Philosophy of Science (2026), **00**, 1–11
doi:10.1017/psa.2026.10246

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

SYMPOSIA PAPER

Against epistemic transparency of algorithms

Juan Manuel Durán

Technology, Policy and Management, Delft University of Technology, Delft, Netherlands
Email: j.m.duran@tudelft.nl

(Received 14 April 2025 UTC; revised 15 June 2026 UTC; accepted 16 June 2026 UTC)

Abstract

Epistemic transparency is often proposed as a solution to algorithmic opacity, wherein revealing the inner logic of an algorithm provides reasons or supporting evidence for the justification of its outputs. I argue that transparency is defective in belief formation and thus inadequate as an epistemology of algorithms. Two objections are developed: *transparency regress* and *bootstrapping*, both rooted in transparency's status as a time-sliced, "outsourced" inferentialist epistemology. While weaker forms—such as contextual transparency—are briefly considered, the central argument remains that transparency fails to provide justification and should be abandoned as an epistemology of algorithms.

1. Introduction

It is widely accepted that algorithms are epistemically opaque. Opacity is here understood in line with Humphreys' definition, whereby the cognitive limitations of human agents prevent them from knowing the epistemically relevant elements of an algorithm that would justify belief in a given output (Humphreys 2009, 618). Now, there would be no special concern about opaque algorithms if it were not for the fact that they are central to many scientific endeavors, progressively displacing humans from the center of knowledge production. The opacity of algorithms, then, triggers significant epistemic anxiety.

To address opacity, many have sought to counteract this lack of knowledge. It is in this context that *transparency* gains traction. Creel, for instance, takes this idea in its strictest form. According to this author, opacity and transparency are "two sides of the same coin: opacity is a lack of transparency and vice versa" (Creel 2020, 569, fn. 1). As we will see, many have folded under this view. And while transparency encompasses a range of methodologies and strategies, there is a common justificatory aim, that is, to provide reasons or supporting evidence for believing the algorithm's output. From where is this justification drawn? From identifying the functions, variables, values, and similar epistemically relevant elements within the algorithm that generate the output. This justificatory aim has been collectively referred to as "showing the inner logic of the algorithm" (Wachter et al. 2018, 843). For example, BenevolentAI identified *baricitinib* as

© The Author(s), 2026. Published by Cambridge University Press on behalf of Philosophy of Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psa.2026.10246 Published online by Cambridge University Press

an effective drug to combat COVID-19 symptoms. For a researcher to be justified in this belief, transparency shows how BenevolentAI internally operates: how it identifies drug-signal blocking, maps causal relations between viruses and drugs, and extracts relevant information from specialized literature (BenevolentAI 2022). In this article, I shall keep the metaphor.

The starting point, then, is that under transparency, individual beliefs about an algorithm's output gain justification by being supported with reasons or evidence. These, as noted, derive from the inner logic of the algorithm—functions, variables, etc.—that are epistemically responsible for the output. The goal is to show that transparency is an inadequate epistemology of algorithms, as it leads to defective belief formation. This is, of course, not to say that transparency is undesirable for other epistemic purposes. It is widely accepted that some level of transparency is required for explanation (Watson and Floridi 2021; Durán 2021), understanding (Sullivan 2022; Räz and Beisbart 2024; Páez 2024), and representation (Kieval 2026; Freiesleben 2026), among others.

Here, the concern is exclusively with the justificatory role of transparency. To this end, and following its main proponents, section 2 situates transparency within an evidentialist epistemology. This is followed by section 3, which discusses *algorithmic regress* and *bootstrapping* as two core obstacles for the proper formation of beliefs. Section 3.3 then addresses the root of the issue, arguing that transparency constitutes an "outsourced," time-sliced, inferentialist epistemology. In section 4, a new form of transparency is introduced—one that incorporates the context within which a belief is justified. Further concerns are raised for proponents of *contextual transparency*. I conclude by suggesting that transparency be abandoned in favor of reliabilism.

This article introduces some new notation to simplify the argument. It uses $\mathscr{A}$, $\mathscr{B}$, etc. to denote any type of opaque algorithm, and $\bar{o}$ for its output. It also adopts a representationalist standpoint, where $\bar{o}$ represents—or fails to represent, to a degree determined by the relevant epistemic community—a fact F, or a fact that entails F. No (anti-)realist commitments should be inferred from this notation or these working assumptions. This is simply a maneuver aimed at simplifying an already complex issue.

