Epistemic Opacity and Scientific Realism and Anti-Realism

Jack Casey, University of Cambridge, jc2605@cam.ac.uk

Abstract

Objections to computer-assisted proofs are remarkably similar to those expressed with regards to machine-learning models; both usually being centered on the epistemic opacity of either methodology. In this chapter, I explore how these objections differ, arguing that their differences indicate that epistemic opacity is a matter of degree. I then argue that one's level of comfort with a given degree of epistemic opacity will be, in large part, determined by one's position with regards to scientific realism. If one is a realist, I argue, one's tolerance for epistemic opacity will be lower. This brings to light the bearing of the well-established debate between scientific realism and anti-realism into contemporary discussions over the acceptability of epistemic opacity in machine-learning.

Keywords: Epistemic Opacity, Machine-learning models, Computer-assisted proofs, Scientific realism, Scientific anti-realism.

I. Introduction

The first computer assisted proof of a theorem was published by Appel and Haken in 1979. Their proof of the four-color theorem initiated a revolution in mathematics, and the use of computer-assisted methods of proof is now commonplace in contemporary mathematics. The proof initiated a philosophical controversy that still rages today, centered mainly on the question of whether computer assisted proofs should be regarded as having the same status as conventional proofs in mathematics. More recently, superficially similar epistemic concerns have been expressed with regards to knowledge gained through the modelling of natural systems by machine learning algorithms. For that reason, then, it's common to see a discussion of the epistemic worries generated by computer-assisted proofs feature alongside a discussion of the epistemic worries generated by the employment of computational resources in natural science – even if this is often accompanied by tacit acknowledgement that we might be dealing with very different concerns (Humphreys, 2009: 617; Durán and Formanek, 2018: 649).

Arguments against the equal acceptance of computer-assisted proofs in mathematics often center on the apparent *unsurveyability* of computer-assisted proofs. Roughly, a proof is surveyable if its possible (even if only in principle) for a rational agent to understand the proof in its entirety, and thereby come to *know* the conclusion. Tymoczko names surveyability as one of three necessary characteristics of proofs (1979: 59) and traces the unsurveyability of (some) computer-assisted proofs as being the crucial difference between computer-assisted, and traditional proofs (1979: 73). It's for this reason, primarily, that computer-assisted proofs are usually regarded with either suspicion, or as having a lesser status in comparison with traditional proofs (Brown, 2010; Tymoczko, 1979; Kitcher, 1983; also see McEvoy, 2008; Parshina, 2023).

Similarly, in the case of machine learning models, epistemic concerns center on the *black box* nature of some machine learning algorithms (see Durán and Jongsma, 2021). The common complaint with regards to the lack of opacity in machine learning contexts amounts to a complaint that machine learning algorithms are also unsurveyable – being either too complex, or too large, for a single (or even multiple) person(s) to work through the entirety of their architecture.

My contention in this chapter is twofold; firstly, that the epistemic concerns around computer-assisted proofs in mathematics, and the employment of machine learning algorithms (MLAs) in natural science both stem from the fact that both processes are *epistemically opaque*. This is not a novel idea (Humphreys, 2009). What has not been considered, however, is how those concerns are importantly dissimilar, and what this tells us about epistemic opacity. In this chapter, I will argue that, though indeed both stem from concerns over epistemic opacity, the different conclusions that are drawn on the basis of their being epistemically opaque indicates that opacity is a matter of degree.

Secondly, after arguing that epistemic opacity is a matter of degree, I then argue that one's acceptable degree of epistemic opacity will be (at least partly) determined by one's commitment to either scientific realism or scientific anti-realism. Realists, I argue, will have a low tolerance for epistemic opacity. Anti-realists, on the other hand, I argue will be much more comfortable with epistemically opaque processes.

Dialectically, this is notable. It's natural to think that the arguments against the implementation of computer-assisted proofs on epistemic grounds might similarly serve those positioned against the employment of machine learning models in natural science. This chapter then will go some way to explaining why those potential lines of criticism are not translatable into the latter context. Furthermore, it opens up room for the reassessment of the scientific realism/anti-realism debate in a novel context.

II. Computer-assisted Proofs

The first computer-assisted proof of a theorem was the proof of the four-color theorem (4CT) in 1977 by Appel and Haken. A small literature bloomed in the wake of Tymoczko's paper concerning the philosophical implications of the proof (1979).[1]

The four-color problem is as follows: is it possible, using only four colors, to color any given map such that no two adjacent areas of the map are the same color? The problem was first posed by Guthrie to De Morgan in 1852 (Parshina, 2023: 108; MacKenzie, 2004: 103). It stood unsolved for more than 100 years – with a brief period in between in which the theorem was believed to be proved, before a mistake was recognized -- before Appel and Haken published their computer-assisted proof in 1977.[2]

The proof begins by translating the problem from one concerning maps to one concerning *graphs*. Regions on a map become *vertices* – points on the graph. That a border exists between two regions are indicated by *edges* – lines joining one vertex to another. The number of edges meeting at any vertex is called the *vertex degree.* Specifically, we're considering *planar graphs* – graphs on which edges cannot intersect, other than at vertices. In essence, the proof consists of reducing all possible combinations of edges and vertices to a finite set of configurations and demonstrating that all possible configurations are such that all vertices can be colored differently from any other vertices to which they are joined by an edge, using only 4 distinct colors in total.

