SERRC
Social Epistemology
Review & Reply Collective

http://social-epistemology.com
ISSN: 2471-9560

Science Based on Artificial Intelligence Need not Pose a Social Epistemological Problem

Uwe Peters, Utrecht University, u.peters@uu.nl

Peters, Uwe. 2024. "Science Based on Artificial Intelligence Need not Pose a Social Epistemological Problem." *Social Epistemology Review and Reply Collective* 13 (1): 58–66. https://wp.me/p1Bfg0-8vp.

SERRC
Social Epistemology
Review & Reply Collective

13 (1): 58–66. 2024.
https://wp.me/p1Bfg0-8vp

Abstract

It has been argued that our currently most satisfactory social epistemology of science can't account for science that is based on artificial intelligence (AI) because this social epistemology requires trust between scientists that can take full responsibility for the research tools they use, and scientists can't take full responsibility for the AI tools they use since these systems are epistemically opaque. I think this argument overlooks that much AI-based science can be done without opaque models, and that agents can take full responsibility for the systems they use even if these systems are opaque. Requiring that an agent fully understand how a system works is an untenably strong condition for that agent to take full responsibility for the system and risks absolving AI developers from responsibility for their products. AI-based science need not create trust-related social epistemological problems if we keep in mind that what makes both individual scientists and their use of AI systems trustworthy isn't full transparency of the internal processing but their adherence to social and institutional norms that ensure that scientific claims can be trusted.

AI is now often used in different domains of science. What may be the implications for the social epistemology of science?

Inkeri Koskinen (2023) argues that according to our currently best supported social epistemology of science, i.e., the "necessary trust" (NT) view, collective scientific knowledge production requires trust between agents who can take responsibility for their actions and research tools because scientific knowledge is largely produced collectively, and scientists depend on each other's expertise. Yet, Koskinen continues, scientists can't trust AI tools in this way because these systems aren't responsible agents and, crucially, they are "epistemically opaque", i.e., it "is impossible" for scientists to "know all the epistemically relevant elements" affecting AI outputs (7).

It might seem that AI developers or operators are agents that could take responsibility for the relevant systems, thus ensuring a basis for trust in AI-based science. However, for Koskinen, even for AI developers or operators, these systems remain opaque, and so even they can't take full responsibility for the systems in the way the NT view requires. Since the NT view is currently our best social epistemology of science and this view can't account for AI-based science, "we currently have no satisfactory social epistemology of AI-based science", Koskinen concludes (15).

There's much to agree with in her paper. But I'm less pessimistic and think that there are some problematic internalist assumptions underlying Koskinen's argument. If we remove these assumptions then the version of the NT view that is left needn't have a problem accounting for AI-based science.

Some Background on AI Systems

AI comprises many different computer systems, including machine learning (ML) systems, deep learning models, and neural networks. ML systems are a subset of AI, deep learning is a subset of ML, and neural networks constitute the backbone of deep learning algorithms (which, unlike a single neural network, have more than three network node layers) (Bigelow 2023). Not all of these systems are epistemically opaque. In fact, many ML systems used in science, for instance, linear and logistic regression, decision trees, general additive, or Bayesian models, are transparent, hence fully explainable to scientists (Arrieta et al. 2020). Since much AI-based science (e.g., science relying on traditional ML) isn't dependent on opaque systems and so needn't pose a problem for the NT view, claiming that "we currently have no satisfactory social epistemology of AI-based science" (Koskinen 2023, 15) is perhaps too strong, as it may suggest that all instances of AI-based science are problematic for the NT view.

Koskinen seems primarily concerned with the most powerful AI systems (e.g., deep neural networks), which tend to be opaque. But why not avoid these systems in science and use transparent ones instead? Koskinen suggests that "reducing the opacity of AI applications […] risks reducing their usefulness", i.e., predictive performance (9). However, this needn't be the case. Simpler, transparent models often match or exceed the performance of more complex ones (e.g., Gosiewska et al. 2021; Janela and Bajorath 2022), leading some AI experts to hold that it is "a myth that there is necessarily a trade-off between accuracy and interpretability": When the data are structured, with a good representation in terms of naturally meaningful features, there is "often no significant difference in performance between more complex classifiers (deep neural networks, boosted decision trees, random forests) and much simpler classifiers (logistic regression, decision lists) after preprocessing" (Rudin 2019, 206–207). If scientists explore whether they can replace computationally complex classifiers for their tasks with simpler ones, this may increase trust in AI-based science, and help the environment: the development of some computationally complex models can produce massive carbon footprints (e.g., the training of one DL model may cause CO_2 emissions five times the lifetime emissions of an average American car; Hao 2019).

