Living with Uncertainty: Full Transparency of AI isn't Needed for Epistemic Trust in AI-based Science

Uwe Peters, Utrecht University
[This is a short invited critical commentary forthcoming in
Social Epistemology Review and Reply Collective.
Comments very welcome!]

Abstract

Can AI developers be held epistemically responsible for the processing of their AI systems when these systems are epistemically opaque? And can explainable AI (XAI) provide public justificatory reasons for opaque AI systems' outputs? Koskinen (2024) gives negative answers to both questions. Here, I respond to her and argue for affirmative answers. More generally, I suggest that when considering people's uncertainty about the factors causally determining an opaque AI's output, it might be worth keeping in mind that a degree of uncertainty about conclusions is inevitable even in entirely human-based empirical science because in induction there's always a risk of getting it wrong. Keeping this in mind may help appreciate that requiring full transparency from AI systems before epistemically trusting their outputs might be unusually (and potentially overly) demanding.

Keywords: epistemic opacity; AI; XAI; uncertainty; epistemic trust

1. Introduction

Inkeri Koskinen thinks that "we do not have a satisfactory social epistemology of AI-based science". Her argument is that any available satisfactory social epistemology of science requires epistemic trust between scientific agents, and "[w]hen the AI applications used in science are epistemically opaque, there can be no rationally grounded relationships of trust, because AI applications are not appropriate trustees, and human agents cannot take appropriate responsibility of their workings, as the applications are epistemically opaque", i.e., the internal epistemically relevant details of how these applications produce their outputs remain unknown (2024, 9).

I agree that science requires epistemic trust between agents (henceforth the "necessary trust" (NT) view). But I disagree with Koskinen that when AI systems are epistemically opaque, human agents cannot take appropriate responsibility (to be epistemically trusted) for the processing of these systems. In a recent commentary (Peters, 2024), I offered several arguments against Koskinen's view. She has now replied. I think her replies don't succeed against my initial points.

2. Overlooking similarities between human and AI opacity

In her reply, Koskinen insists that her favored

> argument for the necessary trust view requires that someone is able to describe, upon request, the basis of the classifications used, and to explain the criteria for individual classifications. If we know that they are unable to do so, trust in the researcher cannot be considered rationally grounded, as they are unable to guarantee that the value-relevant choices that have happened during the classification process are acceptable. […] Therefore, a researcher using an AI application for the classification of big data cannot take informed responsibility for the choices made within the tool in a way that would make trust in their decisions rationally grounded. (2024, 12)

However, Koskinen seems to overlook that if human scientists did the classifications instead, there would also be no "guarantee" that, in their classifications, they processed only and all relevant input features, because human cognition is also epistemically opaque and irrelevant features can unconsciously influence human classifications. Yet, this doesn't undermine the view that trust in human scientific classifiers can be rationally grounded. Why not?

Koskinen claims that "scientists can provide publicly everything that other scientists need to be able to evaluate whether the conclusion is valid" (2024, 12). However, if "providing publicly everything that other scientists need" for this purpose includes an accurate account of all epistemically relevant internal processes that determined the conclusion, then Koskinen's claim is false (scientists' minds are epistemically opaque). If it doesn't include this, however, then it's no longer clear what the social epistemological problem (for the NT view) with AI-based science would be, as the key property that Koskinen claims to be the source of the problem, i.e., epistemic opacity, has dropped out of the picture.

So, what exactly is meant by "everything researchers would need to be able to evaluate whether the result is valid"? If after viewing a patient's brain scans, a neuroradiologist supports her preliminary diagnosis of Alzheimer's by (a) highlighting the amyloid plaques in a specific brain region, (b) noting that she's been trained in line with rigorous scientific norms on thousands of Alzheimer's patient scans, and (c) clarifying that she's been highly accurate in all her previous brain scan classifications, is this "everything" that's needed? Arguably, (a) to (c) do provide strong justificatory reasons for the neuroradiologist's conclusion. But then opaque AI models that are supplemented with XAI systems may in principle provide the relevant sort of public justifications too, because some XAI systems are already routinely trained to produce explanations of AI outputs that (e.g.) highlight the most salient features that a classifier used, that may specify details of the model's training, that indicate model accuracy, and that are increasingly more faithful to the approximated model's reasoning (Mariotti et al., 2023). Koskinen claims:

> XAI analyses [...] do not offer everything researchers would need to be able to evaluate whether the result is valid – in other words, they do not provide justificatory reasons. An XAI analysis of how an AI application classified a large data set does not offer researchers even the theoretical possibility of being able to check whether classification decisions are morally and socially acceptable. (2024, 13)

But this isn't quite right. Common post hoc XAI models do provide AI developers and researchers with ways of checking whether automatic classifiers produce decisions that are moral or socially acceptable by, in some cases, providing similar sorts of "justificatory reasons" as the neuroradiologist reports in the example above. For instance, LIME or SHAP can show how important a specific input feature was for an opaque AI to produce a particular output, allowing developers and users to assess whether sensitive features (e.g., gender, race) unduly affected an opaque model's classifications (Yang et al., 2022; Ali et al., 2023).

Koskinen seems to assume that AI developers are completely in the dark on whether XAI models track the real, causally determinative features underlying an opaque AI's output. They aren't. AI developers often use various methods to determine if the features highlighted by XAI systems are causally relevant. For instance, they can use (1) cross validation and consistency checks (e.g., use LIME with different models trained on the same data and see if the explanation differs), (2) comparative analyses (e.g., if LIME, SHAP, Grad-CAM, etc. all highlight similar features, this can increase confidence in highlighted features), (3) perturbation tests (e.g., mask the highlighted features and observe the impact on the model's performance; performance decline would suggest these features are important), (4) causal inference techniques (e.g., produce counterfactual instances where certain features are changed and check effects on the model's output, or explicitly model and test the causal relationships between features and

outcomes), (5) robustness analysis (e.g., test LIME on different subsets of the data or add noise to the input features to see if the model's predictions and the features highlighted by LIME remain stable), and (6) sensitivity analyses (e.g., test how sensitive the model's output is to small perturbations in highlighted features) (Nauta et al., 2023). No single XAI method is foolproof and the faithfulness of XAI outputs may be limited (Slack et al., 2020). But a multi-method approach can provide strong validation, providing confidence that the features identified by a given XAI method are determinative of an opaque AI's outputs.

Granted, the relevant XAI models produce only post hoc explanations that can be viewed as "rationalizations", which I construed neutrally as retrospective descriptions of the basis of information processing aimed at producing the best explanation of the available evidence (e.g., input and output data) (Peters, 2024). But this needn't be an epistemological problem. Providing a rationalization (thus understood) can be "providing everything that other scientists need to be able to evaluate whether a conclusion is valid", because it may allow other scientists to assess whether a conclusion is valid (e.g., a gradient-based post hoc salience map might correctly highlight some features in an image that are irrelevant to an opaque AI's classification thus indicating that the classification process is likely not valid; Ribeiro et al., 2016). Post hoc XAI outputs can hence be public justificatory reasons.

Moreover, even if XAI methods produce idealizations, i.e., distortions of real-world features that are present in a model or theory, this can be compatible with and useful in science (Sullivan, 2024). For instance, frictionless planes in physics or perfectly rational agents in economics are idealizations but seen as helpful and successful scientific tools (e.g., they facilitate distinguishing relevant from irrelevant features), meaning that strict fidelity to the truth isn't always desirable in science and that requiring XAI to be fully faithful may be uncalled for and even scientifically disadvantageous (ibid).

Correspondingly, Koskinen's claim that "[t]here is no parity" between "post-hoc rationalizations offered by XAI models and the demand that scientists publicly justify their conclusions" is questionable (2024, 13). It overlooks that (a) post hoc XAI outputs can have a robust scientifically acceptable evidential basis or be successful idealizations, and that (b) when scientists provide everything that others need to evaluate a conclusion's validity, what they provide might often also only be rationalizations with limited faithfulness: since their minds are opaque, neither they nor we could tell for sure whether what they report were the only and all factors causally determining their (e.g.) classifications. I think the epistemic opacity of scientists' minds provides a *reductio* of Koskinen's favored version of the NT view because this view implies (absurdly) that no scientist can be trusted. This *reductio*, in turn, supports the alternative version of the NT view that I proposed in my commentary and that I'll now revisit.