Using Euler's formula, it's demonstrable that the minimal counterexample (i.e., the most difficult instance to prove 4-colorable) to 4CT would be found after a graph undergoes *triangulation* (Tymoczko, 1979: 64). Any region of the graph bounded by four or more edges will have at least two non-adjacent vertices on the boundary. We can join these vertices with a new edge, not intersecting any other edges. Doing this until all we are left with is three-sided regions *triangulates* the graph. Doing so only makes

[1] See Parshina (2023) for an overview of the debate.

[2] I wish to avoid the technical details of the proof here as much as possible and will include details only to the extent that they are philosophically relevant for the discussion at hand. Those interested should consult Haken and Appel's original proof (1977), and Tymoczko's overview (1979: 63-69). I here take my description of the proof from Tymoczko's overview.

it more difficult to color the graph using only 4 colors, since it only further restricts possible colorings. If 4CT holds for triangulated graphs, it holds generally. The strategy then is to identify the minimal counterexample, assume that it is not four-colorable, and prove that it is (thereby proving that any other configuration is also four-colorable). It's trivial to demonstrate that vertices of degree 3 are four-colorable – given three edges, it's clearly possible to color each vertex a different color. It's also not too difficult to prove the same for vertices of degree 4 using the *Kempe chain argument*. The proof for 5 is much more difficult.

The computer-assisted portion of Haken and Appel's proof appears here, then. Following work by Birkhoff (1913), they first identified a finite set of configurations of vertices and edges and demonstrated that any planar graph *must* contain one of these configurations, calling this set of configurations *unavoidable.* Secondly, they then demonstrated that these configurations were *reducible* – a configuration being reducible if the 4-coloring of any graph containing that graph as a subgraph is entailed by the 4-colorability of any graph with fewer vertices. The route for providing a proof is then clear: if it can be demonstrated that presence of these subgraphs indicates 4-colorability, and it can also be demonstrated all graphs contain at least one of these configurations as subgraphs, it indicates that no counterexample to 4CT is possible.

Generating the *unavoidable* set of possible configurations was assisted by computer. Though the set is large – there were 1834 configurations in the unavoidable set in 1977 – after generating the set, there are not so many configurations as for it to be unsurveyable.[3] Assessing the reducibility of the set, however, *is* too complex a task for a single human being. To check the reducibility of one configuration alone, approximately 500,000 logical operations are performed (Parshina, 2023: 109). As a result, Appel and Haken had a computer fulfil the task (using a methodology called *discharging*), with the computations taking over 1200 hours (Tymoczko, 1979: 68).

The resultant proof then is well understood from a theoretical standpoint. As Tymoczko says, the proof is far from being a simple brute force proof – the generation of the unavoidable set, and the method for testing them, rests upon 'novel and sophisticated theory developed by the authors' (Tymoczko, 1979: 68). Nonetheless, it's far beyond a human being's capabilities to assess each configuration and see whether it is indeed reducible; and it's the fact that it cannot undergo such an assessment that generated skepticism with regards to its status as a *proof,* in the traditional sense.

I'll here focus primarily on those objections that stem from the *unsurveyability* of 4CT (and, by extension, the objection to the classification of all computer-assisted unsurveyable proofs as proofs simpliciter).[4] The first to focus on the unsurveyability of 4CT as being the reason it stands apart from conventional proofs was Tymoczko (1979). Tymoczko's argument is that one major characteristic of proofs is that are *surveyable*. That is, proofs must be 'comprehended by mathematicians' to count as proofs (Tymoczko, 1979: 59). He very firmly places a requirement on the possibility that a proof can be 'looked over, reviewed, verified by a rational agent', in order to count as being a proof in the traditional sense (Tymoczko, 1979: 59). Tymoczko credits surveyability as being the source of the *aprioricity* of mathematical theorems. In the case of Appel and Haken's proof, there is very clearly a step within the proof to which we, as rational agents, are necessarily ignorant.

There is intuitive pull to the idea that there's very clearly a difference between a proof in which part of the proof is computer-assisted, and one on which we are fully acquainted with the chain of reasoning. Tymoczko argues that computer-assisted proofs differ in so far as the justification for our believing the computer-assisted step is necessarily grounded in empirical knowledge. For example, the justificatory

[3] Though the set has subsequently been reduced significantly, first by Appel and Haken themselves, and then by Robertson, et al. (1996), the computational task of assessing reducibility still stands way beyond the capabilities of a single human being.
[4] Other forms of objection have been made. See Parshina, 2023 for a classification.

power of the step relies on our understanding of the physics that describes the way computers function, for example. We only trust this step because of our knowledge of physics that undergirds our understanding of how computers function generally, and what that computer is doing in this specific case. All this knowledge, Tymoczko notes, is necessarily *a posteriori*. On this basis, Tymoczko argues that Appel and Haken's proof should not be considered as standing alongside traditional proofs, which are typically justified through *apriori* means alone.[5] Nonetheless, Tymoczko takes this development in stride; he suggests that, though 4CT might require us to reevaluate the status of mathematics as a purely *apriori* endeavor, this is not necessarily a concerning development (Tymoczko, 1979: 80). Nonetheless, he maintains that the fact we are ignorant of a step in the proof process of 4CT suggests that this proof is of an importantly different kind, and that it is of a different kind is directly attributed to the necessity of our ignorance of part of the justificatory process, and our reliance on empirical knowledge to serve as justification for our belief that this part does indeed stand in the relevant relations of logical inference, even if we can't directly see that it does so.