Perhaps not all opaque AI models that are useful in science can be replaced with transparent ones without performance loss. Would the use of these models in science be problematic because we don't fully understand how they produce their outputs? Koskinen doesn't consider this question. She just focuses on determining whether the NT view can account for AI-based science. Nonetheless, it might be worth noting that AI models can be unproblematic for scientific knowledge production even if they are opaque. They could be used in the "context of discovery" as a "source of inspiration" for scientific investigation: Their outputs may help researchers conceive new ideas, hypotheses, or theories (Krenn et al. 2022). In science, the origin of an idea is commonly irrelevant because it doesn't affect the 'context of justification', i.e., the part of science where data are analyzed, and ideas evidentially supported. Hence, focusing on the context of discovery, using opaque AI systems can be unproblematic in science (Duede 2023).

In fact, even in the context of justification, scientists often allow claims with opaque origin to have evidential weight and therefore be able to make a difference in justifying scientific

SERRC
Social Epistemology
Review & Reply Collective
13 (1): 58–66. 2024.
https://wp.me/p1Bfg0-8vp

belief formation. Consider speculations in science, i.e., assertions whose epistemic status is uncertain due to insufficient evidential support. Despite their lack of support and potentially opaque origin, speculations can (apart from generating new ideas) provide opportunities to link together existing but seemingly unconnected knowledge areas, thus providing incremental additional support for hypotheses (e.g., see defences of 'story-telling' in historical reconstruction; Currie 2023). That is, speculations can have a "confirmatory function", aiding in "overcoming local underdetermination by forming scaffolds from which new evidence becomes relevant" (Currie and Sterelny 2017, 14). Since speculative claims can, independently of their origin, enter scientific justification by increasing evidential power through tying together existing knowledge branches, and since outputs of opaque AI may act as such claims, these outputs, too, may in some cases enter scientific justification. Hence, even in the context of justification, outputs of opaque AI models can sometimes be unproblematic in science.

Returning to the NT view, if the NT view implies that the source of information that shapes epistemic justification in science needs to be fully transparent, this view may already have a problem accounting for the currently accepted use of speculations in science. But I don't think the NT view implies such thing. This is why I find Koskinen's argument not entirely convincing. The next section unpacks the point.

What Does 'Taking Responsibility' Mean?

The NT view says, according to Koskinen, that "relationships of trust are a prerequisite for […] contemporary science" and these relationships need to hold between agents, or between agents and tools such that the agents can take full responsibility for the tools (4). But advocates of the NT view may mean different things by 'taking full responsibility'. Here are two ways in which one may 'take responsibility' for an instrument or system S that isn't itself responsible for its behavior:

> *Notion (1).* For an agent A to take responsibility for S, A needs to have a full understanding of how S works internally, i.e., of "all the epistemically relevant elements" in the "computations" that lead to S's output(s) (7).
>
> *Notion (2).* For A to take responsibility for S, A needs to be able to design, monitor, control S so that its computation and processing follows accepted procedures in the design and is reliable.

Koskinen's argument relies on notion (1). In a sentence that doesn't describe the NT view but expresses her own perspective, she writes that:

> When AI applications that are essentially epistemically opaque are used in scientific knowledge production, the scientists who use them are epistemically dependent on the instruments, but there is no accountable agent who could fully grasp how the instrument works and therefore be able to take full responsibility of it, and thus no one the scientist could trust, in a rationally grounded way, to have all the relevant knowledge (7–8).

Koskinen seems to make fully grasping how the instrument internally works a necessary condition for the ability to take full responsibility for the instrument, and she (earlier) clarifies that this responsibility is a moral (not merely epistemic) one (4). She then ascribes this notion to the NT view. However, as far as I can see, this condition isn't implied by the NT view even with its 'thick' concept of trust. What (at least some) advocates of (some version of) the NT view seem to say is just that in scientifically trustworthy human-tool hybrids, tools "are 'the *designed product* of a good deal of epistemic work by others who *are* epistemic subjects in their own right'"—a view that Koskinen herself takes to be "in line with the accepted view in the epistemological literature on trust: when we cross a bridge, it is not the bridge we trust, but the engineers and builders responsible for it" (6). Whether we trust the engineers because they can take moral responsibility for their products in the sense of notion (1) is another question. Koskinen suggests that advocates of the NT view are committed to this notion. But I think this notion is untenable and so perhaps not something advocates would or need to accept. Here are two examples.