3. A mischaracterization

Koskinen writes: "Peters argues that it is enough for the researcher to have verified the reliability of the application used" (2024, 11). I didn't say this. I think model reliability is just one property needed for scientists to epistemically trust AI developers that their opaque models are appropriate for use in science. Alignment of the models' processing with other general scientific norms, including explainability, is needed too. I wrote that for scientists to epistemically trust AI developers, the developers need to ensure their models are (1) accurate, (2) verified and validated by the relevant AI experts, (3) designed and updated in collaboration with scientific experts, (4) monitored, controlled, optimized, and constrained so that their performance is reliable, (5) their post hoc explanations are understandable, plausible, consistent, and sufficiently faithful to and complete in capturing the opaque model's processing.

What 'sufficiently faithful and complete' means will need to be settled in collaborations between scientists and AI developers and may differ depending on what's at stake in a scientific study that plans to use opaque

AI. Moreover, it might include that the XAI methods used in science have passed an evaluation of whether they engage in successful idealizations or deceptive explanations of opaque AI outputs (Sullivan, 2024).

Given this, I don't deny that current XAI methods may still have a long way to go to fully satisfy these criteria. The point is simply that full transparency of all the internal processing of the model and strict faithfulness of XAI outputs aren't necessarily required for scientists to be able to epistemically trust the developers, because such a requirement would even undercut epistemic trust in scientists. Rather, what is needed is that developers use as many available methods as possible to ensure the features highlighted by XAI systems are causally relevant for opaque models' outputs (e.g., cross validation, comparative analyses, perturbation tests, causal inference techniques, etc.) such that scientists can draw robust inferences on whether the models' classifications are morally, socially, and epistemically acceptable. My suggestion was and is that *if* AI developers act in these ways, scientists can epistemically trust them, and we have a satisfactory social epistemology of AI-based science, namely a version of the NT view.

A degree of uncertainty about the basis of opaque AI outputs will likely remain with any post hoc XAI. But in the next section, I argue that any satisfactory social epistemology of science will need to allow for some degree of uncertainty even only of human scientific conclusions anyway, meaning that this shouldn't undercut epistemic trust, provided the uncertainty at issue is rigorously minimized through accepted methods.

4. Absolving AI developers from epistemic responsibility

In my earlier commentary, I noted that Koskinen's view that human agents can't take responsibility for the processing of AI systems when these systems are epistemically opaque has the unattractive upshot of enabling AI developers to deny responsibility for their AI systems (by insisting the systems are epistemically opaque) even when they clearly have it. Koskinen now replied:

> the concern is easily addressed. The kind of ability to take informed epistemic responsibility that the necessary trust view requires from scientists is not something that moral or legal responsibility would generally require. In fact, I would say that AI developers are morally and undoubtedly in some cases legally responsible for developing tools for which they are unable to take informed epistemic responsibility. (2024, 13)

The thought seems to be that AI developers can be held *morally* responsible for their AI systems even if these systems are epistemically opaque. They just can't take informed *epistemic* responsibility and so can't be held epistemically responsible for them in such cases.

But this simply shifts the initial problem to informed epistemic responsibility: Koskinen's view allows AI developers to deny epistemic responsibility for their opaque models, even though they *are* epistemically responsible for them. Why? Because even if their systems are opaque to them, AI developers can design, monitor, and control these systems so that their development and processing follow accepted epistemic norms and procedures, and the systems' outputs are accurate and reliable (all of which are epistemic properties). Since it is the developers' choice, and in their power, to design, monitor, etc. their AI systems in ways that make them less epistemically harmful and their processing more conform with epistemic norms, AI developers are epistemically (not only morally) responsible for their models even if they are opaque. Koskinen's view obscures this.

Again, AI developers have only limited knowledge of whether their opaque models use only and all relevant input features for their classifications (predictions, etc.). However, through the selection of training data, the model training regime, model testing (e.g., through adversarial attacks), and debugging

cycles, AI developers shape the final AI product and determine the structures that constrain their models' feature selection for classifications. For instance, after training an opaque AI image classifier, developers may (a) apply SHAP values to understand the influence of each input feature to individual classifications, (b) determine features that have little or no impact on the classifications, (c) consider removing or changing them, (d) retrain the model with the altered feature set, and (e) re-assess the model using SHAP again to confirm the changes (Lundberg & Lee, 2017). Since AI developers channel their models' learning, processing, feature detection, etc., they can with some certainty fix and describe significant parts of the basis of individual classifications and so can be held epistemically responsible for them even if a degree of uncertainty about the basis of their AI systems' outputs remains.