Brown (2010) shares similar concerns. He bases his skepticism over the status of Appel and Haken's proof in the *fallibility* of computers. Similarly to Tymoczko, his contention is that the source of fallibility in computer-assisted proofs is found in empirical science. We have hypotheses over how computers function, and those hypotheses may be wrong (even if we consider that possibility remote). The portions of unsurveyable proofs whose justifications stand on our understanding of how computers function inherit this uncertainty. While we might be very confident of our understanding of how computers function, we can never get to the certainty inherent in *apriori* justifications; our confidence in the reliability of the unsurveyable portions of proofs can never extend beyond that afforded to well-confirmed scientific theories, which undergird our understanding of how computers function. The fact that there is *some* possibility that we are mistaken, means the status of 4CT as *proof simpliciter* cannot be maintained, according to Brown. Brown differs from Tymoczko, in that he is not as sanguine towards the prospect of the quasi-empirical turn in mathematics apparently demanded by the development of computer-assisted proofs.

III. Machine Learning Algorithms, and Epistemic Opacity

Superficially similar epistemic concerns have been raised with regards to machine learning algorithms. Before we begin, I'll give a general overview of how machine learning algorithms function; technical details are omitted – the level of granularity required to see the philosophical implications for our purposes is not high. Typically, machine learning algorithms function by first ingesting training data, to which they attenuate their internal model – adjusting their internal parameters according to that data. Dependent on the degree to which they are predictively successful given recognition of relevant features – with their success being judged externally -- they further adjust their internal parameters, either strengthening or weakening the weight they place on particular features as indicators for whatever predictive output they gave, based on whether or not that particular weighting led to a successful prediction or not, respectively. Through this process, they are able to 'learn' autonomously, in so far as they can come to recognize relevant features of input data on which to make predictions, through the weightings of their internal parameters, without those parameters being set by an external agent.

Machine learning algorithms have now been deployed in numerous areas of empirical enquiry and are increasingly complex. Given sufficient complexity, it becomes impossible to follow the decision procedures internal to them. That is, it is impossible to tell *why* they made the prediction they did. Algorithms with sufficient complexity to preclude understanding of them are often described as being

[5] Some have pushed back against this claim, notably Burge (1998).

black-boxes.[6] To use one oft-cited example to illustrate this issue, consider the example of the employment of a machine learning algorithm to detect skin cancers (Esteva, et al. 2017). The algorithm is trained on labelled dataset consisting of 129,540 images of skin lesions. After this training, it is then tasked with classifying novel images according to the type of skin lesion found in the image. Whilst the study suggests the model outperformed physicians involved in terms of correct classification, no justification for *why* it gave the classification that it did is available. There are saliency maps, but they offer nothing greater than an understanding of which pixel in the image the model took to be important; we lack any ability to understand *why* that pixel was important. As Durán and Jongsma note, there is intuitive pull to the idea that it's objectionable for us to be reliant on a tool for cancer diagnosis, when we are unaware of *how* the diagnosis was made, in so far as the models complexity means we're unable to identify which feature the model identified as being the reason for a particular diagnosis (Durán and Jongsma, 2021: 330). It is often on the basis that MLAs are increasingly used in highly impactful contexts such as these that objections to their inherent complexity are often raised.

Objections to the use of MLAs in empirical science, then, usually rely upon this lack of access to the internal workings of black-box algorithms. As the thought usually goes, in so far as the model is too complex to comprehend, we lack some corollary of a chain of reasoning on which to understand the basis on which the prediction was made, and as a result we should be suspicious of that prediction. If a model picks out a particular pixel as salient, we want to know *why* that pixel is salient, but the complexity of many models precludes us attaining such an understanding.[7]

This class of objections, then, look remarkably similar to those expressed with regards to computer-assisted proofs. We can understand how the two are related by casting those concerns in terms of Humphreys' notion of *epistemic opacity*:

> [A] process is epistemically opaque relative to a cognitive agent X at time t just in case X does not know at t all of the epistemically relevant elements of the process. (Humphreys, 2009: 618)

In the case of 4CT, there is a step within the proof which is epistemically opaque. That is, in affirming the *reducibility* of configurations in the unavoidable set, we are reliant on a process to which we're not directly epistemically acquainted. The computer-assisted part of the proof is *epistemically opaque*. Similarly, in the case of machine learning models, the reason for discomfort in their inclusion in a clinical setting is that there is an epistemically relevant part of the process (namely, the inference made from an image to a diagnosis of cancer) to which we are necessarily ignorant given the complexity of the model. All this is to say that, in both cases, the source of the epistemic concerns that have been registered seem plausibly to stem from the epistemic opacity present in both cases. Furthermore, that opacity seems particularly stubborn; in both cases, the opacity stems from the limitations of our cognitive resources as human beings.[8]

---

[6] From this point on, I'll refer to black-box MLAs as simply 'MLAs'.