Example (1)

Since the focus is on an instrument or system S that isn't itself responsible for what it's doing, consider, instead of an instrument, a biological system that isn't held responsible. Consider children. "Children are typically not yet considered morally responsible for what they do" (Brandenburg 2022, 472). That is, despite displaying some agency, very young children aren't yet what Koskinen calls "agents who can take responsibility for their actions" (2).

Suppose, then, you have a son aged 3. It seems plausible that in some cases you don't fully grasp how or why he acts the way he does, and, clearly, his brain is more complex and so computationally opaque than any existing AI. Yet, despite your lack of a full understanding of all epistemically relevant elements in the 'computations' that lead to your son's behavior, if his behavior results in damage to other people, as long as he is under the age of 14, you will (commonly) be legally liable for it. By common moral and legal standards, young children aren't morally responsible for their actions. Their caretakers are. According to these standards, then, people can (and need to) take moral responsibility for some system S without fully grasping all the epistemically relevant elements that lead to S's responses. Hence, holding that for A to take full moral responsibility for S, A needs to fully understand how S works (at the computational level)[1] is too strong.

Young children aren't scientific instruments. So, here's another example.

Example (2)

Recall, for Koskinen, when scientists use an opaque AI instrument, there is "no accountable agent who could fully grasp how the instrument works and therefore be able to take full responsibility of it, and thus no one the scientist could trust, in a rationally grounded way, to

[1] You might understand your son's behavior by explaining it via the intentional stance, i.e., via ascribing mental states to him. However, many popular explainable AI systems can also give intentional stance explanations of opaque AI models' outputs (Peters 2023).

SERRC
Social Epistemology
Review & Reply Collective
13 (1): 58–66. 2024.
https://wp.me/p1Bfg0-8vp

have all the relevant knowledge" (7–8). Now, Koskinen also rightly notes that "epistemic and cognitive processes of individual researchers are not fully transparent even to themselves. Our cognitive capabilities are limited, and, for instance, it is not possible for an individual researcher to be fully aware of all the background assumptions on which they build their work" (3).

Suppose, then, we replace the term 'instrument' in the first statement with 'individual researcher'. We'll get: 'there is no accountable agent who could fully grasp how the individual researcher works and therefore be able to take full responsibility of the individual researcher, and thus no one the scientist could trust, in a rationally grounded way, to have all the relevant knowledge'. Granted, unlike in the AI case, there are now two agents. But while both may take responsibility for their action, neither fully grasps how the 'instrument' involved works, as the individual researcher isn't fully transparent to herself or others. Hence, if, as Koskinen assumes, A can only take responsibility for S if A fully understands all epistemically relevant elements that lead to S's output, there could never be an agent that can take full responsibility for an individual researcher's 'output', not even the researcher herself, which seems absurd.

Advocates of the NT view may therefore be unlikely to endorse Koskinen's notion of 'taking responsibility'. But if, for instance, the second, weaker notion is endorsed, then the NT view needn't have a problem in accounting for AI-based science because AI developers and operators can design, monitor, control their models so that their processing follows accepted procedures in the design and is reliable.

Aren't There Differences?

It might be suggested that the notion (2) of 'taking responsibility' yields a concept of trust too thin for advocates of the NT view and overlooks that scientists trust in each other because, unlike opaque AI models, scientists can publicly justify their conclusions with reasons. But here's how advocates of the NT could respond to this concern: They may note that while researchers could perhaps often provide reasons for their 'outputs', this doesn't mean these reasons are the factors that determined the 'outputs' because (again) the epistemic and cognitive processes of individual researchers aren't fully transparent to themselves. Indeed, given their own opacity, researchers may often (unwittingly) offer post-hoc rationalizations that seem compelling because they emerge from pressure to provide claims that are plausible, consistent, and aligned with existing data. Importantly, many existing explainable AI (XAI) systems can do likewise for opaque models' outputs, i.e., they can provide post-hoc rationalizations for the outputs (see Binns et al. 2018). Koskinen seems to overlook this parity when she writes:

> While various epistemically opaque processes happening in human minds influence scientific knowledge production in many ways, the justification of a claim […] has to be public, as it has to be scrutinisable. But if a claim is the result of an epistemically opaque process within an AI application, there may be no way to produce an independent, public justification—a more or less accurate, post-hoc analysis of some of the central features of the opaque

> process may well be all we can have. If such claims are accepted and used in science, the processes happening within epistemically opaque AI applications are treated differently than processes happening in the epistemically opaque minds of human beings (13).