This uncertainty needn't be a problem for AI developers to take "informed" epistemic responsibility, because some uncertainty about conclusions is inevitable even in entirely human-based empirical science: Empirical science inevitably involves incomplete evidence, meaning there's always some chance of false positives or false negatives (Elliot & Richards, 2017). Consequently, "informed" epistemic responsibility and trust even in only human-based (empirical) science must be compatible with some degree of incompleteness and uncertainty about the truth of a scientific agent's or tool's claims. The fact that AI developers face some uncertainty about the causally determinative factors in their AI models' processing needn't prevent them from taking informed epistemic responsibility for the AI's processing and be held epistemically responsible, because a degree of uncertainty about scientific conclusions also doesn't prevent empirical scientists from being held epistemically responsible for their conclusions.

To add a final point, Koskinen seems to think that epistemic trust and responsibility are binary in that one can either epistemically trust or take responsibility or one can't. But I think that since uncertainty about the basis of (e.g.) classifications (whether by humans or AI) is gradable, it may be more useful to think of epistemic trust and responsibility as gradable (varying with the level of uncertainty) too. Correspondingly, the NT view might be given a gradualist interpretation as well such that the trust necessary for science is taken to come in degrees. On this view, AI-based research would satisfy the necessary trust condition (provided the criteria outlined above are met) albeit to a potentially lower degree than solely human-based research.

5. Conclusion

Koskinen insists that on her favored version of the NT view, if agents or tools are epistemically opaque, they can't figure in relationships of epistemic trust that are required for science. This version of the NT view is (a) a non-starter because it implies that even entirely human-based science can't be sufficiently epistemically trusted (as humans too are epistemically opaque), and (b) risks absolving AI developers from epistemic responsibility for their AI systems that they clearly have.

Koskinen proposes that scientists can, despite being epistemically opaque, be epistemically trusted because they can provide public justificatory reasons that enable other researchers to evaluate whether their results are valid. But I noted that AI-based science then needn't be a problem even on Koskinen's favored version of the NT view because at least some XAI models can in principle also provide such reasons. A degree of uncertainty about the accuracy of XAI outputs will remain. But uncertainty can't be completely avoided even when it comes to scientists' empirical conclusions. Scientists become epistemically trustworthy when they minimize the uncertainty about their claims by using robust methods to do so. Similarly, if AI developers make efforts to reduce the uncertainty about the features causally determining their opaque models' outputs by combining rigours XAI methods, they too can be epistemically trusted in ways that the NT view requires for AI-based science.

References

Ali, S., Akhlaq, F., Imran, A. S., Kastrati, Z., Daudpota, S. M., & Moosa, M. (2023). The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. *Computers in biology and medicine*, 166, 107555. https://doi.org/10.1016/j.compbiomed.2023.107555

Elliott, K.C. & Richards, T. (2017) (eds.). *Exploring inductive risk: case studies of values in science*. Oxford: Oxford University Press.

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2022). Explainable AI Methods - A Brief Overview. *xxAI - Beyond Explainable AI* (pp.13-38)

Koskinen, I. (2024). We Still Have No Satisfactory Social Epistemology of AI-Based Science: A Response to Peters. *Social Epistemology Review and Reply Collective* 13 (5): 9–14.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017) (pp. 4765-4774).

Mariotti, E., Sivaprasad, A., Moral, J.M.A. (2023). Beyond Prediction Similarity: ShapGAP for Evaluating Faithful Surrogate Models in XAI. In: Longo, L. (eds) *Explainable Artificial Intelligence*. xAI 2023. Communications in Computer and Information Science, vol 1901. Springer, Cham.

McMullin, E. (1992). *The Inference that Makes Science*. Milwaukee WI: Marquette University Press.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput.* Surv. 55, 13s, 295.

Peters, U. (2024). Science Based on Artificial Intelligence Need not Pose a Social Epistemological Problem. *Social Epistemology Review and Reply Collective* 13 (1): 58–66.

Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You?': Explaining the Predictions of Any Classifier. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135– 44. New York: Association for Computing Machinery.

Slack, D. Hilgard, S., Jia, E., Singh, S. & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 180-186.

Sullivan, E. (2024). SIDEs: Separating Idealization from Deceptive 'Explanations' in xAI. *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '24), June 3–6, 2024, Rio de Janeiro, Brazil. ACM, New York, NY, USA.

Yang, G., Ye, Q., & Xia, J. (2022). Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *An international journal on information fusion*, 77, 29–52.