[7] Much work has gone into the task of making MLAs *explainable,* or *interpretable*, where a system is interpretable if its 'operations can be understood by a human, either through introspection or through a produced explanation' (Biran and Cotton, 2017: 8; also see Beisbart and Räz, 2021: 1). Miller goes some way in demonstrating the amount of work required at the level of the conceptual foundations of the field, in so far as the notion of *explainability* requires much greater clarification (Miller, 2018). Durán and Jongsma argue more forcefully that XAI simply removes the problem by one step (2021: 330).

[8] In contrast with more mundane forms of epistemic opacity, such as when a proprietary model is used, precluding us from gaining an understanding of its inner workings (see Burrell, 2016).

Much of this has been said before. What has not been discussed is the difference between the two cases. Here I wish to draw attention to the difference in the conclusions that are usually drawn given epistemic opacity in either case. In the case of computer-assisted proofs, on the whole, the consensus is *not* that, given epistemic opacity at some stage in a computer-assisted proof, the proof should be rejected wholesale. Rather, the more limited recommendation is usually made – as we saw with Tymoczko -- that computer-assisted proofs should have a lesser status than conventional proofs, or that mathematics reliant on computer-assisted proofs must be recognized as quasi-empirical (Lam, 1990). That is, the truth of the result is not really at issue in computer-assisted proofs.[9] Commentators often go to great lengths to express that their incredulity stems from the mere *possibility* that computer-assisted proofs *could* be mistaken, not that they think that they actually are. As we saw, most arguments against computer-assisted proofs are often predicated on the fact that relying on computer-assisted proofs means we're then partially reliant on those physical theories that undergird our understanding of how those computers function, which introduces an unacceptable level of jeopardy, but only by virtue of the fact there is a possibility our best physical theories might turn out to be false. Those arguments only run when one holds that *no* jeopardy should be (or is) the natural state of mathematical proofs.

A much stronger conclusion is usually drawn in the case of MLAs in empirical science; in many cases, skepticism regarding the truth of the result is expressed. Essentially: we should not trust these processes because we don't have an understanding of how they work, and as a result they could *actually be wrong*. The sense of possibility here is a different one. Rather than some vanishingly small chance that our best physical theories turn out to be mistaken, objectors believe that some erroneous correlation might actually be made, and our inability to survey those models means we cannot detect it. In the case of machine learning algorithms, a far stronger recommendation is usually made. In so far as they're employed to recognize patterns in the world, skepticism arises as to whether a pattern has indeed been properly recognized. The epistemic issue is much more immediate. Indeed, the objection that the employment of machine learning algorithms makes the knowledge derived from them *a posteriori,* given epistemic opacity, has no force in this context – all knowledge gained via *any* empirical means is already *a posteriori* to begin with.

What does this indicate, then? First of all, that a process is epistemically opaque does not entail skepticism of the truth of claim that is justified on the basis of that opaque process. If it did, then we would see far greater skepticism with regards to the truth of 4CT, rather than mere skepticism with regards to its status as a proof. Moreover, that this stronger form of skepticism *is* expressed with regards to MLAs indicates that sufficient epistemic opacity *can* entail skepticism with regards to the truth of prediction.

One way to account for this difference is that epistemic opacity comes in degrees.[10] In case of the computer-assisted proofs such as 4CT, the epistemic opacity is fairly low. We have a good theoretical understanding of how the proof functions, it's just that we can't sort through the vast number of configurations ourselves. Nonetheless, the computer-assisted portion of the proof is epistemically opaque *enough* that it warrants a downgrade in the proof's status, given the high threshold established by the fact mathematics is usually taken to be a purely *apriori* discipline. In the case of machine learning algorithms, the epistemic opacity is often high enough to warrant suspicion of the truth of the prediction – we lack a clear theoretical basis on which the function of the algorithm stands. It's no surprise that

[9] Note here, if 4CT is indeed true, for Tymoczko, it is a *necessary a posteriori* truth. The possibility for error only allows for 4CT to be *necessarily false,* or, alternatively, for our apparent justification of 4CT to be misguided – that is, 4CT might be true, but the justification for believing it to be true might be missing, for example if there was an undetected error in the unsurveyable portion of the proof (Tymoczko, 1979: 77, also see Kripke (1980)).

[10] San Pedro (2020) also argues that epistemic opacity is a matter of degree. Durán and Formanek (2018) also talk of 'degrees of opacity'.

recommendations for reducing opacity often proceed along lines suggesting that we make them coherent with existing theoretical foundations (see Rudin, 2019). That epistemic opacity is a matter of degree is in line with other suggestions, elsewhere, too. Sullivan, for example, argues that fully-fledged transparency in a model isn't a desirable goal. Some details don't matter (2022: 1). In so far as *fully-fledged* transparency isn't a desirable goal, though *some* is, this implies that epistemic opacity is not an absolute matter.