2. The evidentialist flavor of transparency

Transparency emerges as a response to the opacity of algorithms—epistemic, methodological, and representational. The chief concern is that algorithms may not fully participate in scientific knowledge unless we know how their outputs are generated. Transparency addresses this concern by seeking reasons or evidential support for believing the algorithm's output. This is achieved either by directly making visible the algorithm's inner logic, or by externally reconstructing it. Here are some examples.

Proponents of explainable AI (XAI) have made significant progress toward making visible the inner logic of algorithms. For some, key explanatory questions—such as "Which factors influence a given output?" and "How can the algorithm's behavior be reconstructed?"—are paths to secure justification (Zednik and Verreault-Julien 2025). Consider the "explanation game" (Watson and Floridi 2021), a causal-manipulationist approach in which $\bar{o}$ is explained in virtue of answering *w*-questions about how $\mathscr{A}$ would respond to hypothetical scenarios in the vicinity of $\bar{o}$, under conditions of

https://doi.org/10.1017/psa.2026.10246 Published online by Cambridge University Press

simplicity, accuracy, and relevance. According to the authors, "with the right tools [the explanatory game], we may conclusively reveal the true reasoning behind consequential decisions [of algorithms]" (Watson and Floridi 2021, 9236).

Counterfactual explanations operate in a similar way (e.g., Wachter et al. 2018). These explanations exhibit patterns of counterfactual dependence, describing how $\bar{o}$ could have changed had the input variables, values, or internal procedures of $\mathscr{A}$ been different. Similarly, popular XAI methods such as Local Interpretable Model-agnostic Explanations (LIME) approximate the behavior of a complex model by generating local surrogate models around a given instance (Ribeiro et al. 2016). These surrogates—typically sparse linear models—are designed to capture how slight perturbations to the input affect the output. While considered explanatory, the primary function of LIME is to provide insight into the local decision boundaries of the model, thereby offering a simplified representation of how the output is produced in a specific region of the input space.

Finally, transparency is also pursued through *Qualitative Input Influence* (QII) methods, which measure the degree of joint influence that input variables exert on outputs. A *transparency report* is produced to specify how particular profile variables and functions (e.g., age, education) influence the prediction. The ultimate goal is to provide "access to the decision-making system and knowledge of the input dataset on which it [the algorithm] operates" (Datta et al. 2016, 599).

On the philosophical front, transparency concerns our lack of knowledge about how $\mathscr{A}$ generates $\bar{o}$. Boge, for instance, is explicit on this point: "I take it for granted that epistemic opacity is relative to an agent and involves a lack of knowledge"[1] (Boge 2022, 15), adding that its treatment "may concern all three forms of transparency in complex computational systems identified by Creel" (Boge 2022, 15, fn. 18). The tripartite structure referred to here comes from Creel's categorization of various levels of transparency: *functional transparency*, *structural transparency*, and *run transparency*. In all three cases, knowledge arises from making visible the relevant epistemic elements of the algorithm at each corresponding level: "knowledge of the algorithmic functioning of the whole," "knowledge of how the algorithm was realized in code," and "knowledge of the program as it was actually run in a particular instance, including the hardware and input data used" (Creel 2020, 569–70).

Zerilli (2022) also thinks in line with these authors, advocating for "fathomability," that is, the degree to which one can immediately grasp the relationship between features in the model and its predictions. Humphreys, in turn, articulates a notion that aligns closely with his account of epistemic opacity, where a representation $\bar{o}$ is transparent to an agent "in case the states of a system or process are represented in a way that is open to explicit scrutiny, analysis, interpretation, and understanding by [that agent]" (Humphreys unpublished, 10).