This overlooks that if individual researchers are opaque to themselves and others, then for them, too, often 'there may be no way to produce an independent, public justification—a more or less accurate, post-hoc analysis of some of the central features of the opaque process may well be all we can have'. Yet, we commonly accept and use in science such claims provided they are plausible, consistent, and sufficiently aligned with existing data. Withholding acceptance, in these cases, may look like science denialism, as we might be dealing with abduction (i.e., inferences to the best explanation), when the conclusions may not follow logically from the premises, or even from all the information one has, but be warranted because, if true, they best explain the relevant phenomena. Given this, if we also accept post-hoc analyses from XAI systems (that supplement opaque AI) then—provided they are plausible, consistent, checked against existing findings, sufficiently faithful to the approximated model's reasoning, and the opaque model's outputs are verified, validated, designed in collaboration with scientific experts—we will not (*pace* Koskinen) treat their processing differently than the processing in humans. But then, to the extent that the NT view is adjusted so as to avoid making the relevant explanatory requirements higher for opaque models than for scientists, the NT view may accommodate (opaque) AI-based science.

I do not deny that differences will remain between the two cases and explanations. The point here is simply that, in both cases (researcher vs. AI), epistemic transparency of *internal processing* and so public justifiability of outputs is inevitably limited. To note one difference, unlike current XAI models, individual researchers can still regulate their behavior in ways that align their future thinking and acting with their explanations (Peters 2023). However, in the scenarios relevant here, there are humans that can take over the regulative interventions on the opaque AI models, i.e., the AI developers. They can monitor, control, optimize and constrain their models such that their performance is reliable, and their post-hoc rationalizations are plausible, consistent, tested against previous findings, and sufficiently faithful (Kroll 2018). Developers have mechanisms to 'police' opaque AI models' outputs just like scientific communities can 'police' individual researchers' outputs. These mechanisms include verification, validation methods, reproducibility, robustness analyses, histories of implementations, expert knowledge as sources for attributing reliability to computations, but also, for scientists, methodological training, reporting guidelines, checklists, and so on. It therefore seems that with respect to external constraints that may ground epistemic justification, (opaque) AI-based and human-based science may not differ much.

In sum, Koskinen assumes that the NT view implies that scientists can only sufficiently trust each other (or their tools) if they can take full responsibility for their conclusions where this implies understanding all the epistemically relevant elements in the computations that lead them (or their tools) to their conclusions. However, *if* the NT view of human-based science had this implication, it would be a non-starter, as neither individual researchers nor their peers have the required understanding to take the relevant responsibility. But I think, unlike Koskinen seems to assume, the NT view doesn't have to have this implication. Advocates of

13 (1): 58–66. 2024.
https://wp.me/p1Bfg0-8vp

the NT view may adopt notion (2) of 'taking responsibility' instead. In fact, I think it's precisely because many scientists have known all along that human minds are opaque that the just mentioned external policing mechanisms were put in place to ensure that, even without full transparency about all epistemically relevant factors, scientific claims can be trustworthy. But anything that is external to the 'computations' in this way can perhaps, in principle, also be put in place by AI developers in the AI training or fine-tuning and then be checked by independent AI auditing bodies, ensuring that the outputs of opaque systems are sufficiently justified for scientific knowledge production.

Perhaps the external constraints that AI developers currently impose on opaque AI and the relevant auditing or validation bodies aren't yet as comprehensive and powerful, respectively, as their equivalents for individual scientists' belief formation and research practices. Social epistemologists thinking about the use of opaque AI systems in science are rightly concerned. But the concern seems misdirected and potentially harmful if it focuses too much on internal opacity in AI. Relatedly, Koskinen's assumption that, since there "is no accountable agent who could fully grasp how the [opaque AI] instrument works", there is no agent "able to take full responsibility of it" (7) seems to absolve the AI designers, controllers, companies that build and field the AI from full responsibility for their products when they clearly have it. After all, how AI models work is the result of choices by these designers, controllers, and companies. Obscuring this can result in designer complacency and models that are less thoroughly checked than they could and should be.