Indeed, we can consider what would be the case if we raised the degree of epistemic opacity in the case of computer-assisted proofs to something closer to the degree of epistemic opacity we see in machine learning algorithms – Tymoczko consider just such a case, and perhaps unsurprisingly, the skepticism of the result would be much more severe (Tymoczko, 1979: 72). The case he considers is one concerning a genius mathematician, Simon, whose ability fair outstrips his peers. Simon begins by proving many new theorems by traditional methods. Eventually, given his ability, his proofs become too long to comprehend. He begins to handwave to portions of proof that are too extensive with the simple phrase 'Simon says'. Tymoczko's claim is that the logic of 'Simon says' is the same as the appeal to computer in a computer-assisted proof; both being, in essence, appeals to authority (Tymoczko, 1979: 72). Nonetheless, there's an intuitive pull to the idea that computer-assisted proofs are on a far safer epistemic footing than appeals to authority such as 'Simon says'. Viewing this issue through the lens of epistemic opacity, though both are indeed formally similar, we can account for this difference in so far as there is much greater degree of epistemic opacity in Simon's case. And it is because of this greater degree of opacity that we are rightly more skeptical of cases like Simon. Indeed, even if it turned out that Simon was universally correct, and the 4CT proof turned out to be false, we would still suggest that we were right to be more suspicious of Simon.

IV. Epistemic Opacity, Realism and Anti-Realism

What this tells us then is that epistemic opacity is a matter of degree. Processes of knowledge production can be epistemically opaque to some particular degree. Given greater opacity in a process, a greater degree of skepticism in the truth of the claim made on the basis of that process is warranted.

Two questions naturally arise, then: how do we gauge the level of epistemic opacity in a process, and what determines what level of epistemic opacity is acceptable? My suspicion is that both have highly complex answers. The acceptable level of opacity will probably be determined by a variety of context-dependent factors. For one, it seems likely that it will partly be determined by what one wants to do with the knowledge. If we're considering a low-stakes situation, the threshold might be high. If I'm using the algorithm to determine whether or not a person has cancer, the threshold might be much lower. Relatedly, the availability of extant, alternative explanations: if we're considering phenomena whose dynamics we are largely ignorant of, then perhaps a higher degree of opacity will be palatable. If we lack a theoretical basis for how a particular disease functions, for example, but we have an opaque model which is apparently predictively successful, we are more likely to be comfortable with its employment, even given a degree of epistemic opacity we might otherwise deem intolerable. In relation to computer-assisted proofs: though a computer-assisted proof might not be palatable as a proof itself, its existence might be enough to encourage mathematicians that the aim of finding a conventional proof for the same theorem is worth dedicating their time to. Similarly, measuring opacity seems possible – my argument partly rests on the idea that we can at least recognize relative differences in cases.[11]

For now, I want to sidestep both of these rather thorny issues. These foregoing considerations seem to be largely instrumental in nature. I will here focus on whether there any *intrinsic* reasons that affect the acceptable degree of epistemic opacity. Primarily, what I want to consider is whether there are any

[11] San Pedro (2020) considers the relation between epistemic opacity and actual scientific modelling and simulation practices.

background philosophical positions that will act to generally determine one's acceptable level of epistemic opacity. One such belief which I take to act as a significant determinant of one's general acceptance of epistemic opacity is one's position with regards to the debate between scientific realism and anti-realism.

If one's position with regards to scientific realism and anti-realism are as impactful as I will argue, this is useful to consider as i) given that these are fundamental, background philosophical beliefs, these will apply generally, across most, if not all, contexts mentioned. Given that epistemic opacity appears likely to be highly context-dependent, this might give us some firmer footing on which to tackle this much more complicated question, and ii) dependent on which one of these views is correct, this might help us establish general normative standards with regards to what *should* be the acceptable of epistemic opacity. That is, if it should turn out that subsequent argumentation strongly favors one particular position over the other, then it might be that we *should* be more lenient (or stringent) *generally*, with regards to what we consider to be acceptable levels of epistemic opacity.

To begin, scientific realism and anti-realism are usually framed as positions on the correct epistemic attitude to adopt with regards to scientific theories (Kukla, 1998: ch. 1; Psillos, 1999; Chakravartty, 2011: 1.2). Realism is typically marked by a positive attitude to the existence of entities whose existence is entailed by scientific theories. The reason for this positive attitude is usually ascribed to the realist commitment to the *literal* or *approximate truth* of scientific theories. For example, if our best physical theories talk of *electrons*, the realist suggests we should believe that electrons exist, in just the same sense in which we readily commit to the existence of non-theoretical objects we're directly acquainted with (tables, chairs, etc.). The primary argument for the position, termed the *no-miracles argument,* is that the best explanation for the predictive success of our best scientific theories is that what they say is *true* – not ascribing truth to them makes that predictive success *miraculous* (Putnam, 1975: 73). In so far as they are true, what our best theories refer to must exist for them to be true.[12]