Now, transparency functions as an epistemology in virtue of the role it has been assigned in belief formation: to provide reasons or evidential support that an algorithmic prediction represents a fact in the world. This is, in so many words, what

[1] Boge distinguishes between *h-opacity*, concerned with "how the machine learns" (Boge 2022, 15), and *w-opacity*, focused on "what was learned by the machine" (Boge 2022, 16). The objections to transparency discussed here apply to both forms of opacity.

https://doi.org/10.1017/psa.2026.10246 Published online by Cambridge University Press

Creel affirms with respect to LIME: "because the diagnostic systems do not give an explanation or reason for the diagnosis, doctors often deem them untrustworthy and avoid them. Applying LIME to such a system gives doctors a rationale for the existing system's diagnoses" (Creel 2020, 583).[2] This view is also upheld by Guidotti and colleagues, who emphasize the role of transparency in revealing the reasons behind a decision or an event (Guidotti et al. 2018, 93:2). Similar views are also found in Ribeiro et al. (2016), 1135 and Burrell (2016), 9, but also in the cited work of Watson and Floridi (2021) and Zerilli (2022), among many others.

As a working assumption, I hold that, methodologically, transparency consists in showing the inner logic of the algorithm that generated $\bar{o}$; epistemically, transparency provides reasons or supporting evidence to form the belief that $\bar{o}$ represents, to the degree permissible by the relevant community, a fact in the world. In this sense, transparency functions as an evidentialist epistemology for the justification of beliefs. I argue that transparency is epistemically inadequate as an epistemology of algorithms for belief formation.

3. Against transparency as an epistemology of algorithms

3.1. Transparency regress

Transparency regress is a form of foundational skepticism: the problem lies in deciding on a transparent algorithm that can serve as epistemically foundational without appealing to further instances of transparency. To illustrate this, consider $\mathscr{A}$, a medical machine learning system whose complexity exceeds any feasible surveyability of its inner logic. Suppose that $\mathscr{A}$ generates the output $\bar{o}$ = "the probability that m is a melanoma is x%." What considerations give us reasons to believe that $\bar{o}$ represents, to the intended extent, a fact about the world? The partisan of transparency, S, proposes a third-party algorithm—call it $\mathscr{B}$—with the capacity to reveal $\mathscr{A}$'s inner logic; that is, to provide S with reasons to believe that $\bar{o}$ (Guidotti et al. 2018). One possible toy reconstruction is as follows:

$$\begin{aligned}\mathscr{B}(\mathbf{input}) = \mathscr{A}\\ \mathscr{B}(size) = \texttt{size > 6mm};\\ \mathscr{B}(shape) = \texttt{shape == asymmetrical};\\ \ldots;\\ \mathscr{B}(borders) = \texttt{borders == irregular}\end{aligned}$$

Equation 1. Example of algorithmic transparency in melanoma detection.

S then presents evidence supporting that $\bar{o}$ = "the probability that m is a melanoma is x%" represents a fact in the world. But does it? Transparency regress is the claim that the process by which S obtained the supporting evidence has not yet been secured, as $\mathscr{B}$ is itself an opaque algorithm. In view of this, S must either admit that $\mathscr{B}$ is epistemically unproblematic (e.g., not opaque), or accept that another

[2] Creel has also linked the absence of reasons with a lack of transparency; see Creel (2020, 579).

https://doi.org/10.1017/psa.2026.10246 Published online by Cambridge University Press

algorithm—call it $\mathscr{B}_0$—is needed to reveal the inner logic of $\mathscr{B}$ itself. The former option is, in principle, untenable for *S*, as it amounts to an admission that transparency requires an arbitrary stopping point in the regress, and thus cannot adequately secure justification. The latter option, on the other hand, places *S* in the uncomfortable position of shifting the question from the transparency of $\mathscr{A}$ to that of $\mathscr{B}$, thereby committing to a transparency chain:

$$\mathscr{B}_n \rightarrow \mathscr{B}_{n-1} \rightarrow \cdots \rightarrow \mathscr{B}_0 \rightarrow \mathscr{B} \rightarrow \mathscr{A}$$

which can only terminate at some $\mathscr{B}_n$, a foundational algorithm so simple that it is epistemically secured by *S* themselves. Let me now discuss each case in turn, beginning with the scenario in which a transparency chain is required.

