain medical interventions, allowing physicians to disagree and go for alternative treatment options" (p 17). In addition, XAI can be useful in improving physicians' shared decision-making with patients. From an ethical point of view, XAI not only improves patient compliance, but also fosters patient autonomy "by allowing more individual decision-making and, hence, lifestyle-compatible interventions" (p

[1] Institute of Philosophy and Scientific Method, Johannes Kepler University Linz, mnl.ratti@gmail.com
[2] Department of Arts and Humanities, Technion Israel Institute of Technology

22). All in all, these epistemological and ethical benefits XAI will in principle result in better health outcomes.

I am very sympathetic with Herzog's views. This commentary is an attempt to delve into some of the issues raised by his project. His ethical analysis is timely and stimulating, but here I am more interested in his epistemological considerations. In particular, I find the claim that XAI is an explanatory interface very compelling, but in need of a more encompassing framing that can be adapted to other scientific contexts beyond the clinical context. The examples used by Herzog in the article are useful, but it is possible to zoom-out and sketch a framework that would make the epistemological role of XAI clearer. Here I suggest a way in which the epistemological role of XAI as an interface in science and scientific research can be understood.

In order to introduce my view, let me start by going through the goals that are usually ascribed to XAI tools. We can distinguish between proximate and ultimate goals. An example of an ultimate goal would be to say that, in principle, XAI can make AI tools more *trustworthy*. However, it seems to me that such a claim should not be accepted lightly. This is because 'trust', at least to my knowledge, applies to relationships between human beings (Kelly 2018). This is apparent in the very simple fact that violations of trust are not just disappointing: they are blameworthy. When we trust, we expose our own vulnerabilities to someone. We expect those individuals we entrust with such vulnerabilities to respect our own weaknesses and care for them as theirs. This is why trustworthiness is central in many professions, medicine in particular. But, unlike human beings, artifacts in principle do not respect nor care for anything. In fact, it would be odd to demand respect and care from tools like AIs. An alternative ultimate goal might be *reliability*: XAI can, in principle, give us assurance that AI tools are reliable. This is based on the idea that we "do not trust unanimated objects, we rely on them" (van Wynsberghe and Robbins 2019, p 727). This is intuitively plausible, but there is not much discussion on the criteria we should use to evaluate the performances of XAI models, and how these can be related to the performance of AI tools themselves (Watson 2022)[3]. Things are not better if we turn to proximate goals. In fact, it is not clear whether XAI should explain data-generating processes or the way AI tools have generated a certain output (Watson 2022). From these brief considerations, it looks like that XAI is a promising tool in need of a goal. Herzog identifies more precise ethical and epistemological goals for XAI in the medical context. Here

[3] In fact, the reliability of AI tools may just be a matter of whether they have been designed following good practices and/or they meet certain performance metrics (Ratti and Graves 2022; London 2019). Admittedly, there is a great deal of disagreement on these issues.

I make a further step by considering XAI in scientific research and scientific practice in general. I claim, developing Herzog's suggestions, that the goal of XAI in scientific research and practice is to integrate AI tools in the wider scientific context in which such tools are employed. But what does this mean exactly?

Works integrating philosophy of science and philosophy of technology can be very useful here. In discussing role functions of engineering artifacts in the context of mechanistic philosophy, van Eck (2015) distinguishes between *effect function* and *purpose function*. For instance, the effect function of an electric screwdriver can be to loosen/tighten screws, while its purpose can be to facilitate the process of hanging paintings in my living room. As van Eck argues, effect functions are internal to the artifact, while purposes are external. Effect functions can be instrumental to purpose functions, in the sense that an effect can contribute to the realization of a purpose, which should be seen as the role that an artifact plays in a wider context. But sometimes the effect alone is not sufficient to realize the purpose. In the case of medical AI tools, effect functions are often simple outputs, e.g. binary classifications. But effect functions do not usually exhaust what AI tools are supposed to provide. Often, AI tools have a purpose that goes beyond the mere effect or output, in the sense that the effect alone does not ensure that the AI tool will realize its purpose, which is the role it is supposed to play in a wider scientific (or medical) process.