Anti-realism, on the other hand, is usually marked by either a skeptical, or, alternatively, a dismissive attitude to questions regarding the truth of our best scientific theories. Taking their cue from the so-called *pessimistic meta-induction*, or the recognized historical tendency for our best scientific theories to be supplanted by novel, incommensurable theories over time, anti-realists recommend that we take a more conservative attitude to question of the truth of our best scientific theories.[13] Having been proven wrong in the past, anti-realists are skeptical of the claim that predictive success can underwrite a belief that those theories are true. To explain what differentiates our best scientific theories from unsuccessful ones, they typically point to their predictive success alone, or to the coherence of a theory with observable statements to which they are more comfortable committing to the truth of. Van Fraassen's view of science as 'a jungle red in tooth and claw' is often appealed to (Van Fraassen, 1980: 40); the success of our best theories is not miraculous, they're just the only ones that have managed to survive (see Wray, 2010).

Though strictly speaking realism is usually cast, as mentioned, as a position with regards to the correct epistemic attitude to our best scientific theories, there are underlying motivations for the position that find their expression in that explicit claim. Laudan, for example, describes realism as follows:

> At its core, realism is a normative doctrine about what the aims or values of science ought to be. Specifically, the realist maintains that the goal of science is to find ever truer theories about the natural world. But the modern-day realist typically conjoins to this axiological thesis a

[12] I omit some detail with regards to exact formulations of realism here, simply because it has no impact on the argument. Some realists, for example, commit to the *semantic thesis* of realism (namely, that our best theories are literally (approximately) true, whilst reserving from committing to the mind-independent existence of that which they describe (see Chakravartty, 2011, 1.2; 2007: ch. 6).
[13] See Wray (2015) for an overview of the various formulations of the argument.

> descriptive one: to wit, that the history of science, especially in recent times, can best be understood as an exemplification of the programmatic ideas of realism… [A]lthough the normative claim of the realist is logically independent of the descriptive claim, the former is epistemically parasitic on the latter. If it should turn out that the descriptive claim is false… then serious doubts will be raised about its normative counterpart. (Laudan, 106).

According to Laudan's characterization, the realist motivation is deeper than one merely concerning the correct epistemic attitude to our best scientific theories – rather, it is a position with regards to the aims of science itself. The aim of science *ought* to be the identification of truths concerning the natural world. That is, realists take it to be an aim of science to accurately describe reality. This motivation then finds its expression in the commitment to the literal truth of theories. As such, they look likely to be uncomfortable with the utilization of more epistemically opaque methods in scientific enquiry. If, as the realist suggests, a core aim of science is descriptive accuracy, mere predictive success is not enough. In so far as the realist is committed to a conception of science as having a core aim of furnishing us with truth, methodologies which offer mere predictive success, and not an interpretable, accurate description will fail to meet this standard. All this is to say, those of a realist bent who subscribe to a view of science as having as a core aim an accurate description of the world will likely be uncomfortable with epistemic processes with high epistemic opacity, for those processes offer no such descriptions. Understood as a normative claim, then, realism will hold that epistemically opaque processes, however predictively successful, are in violation of the dictum that science *should* furnish us with an accurate description of reality. Therefore, their tolerance for epistemic opacity will necessarily be lower.

This conception of realism as concerning the *aims* of science is not universally shared, however (Chakravarty, 2011: 1.1). And as Lauden notes, the normative claim is logically independent of the descriptive one. Some realists will not frame their position in terms of the *aims* of science and will restrict themselves to a view of realism as merely requiring a commitment to the truth of our best scientific theories. Doing so, perhaps they might find epistemically opaque processes palatable in the following way: we can transpose that position into the current context, substituting interpretable theories of natural science to which they would commit to the truth of for the epistemically opaque models internal to black-box MLAs. It seems possible for the realist to maintain that the only reasonable explanation for the predictive success of those models is that they accurately represent the target phenomena. As such, they might be comfortable committing to the accuracy of those models, even when the model itself is epistemically opaque.

Whether or not the realist *can* maintain their position coherently, however, is not the question. This requirement to commit to the *truth* of epistemically opaque models is quite onerous. Commitment to realism entails a commitment to the truth of models which necessarily lay outside of our epistemic reach. It's one thing to commit to the truth of understandable theories, which sit embedded in a wider framework of theoretical understanding – with our trust in those theories being at least partially dependent on them cohering with other, empirically verified theories. It's quite another to be asked to commit to the truth of something to which we are not acquainted, purely on the basis of predictive success. As such, though the realist *can* coherently commit to the truth of epistemically opaque models, they might not want to. That is, given that they commit to the (approximate) truth of predictively successful models, such a high commitment might mean their threshold for epistemic opacity is lower. If we are to commit to the truth of MLA models, then realists might prudently reserve this commitment to processes with less epistemic opacity.