S's second option raises the question: have we stopped the transparency regress with a foundational algorithm? I do not think so. First, there is the general issue that transparency does not apply to an entire algorithm, but is rather target-directed, in the sense that it is only required to show the inner logic of the next algorithm in the chain. Thus, $\mathscr{B}_n$ reveals the inner logic of $\mathscr{B}_{n-1}$, which in turn reveals the inner logic of $\mathscr{B}_{n-2}$, and so on, until $\mathscr{A}_n$ provides supporting evidence that $\bar{o}$. Thus understood, *S* has one of two options. Either *S* starts from their only source of secured knowledge (i.e., $\mathscr{B}_n$) and works upwards to the transparency of $\mathscr{A}$, or begins with $\mathscr{A}$ and attempts to work down the chain, securing transparency all the way to $\mathscr{B}_n$. I believe both options are epistemically untenable and should be abandoned.

Starting from one end of the chain, *ex hypothesi*, $\mathscr{B}_n$ is epistemically accessible to *S*. This means that $\mathscr{B}_n$ constitutes secured knowledge that can confer target-directed transparency to $\mathscr{B}_{n-1}$. But of which functions, variables, and operations? Of those that generated $\mathscr{B}_{n-1}$'s output—in this case, the transparency of $\mathscr{B}_{n-2}$. And again: *S* utilizes the transparency of $\mathscr{B}_{n-2}$ to secure transparency in $\mathscr{B}_{n-3}$. This repeats until *S* secures transparency of $\mathscr{A}$ and thus supporting evidence that $\bar{o}$.

Now, for any of this to happen, *S* must have had *a priori* knowledge of the functions, variables, and mechanisms in $\mathscr{A}$ that generated $\bar{o}$—otherwise, the right transparency chain cannot be built. Indeed, unless *S* already knows that $\mathscr{B}(size) = \text{size} > 6mm$ contributes to the supporting evidence for $\bar{o}$ = "the probability that *m* is a melanoma is *x*%," *S* has no indication of how to reconstruct the target-directed transparency chain. But not only is this form of knowledge unavailable to *S*, if it were available it would render transparency entirely superfluous.

Suppose now that *S* works down the transparency chain from $\mathscr{A}$ to $\mathscr{B}_n$. This directionality has the advantage of not requiring *a priori* knowledge, as previously discussed. Instead, transparency is secured by $\mathscr{B}$, then by $\mathscr{B}_0$, and so on. The issue, however, is that *S* offers only a piecemeal justification, where transparency is secured between any two adjacent algorithms. Yet to warrant the supporting evidence for $\bar{o}$, *S* is epistemically obligated to also show that the justification holds across the entire transparency chain. That is, *S* must show that $\mathscr{B}_n$—their only source of secured knowledge—provides the right kind of supporting evidence for $\bar{o}$ across the full chain. While this presents practical difficulties, the more pressing concern is that *S* introduces a new form of opacity—this time at the level of justification. It then falls to *S* to show either that there is no specific problem in justifying $\bar{o}$ on the basis of $\mathscr{B}_n$, or that transparency can resolve this form of (meta-)opacity.

https://doi.org/10.1017/psa.2026.10246 Published online by Cambridge University Press

We are then left with the first option: that the transparency regress can be addressed only by arbitrarily stopping the regress. Under this umbrella, three main strategies have been proposed in the literature. These are:

1. $\mathscr{B}$ is epistemically non-problematic and confers transparency to $\mathscr{A}$. The algorithm's output $\bar{o}$ is justified on this basis. Examples include *interpretable predictors* and some versions of XAI (Guidotti et al. 2018).
2. $\mathscr{B}$ is epistemically non-problematic and confers transparency to $\mathscr{A}$'s output directly by approximating it using local surrogate models. Standard examples are *model agnostics*, such as LIME (Ribeiro et al. 2016).
3. $\mathscr{A}$ is transparent itself, and thus able to self-provide reasons to believe $\bar{o}$. This is the case of *interpretable models* (Rudin 2019).

Briefly, options (1) and (2) are epistemically untenable positions. Taking $\mathscr{B}$ to be epistemically unproblematic is unwarranted and arbitrary, lacking any special justificatory status. It follows that the reasons or evidence provided by these methods lack justificatory force, and thus belief formation is necessarily defective—for instance, $\bar{o}$ could represent a fact merely by epistemic luck.[3]

Furthermore, proponents of options (1) and (2) must explain why $\mathscr{B}$ is exempt from the requirement of transparency and accepted by mere epistemic *fiat*. Or, equivalently, why $\mathscr{B}$ is treated as epistemically unproblematic in the absence of any warrant for arbitrarily halting the regress.[4]

The only viable claim is that $\mathscr{B}$ itself is surveyable by S, in which case the regress is halted, as the algorithm now possesses the epistemic status required for justification. However, strictly speaking, $\mathscr{B}$ provides reasons to believe the inner workings of $\mathscr{A}$, rather than $\mathscr{A}$'s output, $\bar{o}$. To justify $\bar{o}$, an additional argument would still be required. Option (3) avoids this issue altogether: it halts the regress without necessitating any further justificatory argument. Let me now turn to this case.