A Way Forward

I think AI-based science doesn't pose new epistemological problems because the relevant AI is internally opaque but rather because AI developers, controllers, companies, and AI-using scientists haven't yet collectively established governance frameworks that formalize and make known the design assumptions, choices, and adequacy determinations linked to particular scientific AI systems (Kroll, 2018). This requires cooperation between AI experts and scientists on how the AI will function in the field, the metrics employed for performance tests, and the procedures for model auditing to avoid biases and errors from going unmitigated. What is needed, then, for trust in science that uses opaque AI systems is explanations specifying why the systems are reliable, task appropriate, and well calibrated to their context. These externalist explanations need to specify the design decisions during a system's development and the system's technical details, showing why we should trust the developers and companies designing it, what efforts they took to avoid errors, and how they included scientific users in the AI development (Theunissen and Browning 2022). If this happens, then perhaps even the NT view can account for science that is based on opaque AI, provided this view ties 'thick' trust among scientists to a realistic notion of taking responsibility.

References

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera. 2020. "Explainable Artificial Intelligence

(XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." *Information Fusion* 58: 82–115.
Bigelow, Stphen J. 2023. "Machine Learning vs. Neural Networks: What's the Difference?" *Tech Target*. https://www.techtarget.com/searchenterpriseai/answer/Machine-learning-vs-neural-networks-Whats-the-difference.
Binns, Reuben, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, Nigel Shadbolt "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions." *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* 1–14. doi: 10.48550/arXiv.1801.10408.
Brandenburg, Daphne. 2021. "Consequentialism and the Responsibility of Children: A Forward- Looking Distinction between the Responsibility of Children and Adults." *The Monist* 104 (4): 471–483.
Currie, Adrian. 2023. "Science & Speculation." *Erkenntnis* 88: 597–619.
Currie, Adrian and Kim Sterelny. 2017. "In Defence of Story-Telling." *Studies in History and Philosophy of Science Part A* 62: 14–21.
Duede, Eamon. 2023. "Deep Learning Opacity in Scientific Discovery." *Philosophy of Science* 90 (5): 1089–1099. doi: 10.1017/psa.2023.8.
Gosiewska, Alicja, Anna Kozak, Przemyslaw Biecek. 2021. "Simpler is Better: Lifting Interpretability-Performance Trade-Off Via Automated Feature Engineering." *Decision Support Systems* 150 (113556). doi: 10.1016/j.dss.2021.113556.
Hao, Karen. 2019. "Training a Single AI Model Can Emit as Much Carbon as Five Cars in Their Lifetimes." *MIT Technology Review*. https://www.technologyreview.com/2019/06/06/ 239031/training-a-single-ai-model-can-emit- as-much-carbon-as-five-cars-in-their-lifetimes/.
Janela, Tiago and Jürgen Bajorath. 2022. "Simple Nearest-Neighbour Analysis Meets the Accuracy of Compound Potency Predictions Using Complex Machine Learning Models." *Nature Machine Intelligence* 4 (12). doi: 10.1038/s42256-022-00581-6.
Koskinen, Inkeri. 2023. "We Have No Satisfactory Social Epistemology of AI-Based Science." *Social Epistemology* 1–18. doi: 10.1080/02691728.2023.2286253.
Krenn, Mario, Robert Pollice, Si Yue Guo, Matteo Aldeghi, Alba Cervera-Lierta, Pascal Friederich, Gabriel dos Passos Gomes, Florian Häse, Adrian Jinich, Akshat Kumar Nigam, Zhenpeng Yao, Alán Aspuru-Guzik. 2022. "On Scientific Understanding with Artificial Intelligence." *Nature Reviews Physics* 4: 761–769.
Kroll, Joshua A. 2018. "The Fallacy of Inscrutability." *Philosophical Transactions of the Royal Society Series A: Mathematical, Physical, And Engineering Sciences* 376 (2133): 1–14. https://doi.org/10.1098/rsta.2018.0084.
Peters, Uwe. 2023. "Explainable AI Lacks Regulative Reasons: Why AI and Human Decision-Making are not Equally Opaque." *AI Ethics* 3: 963–974.
Peters, Uwe. 2022. "Reclaiming Control: Extended Mindreading and the Tracking of Digital Footprints." *Social Epistemology* 36 (3): 267–282.
Rudin Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1 (5): 206–215.
Theunissen, Mark and Jacob Browning. 2022. "Putting Explainable AI in Context: Institutional Explanations for Medical AI." *Ethics and Information Technology* 24 (23): 1–10. https://doi.org/10.1007/s10676-022- 09649-8.

SERRC
Social Epistemology
Review & Reply Collective
13 (1): 58–66. 2024.
https://wp.me/p1Bfg0-8vp