Let me illustrate my point with a concrete example. Diagnosing a disease is a complex medical process that involve several phases, including a medical interview, a physical examination, and the use of diagnostic testing tools (e.g. bloodwork, x-ray exams, etc). These different, iterative phases are navigated through diagnostic reasoning, which often consists in going from effects to causes. AI tools can be used in all these phases, most notably in (but not limited to) interpreting the results of diagnostic tools. In many cases, this is the purpose of AI tools: they assist clinicians or pathologists in assessing evidence gathered through tests (i.e. pathophysiological reasoning). An example of an AI tool for pathologists is Lymph Node Assistant (LYNA), a classificatory tool developed by a team at Google AI Healthcare (Liu et al 2019). LYNA has proved to be very effective in distinguishing normal lymph node from metastatic lymph node, a task that is complex even for trained pathologists. However, the mere effect – the output 'metastatic/not metastatic' – is not enough to achieve the purpose, namely to assist pathologists in contextualizing laboratory evidence with evidence gathered in the other phases of the diagnostic process. In order for a tool like LYNA to really assist diagnosis, its classifications have to be delivered in a way that are integrated in the wider 'medical culture' of diagnosis. Such an epistemological culture includes theoretical frameworks, concepts, and perceptual activities employed in the complex diagnostic processes. Here is where XAI models can be helpful: by providing ‘explanatory interfaces’ (e.g. post hoc ‘explanations’), XAI models aid diagnostic reasoning. For instance, LYNA’s explanatory interface can foster ‘perceptive interpretability’ through saliency “by computing the amount that each pixel affects LYNA’s output prediction” (Liu et al 2019, p 861). Through this, LYNA show which parts of the actual slides were important to make the classification, thereby helping the pathologist to make an actual connection between physiological entities and the condition of the patient. XAI models can indicate, for instance, pleomorphic cells, which have recognizable nuclei, and they often characterize carcinoma of the breast. Knowing not only that something is metastatic, but also the type of cells can be useful for the iterative process of diagnosing a disease and ruling out some hypotheses while prioritizing others. It is important to point out that LYNA’s predictions, even when complemented by XAI models, should not be definitive: they just point to one type of evidence among others. There are several techniques that tools like LYNA can use (see for instance Zednik and Boelsen 2022 for an overview), but the whole point is that XAI models can potentially play the role of facilitating the realization of the purpose of AI tools (e.g. assisting diagnosis) well beyond the mere effect (e.g. the classificatory output). This is what XAI models can do: they can integrate AI tools into a wider scientific process or system of practice in a way that AI tools realize their purpose. We can even distinguish between effect and purpose of XAI models: the effect is to make AI tools more interpretable, but the purpose is to integrate AI tools in, for instance, a scientific reasoning process. In other words, the goal of a XAI model, or its epistemological role in science, is to facilitate the realization of AI’s purpose functions in a wider system of scientific research. This is compatible with Herzog’s important insight that XAI can be conceived as an *interface*, in the sense that it is a tool through which two unrelated systems or entities (e.g. the AI tool and a scientist) interact fruitfully.

Let me close with a final consideration. I am not saying that XAI models are currently understood in this way; rather, I am pointing to one way in which their epistemological role in scientific research can be conceptualized, clarified, and developed. But this idea of XAI models as aligning AI tools to a wider scientific process or practice is, in my opinion, a general characterization of what Herzog describes in his article. There are other works that, to me, seem to go in the direction I have briefly sketched. For instance, Zednik and Boelsen (2022) show how XAI models can contribute to scientific exploration “by facilitating the task of refining target phenomena” (p 225) and by identifying “starting points for future inquiry” (p 228). This is to say that XAI tools can go well beyond not-well-defined explanatory tasks, and that they can actually facilitate the integration of AI tools in scientific research.

## REFERENCES

Herzog, C. (2022). On the Ethical and Epistemological Utility of Explicable AI in Medicine. *Philosophy & Technology*, *35*(2), 50. https://doi.org/10.1007/s13347-022-00546-y

Kelly, T. (2018). *Professional Ethics*. Lexington Books.

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, *49*(1), 15–21. https://doi.org/10.1002/hast.973

Liu, Y., Kohlberger, T., Norouzi, M., Dahl, G. E., Smith, J. L., Mohtashamian, A., Olson, N., Peng, L. H., Hipp, J. D., & Stumpe, M. C. (2019). Artificial intelligence–based breast cancer nodal metastasis detection insights into the black box for pathologists. *Archives of Pathology and Laboratory Medicine*, *143*(7), 859–868. https://doi.org/10.5858/arpa.2018-0147-OA

Ratti, E., & Graves, M. (2022). Explainable machine learning practices: opening another black box for reliable medical AI. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00141-z

van Eck, D. (2015). Mechanistic explanation in engineering science. *European Journal for Philosophy of Science*, *5*(3), 349–375. https://doi.org/10.1007/s13194-015-0111-3

van Wynsberghe, A., Robbins, S. (2019). "Critiquing the Reasons for Making Artificial Moral Agents." *Science and Engineering Ethics* 25 (3). Springer Netherlands: 719–35. doi:10.1007/s11948-018-0030-8

Watson, D. S. (2022). Conceptual challenges for interpretable machine learning. *Synthese*, *200*(1). https://doi.org/10.1007/s11229-022-03485-5

Zednik, C., & Boelsen, H. (2022). Scientific Exploration and Explainable Artificial Intelligence. *Minds and Machines*, *32*(1), 219–239. https://doi.org/10.1007/s11023-021-09583-6

## DECLARATIONS

### Consent for publication

Author consents to the publication

### Availability of data and material

Not applicable

### Competing Interests

No competing interests to declare

### Funding

No funding to declare

**Authors' contributions**

Not applicable