3.2. Level-0 transparency and the bootstrapping of beliefs

Option (3) above has the fundamental advantage of getting rid of any intermediary for the justification. $\mathscr{A}$ can then harness its own reasons that $\bar{o}$. It is with this idea in mind that Rudin argues for a form of *algorithmic interpretability*, where S justifies that $\bar{o}$ if the right functions, variables, etc. are identified within $\mathscr{A}$. According to Rudin, then, $\mathscr{A}$ is itself transparent and thus self-provides the necessary reasons or supporting evidence that $\bar{o}$. Let me call this *level-0 transparency*.

Epistemically speaking, algorithmic interpretability is bootstrapping the same beliefs that it intends to justify. This occurs when a method justifies its own outputs by appealing to its own internal procedures.[5] Here, the source for S's justification that $\bar{o}$ are reasons provided by the inner logic of $\mathscr{A}$, the same algorithm that generated $\bar{o}$. If S's belief that $\bar{o}$ = "the probability that m is a melanoma is x%" is justified solely by the

[3] See Vandenburgh (2023); Schmidt et al. (2025); Durán (2025).

[4] Rudin (2019) also argues that methodologies such as (1) and (2) are defective, albeit for different reasons.

[5] Note that this form of bootstrapping slightly differs from the classic objection to reliabilist epistemologies, in which a method justifies its own reliability by appealing to its own outputs.

https://doi.org/10.1017/psa.2026.10246 Published online by Cambridge University Press

fact that $\mathscr{A}$ generated that output, then the justification is not independent of the process that generated $\bar{o}$—it is actually identical to it. It follows that S cannot assess epistemic warrants of $\bar{o}$ because they do not possess independent access to the functions, variables, etc. that generated $\bar{o}$. The output and functions act as both product and evidence, violating the independence condition for robust epistemic justification (e.g., Elgin 2017).[6] From an evidentialist perspective, the justification proposed by transparency is shallow.

Level-0 transparency then gives S no genuine foundation for justification. The bootstrapping that accompanies it leads to defective formation of beliefs, as no new evidence or reasons—independent from those used to generate the output—have been provided for warranting the justification. This leaves S in the same fragile epistemic position in which they began, with the justification offered being no more compelling than the initial unjustified output. This failure to provide independent support also weakens the justificatory credibility of the belief system S intends to create, leaving it as unsupported as it was to begin with. Level-0 transparency, then, fails to provide a justification that goes beyond the very thing being justified, making the formation of beliefs inherently epistemically defective. Bootstrapping shows that transparency is a self-reinforcing epistemology, illegitimately trying to justify itself using the very inferential structure and without appealing to independent (non-) doxastic sources.

3.3. The root of the problem

Transparency regress and bootstrapping are merely symptoms of a deeper issue. The root problem lies, I believe, in the theoretical assumptions under which transparency operates. Consider first that transparency constitutes an outsourced-inferentialist epistemology. Inferentialism holds that all justified beliefs must be supported by other, prior justified beliefs through reasoning or inference. Transparency adopts this framework but outsources the inferential process to the computation performed by the algorithm—rather than by a human agent. Thus, when S claims to have reasons to believe $\bar{o}$, what S is actually saying is that the algorithm has provided—via such computation—a specific set of beliefs (i.e., functions, variables, etc.) that justifies $\bar{o}$. Since computation can arguably be construed as a logic-mathematical process (Fetzer 1988), and since S neither contributes to nor can intervene in the justificatory process (see section 4 for more on this), it follows that transparency invalidates any non-inferentialist formation of beliefs (e.g., those based on testimony). This would not be problematic were it not for the fact that we do, in fact, require an epistemology capable of supporting justification through non-inferential beliefs—particularly because such beliefs play crucial epistemic and moral roles in belief formation. For example, Pozzi (2023) has convincingly argued that shared hermeneutical resources are necessary for justification in the use of medical machine learning.