If one considers this situation from an anti-realist perspective, the difference is stark. While MLAs might stand beyond our ken, if the sole metric by which scientific theories are to be judged is their predictive success as the anti-realist suggests, then nothing is standing in the way of extending the status of 'science' – whatever that might amount to – to models whose inner workings are beyond our cognitive capabilities. There seems to be no principled reason to require that theories be interpretable

to us; indeed, the position of the anti-realist is that we neither commit to the truth nor the falsity of our best scientific theories. As such, that a model has a high degree of epistemic opacity proves to be no issue. Take, for instance, Van Fraassen's *constructive empiricism*. As long as predictions of the model coheres with observation of observables (Van Fraassen, 1980), where *observables* refers to those entities we can observe unmediated by the use of theory – chairs and tables being observable, electrons (in so far as our observation of them requires the use of technology, with the understanding of how that technology functions being embedded in theoretical knowledge) being unobservable – then the model is *empirically adequate* (in comparison with *true*, according to the realist). The epistemic opacity of the model is simply of no concern – if it works, it works. In so far as anti-realism makes no demand that we commit to the truth of the models, our threshold for epistemic opacity should be much higher – if the model turns out to be false – which, given higher degrees of epistemic opacity, is more likely --, that's fine. We were never committed to its truth in the first place.

That being said, though the situation described seems to fit nicely with the underlying motivations of the anti-realism position overall, the particularities of actual expressions of anti-realism might preclude them being entirely comfortable with highly epistemically opaque processes. For example, Van Fraassen's constructive empiricism ascribes empirical adequacy to *theories*. MLAs do not produce theories to which we can ascribe empirical adequacy. This, however, is perhaps unsurprising, given that these positions were developed in a context in which predictive success only ever emerged as a result of the employment of theories. There seems no principled reason empirical adequacy cannot be extended to MLA developed models, as well. If not true to the letter, it appears true to the *spirit* of anti-realism.

It seems likely, then, that one's comfort or discomfort with epistemic opacity will be, in large part, determined by one's position with regards to scientific realism and anti-realism. In so far as realism demands us to commit to the truth of predictively successful models, a lower threshold for epistemic opacity is likely. In so far as anti-realism ascribes the lesser status of empirical adequacy to models, just in case the predictions of those models agree with facts concerning observables, there's more scope for epistemically risky models, and therefore a higher tolerance for epistemic opacity.

Though these are technical positions, they concern very general considerations that most will have opinions on (even if such opinions have not been reflected upon). They concern a basic conception of what science is, and what its aims are. That is, if one considers science as being in the business of increasing understanding, rather than merely being a methodology for the construction of technology, one is more likely to be aligned with the realism. Those of a more pragmatic stripe, who are dismissive of the notion that any deep understanding of nature is to be gleaned using science – typified, perhaps, by Mermin's famous instruction for to simply 'shut up and calculate', concerning discussion regarding the correct interpretation quantum mechanics (Mermin, 1989: 9)[14] -- are likely to feel more at home in the anti-realist camp. As such, though they might not cast them in terms of realism and anti-realism, most will have preformed opinions that will act as general determinants of the acceptability of epistemic opacity. Philosophical clarification of this issue – ignoring the inherent unlikelihood of such an event – might then act to give general normative guidelines with regards to epistemic opacity. Should argumentation strongly favor anti-realism, for example, then it might be that we have good reason to dismiss our concerns regarding the epistemic opacity of MLAs; don't worry that we don't have any understanding of how they work, that worry is simply a hangover of realism – all that matters is they work. This underlying tension between realism and anti-realism can be seen as expressed in other disagreements. For example, over what models should do; as to whether they should be assessed solely according to their purpose (Parker, 2020), along anti-realist lines, or whether there is a more objective feature of correspondence between model and phenomenon (underlying something like 'link uncertainty', for Sullivan (2022a)), as the realist would have it. The recent focus on *understanding*, and

[14] Often misattributed to Feynman (Mermin, 2004).

the question of whether or not models furnish us with understanding, can too be seen as being driven by underlying realist commitments (see Sullivan, 2022b; de Regt, 2017; Strevens, 2013). Understanding might not even be a desideratum for anti-realists, outside of the instrumental benefits it offers. Up to now, that this may be a purely partisan issue hasn't really been recognized; the presumption that models *should* provide understanding in the first place looks to be fundamentally grounded in unspoken realist commitments.

V. Concluding remarks

There remain questions about the coherency of a realist demand for lower epistemic opacity. While it does indeed seem prudent to reduce epistemic opacity, given one's commitment to the literal truth of a model, if the primary argument for taking theories to be literally true in the first place is the predictive success of the theory, it seems difficult to explain how one might coherently demand lower epistemic opacity. Either predictive success is indicative of truth, or it isn't. As such, this problem seems like it might require a restatement of some realist positions. That, or they extend the umbrella of the no-miracles argument to epistemically opaque models, in which case, the position seems to be at odds with its underlying normative character.

This is an interesting development, in so far as it promises to reinvigorate the debate between realism and anti-realism, transposing the dispute from one concerning the correct epistemic attitude to theories, to a much wider dispute, one which appears much closer to the heart of their motivating principles – whether we should conceive as science as being in the business of furnishing us with truths about our world, or have a much more limited aim, in being predictively successful. Of course, from the perspective of the anti-realist in this dispute, this more 'limited aim' is simply an entailment of the increased predictive power technology offers us, which simultaneously demands our displacement from the center of epistemic considerations.