Another characteristic of transparency is that it embodies a time-sliced epistemology, or synchronic justification (Hedden 2015), in the sense that justification

[6] Note that validation procedures, such as contrasting $\bar{o}$ with a "ground truth," do not amount to independent justification. Neither because justification is the epistemic goal of validation—as it aims at recalibrating the algorithm to be more accurate—nor because validation has justificatory force—it does not provide reasons to believe $\bar{o}$, just to confirm its accuracy.

https://doi.org/10.1017/psa.2026.10246 Published online by Cambridge University Press

does not consider past beliefs or diachronic coherence. Instead, reasons for belief stem from a specific execution of the algorithm's procedures and functions, utilizing a fixed set of input values—for that execution—a particular choice of parameters—for that execution—and so on. *S*'s formation of beliefs is therefore more about how beliefs are currently being sustained than about how they were originally acquired. This implies that the rational status of believing the output is evaluated at a given time-slice of its execution, rather than over the entire performance of the algorithm. Furthermore, under a time-sliced epistemology, justification is highly restricted. While a given transparency chain might justify $\bar{o}$ = "the probability that m is a melanoma is x%," it offers no epistemic warrant that the same chain could justify $\bar{q}$ = "the probability that p is a melanoma is y%." A newly tailored, case-specific transparency chain would need to be produced for $\bar{q}$.

Above all, transparency is inadequate because it buys into the epistemic opacity program wholesale. To address opacity, every epistemically relevant element in the algorithm that provides reasons or supporting evidence for its output must be made visible. "Transparency and opacity," we are told, "are two sides of the same coin" (Creel 2020). Now, this rigidness in justification is not share by every advocate of transparency. Some reject the condition that justification stems exclusively from the algorithm, thereby allowing for beliefs to be formed in relation to the context in which the algorithm is employed—while presumably still preserving transparency as the central epistemic framework. Let me now turn to this form of *contextual transparency*.

4. Contextual transparency

There is a growing sense that rigid transparency is unsuitable for the justification of $\bar{o}$, chiefly due to methodological reasons. This line of thought is found in Zednik, who states, "rendering an opaque system transparent [. . .] require[s] knowledge of the environmental patterns and regularities that are being tracked and of the abstract representational structures that are tracking them" (2021, 268), as well as in Burrell: "[m]any [. . .] take a broad socio-technical approach looking at 'algorithms in the wild.' The algorithms in question are studied for the way they are situated within a corporation, under the pressure of profit and shareholder value, and as they are applied to particular real-world user populations (and the data these populations produce). Thus, something more than the algorithmic logic is being examined" (2016, 2). This is a form of *contextual transparency*, where reasons or supporting evidence for the formation of beliefs are both internal to the algorithm and external to it—including, for instance, design decisions, environmental patterns, and shareholder values.

While I largely agree with the underlying concern, for contextual transparency to succeed, the role of contextual information in justification must be explained. If such information merely plays a corrective role in belief formation without serving as the primary basis for justification, then transparency regress and bootstrapping remain problematic. That is, if contextual information merely provides support for accepting or abandoning the belief, rather than reasons that warrant it, then it amounts to rigid transparency with extra steps. On the other hand, if it is conceded that contextual information carries greater epistemic weight than the algorithm's inner logic, then this appears to be an admission that transparency is insufficient for justification. In

https://doi.org/10.1017/psa.2026.10246 Published online by Cambridge University Press

that case, it remains to be explained why we should be interested in transparency at all—for justificatory purposes, naturally.

There is a third way, one in which contextual information and the inner logic of the algorithm both contribute to justification. But for this third way to succeed, it must be explained what justification is and how it operates, as there is no common currency of evaluation. For instance, how do non-inferential beliefs have justificatory force within an inferentialist epistemology? Or how can a strictly synchronic justification incorporate diachronically formed beliefs? I do not have the answers to these questions, nor do I think the advocate of transparency does.