References

Appel, Kenneth, and Haken, Wolfgang (1977). Solution of the four color map problem. *Scientific American*, vol. 237, no. 4, pp. 108–121.

Beisbart, Claus and Räz, Tim. (2021). Philosophy of Science at sea: Clarifying the interpretability of machine learning. *Philosophy Compass.* 1-11.

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In IJCAI-17 workshop on explainable AI (XAI) (Vol. 8, pp. 8–13).

Birkhoff, George. (1913), The reducibility of maps. *American Journal of Mathematics*. 114-128.

Brown, James Robert. (1999). *Philosophy of Mathematics: An Introduction to a World of Proofs and Pictures*. New York: Routledge.

Burge, Tyler. (1998). Computer proof, apriori knowledge, and other minds: The sixth philosophical perspectives lecture. *Philosophical Perspectives*. Vol. 12. *Language, Mind, and Ontology*, pp. 1-37

Burrell, Jenna. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society.* 3. (1):205395171562251.

Chakravartty, Anjan. (2017). Scientific realism. *The Stanford Encyclopedia of Philosophy.* Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/>.

De Regt, Henk W. (2017). *Understanding Scientific Understanding*. New York: Oxford University Press.

Durán, Juan M. & Formanek, Nico. (2018). Grounds for trust: essential epistemic opacity and computational reliabilism. *Minds and Machines*. 28 (4):645-666.

Durán, Juan M. & Jongsma, Karin R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics.* 47 (5).

Esteva, A., Kuprel, B., Novoa, R. *et al.* (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118. https://doi.org/10.1038/nature21056

Humphreys, Paul. (2009). The philosophical novelty of computer simulation methods. *Synthese.* 169 (3):615 - 626.

Kitcher, P. (1983). *The Nature of Mathematical Knowledge.* New York: Oxford University Press.

Kripke, Saul A. (1980). *Naming and Necessity: Lectures Given to the Princeton University Philosophy Colloquium*. Darragh Byrne & Max Kölbel (eds.). Cambridge, MA: Harvard University Press.

Kukla, André. (1998). *Studies in scientific realism*. New York: Oxford University Press.

Lam, C. W. H. (1990). How Reliable is a Computer-Based Proof. *The Mathematical Intelligencer* 12 (1): 8–12. https://doi.org/10.1007/bf03023977.

Laudan, Larry. (1984). *Science and Values: The Aims of Science and Their Role in Scientific Debate*. University of California Press.

MacKenzie, D. (1999). Slaying the kraken: the sociohistory of a mathematical proof. *Social Studies of Science*. 29(1), 7-60. https://doi.org/10.1177/030631299029001002

McEvoy, Mark. (2008). The epistemological status of computer-assisted proofs. *Philosophia Mathematica*. 16 (3):374-387.

Mermin, N. David. (1989). What's Wrong with this Pillow? *Physics Today*. 42 (4): 9.

Mermin, N. David. (2004). Could Feynman have said this? *Physics Today*. 57 (5): 10–11.

Miller, Tim. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence.* 267:1-38.

Parker, Wendy S. (2020). Model evaluation: an adequacy-for-purpose view. *Philosophy of Science.* 87 (3):457-477.

Parshina, Katia. (2023). Philosophical assumptions behind the rejection of computer-based proofs. *Kriterion – Journal of Philosophy.* 37 (2-4):105-122.

Psillos, Stathis. (1999). *Scientific realism: how science tracks truth*. New York: Routledge.

Putnam, Hilary. (1975). *Mathematics, Matter and Method*, Cambridge: Cambridge University Press.

Robertson, N., Sanders, D. P., Seymour, P., & Thomas, R. (1996). A new proof of the four-color theorem. *Electronic Research Announcements of the American Mathematical Society*, *2*(1), 17–25.

San Pedro. Iñaki. (2020). Degrees of Epistemic Opacity. *Manuscript.*

Strevens, Michael. (2013). “No Understanding Without Explanation.” Studies in History and Philosophy of Science Part A, 44(3): 510–515. https://doi.org/10.1016/j.shpsa.2012.12.005.

Sullivan, Emily. (2022a). Understanding from machine learning models. *The British Journal for the Philosophy of Science*, 73(1): 109–133. https://doi.org/10.1093/bjps/axz035.

Sullivan, Emily. (2022b). How values shape the machine learning opacity problem. In *Scientific Understanding and Representation* Eds. Lawler, Khalifa, Shech (pp. 306-322). Oxford: Routledge.

Tymoczko, T. (1979). The Four-Color Problem and its philosophical significance. *Journal of Philosophy.* 76. 57–82.

van Fraassen, Bas C. (1980). *The Scientific Image*. Oxford: Oxford University Press.

Wray, K. Brad. (2010). Selection and predictive success. *Erkenntnis*, 72(3): 365–377. doi:10.1007/s10670-009-9206-6

Wray, K. Brad. (2015). Pessimistic inductions: Four varieties. *International Studies in the Philosophy of Science*. 29(1): 61–73. doi:10.1080/02698595.2015.1071551