5. Final remarks

I would like to conclude in the same spirit as I began, by citing Humphreys on the prospects of transparency. In his view, "we must abandon the insistence on epistemic transparency for computational science" (Humphreys 2004, 150)—a sentiment I share and have tried to convey in arguing that transparency leads to defective belief formation and is therefore inadequate for the purposes of epistemic justification.

But Humphreys goes further, maintaining that "[w]hat replaces it [transparency] would require an extended work in itself, but the prospects for success are not hopeless" (Humphreys 2004, 150). I, too, agree with this outlook. Such prospects can already be assessed in the development of externalist epistemologies of algorithms, such as my *computational reliabilism* (Durán et al. 2024; Durán 2026), and Humphreys' own *instrumental reliabilism* (Humphreys online).

It is to be expected that any epistemology of algorithms will attract criticism and raise doubts. There remains, undoubtedly, significant ground to cover. Here, I have presented two objections to transparency that ultimately cast doubt on its justificatory success, despite its influence as a proposed solution to epistemic opacity. The value of this debate lies, I believe, in helping us to understand not only the limitations of transparency as an epistemology of algorithms—often taken as unproblematic—but also in prompting reflection on viable alternatives.

Acknowledgment. I am grateful to Emanuele Ratti, Manuel Barrantes, Emma-Jane Spencer, Giorgia Pozzi, Sjoerd Zwart, Chirag Arora, Thijs Latten, Imane Ihaddouchen, and Marc Mela for many helpful discussions on this paper and the ideas surrounding it. This work was supported by the European Commission through the project "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (Grant Agreement 871042). The funders had no role in developing the research and writing the manuscript. To Kass and Diego, in infinite love.

Funding statement. Open access funding provided by Delft University of Technology.

References

BenevolentAI. 2022. "What We Do." https://www.benevolent.com/what-we-do. Accessed November 3, 2020.

Boge, Florian J. 2022. "Two Dimensions of Opacity and the Deep Learning Predicament." *Minds and Machines* 32:43–75. https://doi.org/10.1007/s11023-021-09569-4.

Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1):1–12. https://doi.org/10.1177/2053951715622512.

Creel, Kathleen A. 2020. "Transparency in Complex Computational Systems." *Philosophy of Science* 87 (4):568–89. https://doi.org/10.1086/709729.

https://doi.org/10.1017/psa.2026.10246 Published online by Cambridge University Press

Datta, Anupam, Shayak Sen, and Yair Zick. 2016. "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems." In *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 598–617. https://doi.org/10.1109/SP.2016.42.

Durán, Juan M. 2021. "Dissecting Scientific Explanation in AI (sXAI): A Case for Medicine and Healthcare." *Artificial Intelligence* 297:103498. https://doi.org/10.1016/j.artint.2021.103498.

Durán, Juan M., David van der Vloed, Arnout Ruifrok, and Rolf J. F. Ypma. 2024. "From Understanding to Justifying: Computational Reliabilism for AI-Based Forensic Evidence Evaluation." *Forensic Science International: Synergy* 9:100554. https://doi.org/10.1016/j.fsisyn.2024.100554.

Durán, Juan M. 2025. "In Defense of Reliabilist Epistemology of Algorithms." *European Journal for Philosophy of Science* 15 (2):37. https://doi.org/10.1007/s13194-025-00664-2.

Durán, Juan M. 2026. "Beyond Transparency: Computational Reliabilism as an Externalist Epistemology of Algorithms." In *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*, edited by Juan M. Durán and Giorgia Pozzi, 55–79. Cham: Springer. https://doi.org/10.1007/978-3-032-03083-2_4.

Elgin, Catherine Z. 2017. *True Enough*. Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/11118.001.0001.

Fetzer, James H. 1988. "Program Verification: The Very Idea." *Communications of the ACM* 31 (9):1048–63. https://doi.org/10.1145/48529.48530.

Freiesleben, Timo. 2026. "Artificial Neural Nets and the Representation of Human Concepts." In *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*, edited by Juan M. Durán and Giorgia Pozzi, 273–94. Cham: Springer. https://doi.org/10.1007/978-3-032-03083-2_13.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. "A Survey of Methods for Explaining Black Box Models." *ACM Computing Surveys* 51 (5):93. https://doi.org/10.1145/3236009.

Hedden, Brian. 2015. *Reasons Without Persons: Rationality, Identity, and Time*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198732594.001.0001.

Humphreys, Paul. 2004. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford: Oxford University Press. https://doi.org/10.1093/0195158709.001.0001.

Humphreys, Paul. 2009. "The Philosophical Novelty of Computer Simulation Methods." *Synthese* 169 (3):615–26. https://doi.org/10.1007/s11229-008-9435-2.

Humphreys, Paul. 2020. "Predictive Failures in Neural Nets." Available at: https://philsciindia.wixsite.com/lecture-series/post/8-predictive-failures-in-neural-nets-paul-humphreys. Accessed January 3, 2021.

Humphreys, Paul. n.d. "Epistemic Opacity and Epistemic Inaccessibility." Available at: https://wordpress.its.virginia.edu/Paul_Humphreys_Home_Page/files/2016/02/epistemic-opacity-and-epistemic-inaccessibility.pdf. Accessed January 3, 2021.

Kieval, Phillip H. 2026. "Representation Learning Without Representationalism: A Non-Representational Account of Deep Learning Models in Scientific Practice." In *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*, edited by Juan M. Durán and Giorgia Pozzi, 245–72. Cham:Springer. https://doi.org/10.1007/978-3-032-03083-2_12.

Páez, Andrés. 2024. "Understanding with Toy Surrogate Models in Machine Learning." *Minds and Machines* 34:45. https://doi.org/10.1007/s11023-024-09700-1.

Pozzi, Giorgia. 2023. "Automated Opioid Risk Scores: A Case for Machine Learning-Induced Epistemic Injustice in Healthcare." *Ethics and Information Technology* 25:3. https://doi.org/10.1007/s10676-023-09676-z.

Räz, Tim, and Claus Beisbart. 2024. "The Importance of Understanding Deep Learning." *Erkenntnis* 89 (5):1823–40. https://doi.org/10.1007/s10670-022-00605-y.

Ribeiro, Marco T., Sameer Singh, and Carlos Guestrin. 2016. "'Why should I trust you?': Explaining the Predictions of any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. New York: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High-Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1:206–15. https://doi.org/10.1038/s42256-019-0048-x.

Schmidt, Eva, Paul M. Putora, and Rianne Fijten. 2025. "The Epistemic Cost of Opacity: How the Use of Artificial Intelligence Undermines the Knowledge of Medical Doctors in High-Stakes Contexts." *Philosophy & Technology* 38 (1):5. https://doi.org/10.1007/s13347-024-00834-9.

https://doi.org/10.1017/psa.2026.10246 Published online by Cambridge University Press

Sullivan, Emily. 2022. "Understanding from Machine Learning Models." *British Journal for the Philosophy of Science* 73 (1):109–33. https://doi.org/10.1093/bjps/axz035.

Vandenburgh, Jonathan. 2023. "Machine Learning and Knowledge: Why Robustness Matters." *Preprint*, arXiv:2310.19819. https://doi.org/10.48550/arXiv.2310.19819.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2018. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR." *Harvard Journal of Law and Technology* 31 (2):841–87. https://doi.org/10.2139/ssrn.3063289.

Watson, David S., and Luciano Floridi. 2021. "The Explanation Game: A Formal Framework for Interpretable Machine Learning.: *Synthese* 198 (10):9211–42. https://doi.org/10.1007/s11229-020-02629-9.

Zednik, Carlos. 2021. "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence." *Philosophy & Technology* 34(2):265–88. https://doi.org/10.1007/s13347-019-00382-7.

Zednik, Carlos, and Philippe Verreault-Julien. 2025. "Can XAI Justify?" In *Philosophy of Science for Machine Learning: Core Issues and New Perspectives* edited by Juan M. Durán and Giorgia Pozzi, 107–29. Cham: Springer. https://doi.org/10.1007/978-3-032-03083-2_6.

Zerilli, John. 2022. "Explaining Machine Learning Decisions." *Philosophy of Science* 89 (1):1–19. https://doi.org/10.1017/psa.2021.13.

Cite this article: Durán, Juan Manuel. 2026. "Against epistemic transparency of algorithms." *Philosophy of Science*. https://doi.org/10.1017/psa.2026.10246

https://doi.org/10.1017/psa.2026.10246 Published online by Cambridge University Press