Philosophy & Technology (2020) 33:487–502
https://doi.org/10.1007/s13347-019-00372-9

RESEARCH ARTICLE

Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning

Maya Krishnan[1]

Received: 1 January 2019 / Accepted: 31 July 2019 / Published online 13 August 2019

Abstract
The usefulness of machine learning algorithms has led to their widespread adoption prior to the development of a conceptual framework for making sense of them. One common response to this situation is to say that machine learning suffers from a "black box problem." That is, machine learning algorithms are "opaque" to human users, failing to be "interpretable" or "explicable" in terms that would render categorization procedures "understandable." The purpose of this paper is to challenge the widespread agreement about the existence and importance of a black box problem. The first section argues that "interpretability" and cognates lack precise meanings when applied to algorithms. This makes the concepts difficult to use when trying to solve the problems that have motivated the call for interpretability (etc.). Furthermore, since there is no adequate account of the concepts themselves, it is not possible to assess whether particular technical features supply formal definitions of those concepts. The second section argues that there are ways of being a responsible user of these algorithms that do not require interpretability (etc.). In many cases in which a black box problem is cited, interpretability is a means to a further end such as justification or non-discrimination. Since addressing these problems need not involve something that looks like an "interpretation" (etc.) of an algorithm, the focus on interpretability artificially constrains the solution space by characterizing one possible solution as the problem itself. Where possible, discussion should be reformulated in terms of the ends of interpretability.

Keywords Machine learning · Interpretability · Explicability · Algorithms · Black box

1 Introduction

The usefulness of machine learning (ML) algorithms has led to their adoption prior to the development of a conceptual framework for making sense of them.

✉ Maya Krishnan
maya.krishnan@all-souls.ox.ac.uk

[1] All Souls College, High Street, Oxford, Oxfordshire OX1 4AL, UK

When people and organizations use tools that they do not understand, they risk failures that jeopardize lives, as well as more subtle and invidious failures such as promoting discriminatory outcomes. Growing awareness of such risks has made it commonplace to say that machine learning suffers from a "black box problem." The basic statement of the problem is this: Although classifiers built by new machine learning techniques such as "deep learning" are very successful at making predictions, we do not know how these classifiers actually work. The models underlying the classifiers are "opaque" to us. Both computer scientists and philosophers have addressed various facets of this dilemma, which has also been labeled a problem about intelligibility, explicability, transparency, or interpretability (Leese 2014; Mittelstadt et al. 2016; Burrell 2016).[1]

In this paper, I challenge the widespread agreement about the existence and importance of a black box problem. The first section argues that common belief in the presence of a black problem reflects conceptual unclarity and embeds debatable assumptions. "Interpretability," "explicability," and cognates lack precise meanings when applied to algorithms. Recently proposed definitions have not solved this problem. Moreover, it is not clear that obtaining information about the inner workings of algorithms will be useful to solving the problems cited above. The second section argues that, in many cases in which a black box problem is cited, what is called interpretability or explicability serves as a means to an end, rather than being an end in itself. Focus on interpretability artificially constrains problem-solving efforts by recasting one solution to a problem as the problem itself. Where alternative solutions could be available, it is preferable to frame discussion in terms of the underlying goal.

Although this paper criticizes the conceptual foundations of current research centered around interpretability or explicability, it does not thereby dispute the usefulness of the technical results themselves. It does not provide reason to doubt, for instance, the utility of the introduction of monotonicity constraints on input data or generated models (Hall et al. 2017), the use of local linear models to summarize decisions made by non-linear models (Ribeiro et al. 2016), or the introduction of techniques for visually representing relations in high-dimensional data sets in 2- or 3-dimensional space (Hall et al. 2017). Rather, it provides reason to doubt whether there is one single quality such as interpretability or explicability that unites these results in a significant way. Interpretability could still be a useful term for providing an approximate characterization of the way in which these projects seem broadly similar.[2] This is a loose and deflationary use of interpretability whose sense is supplied via rough analogy with uses of interpretation in different domains. But insofar as interpretability is understood as a substantial

[1] Many scholars have proposed fine-grained distinctions between these terms (e.g., "explicability" versus "interpretability," Mittelstadt et al. 2016, footnote 26) or define one term in relation to a similar term (e.g., defining transparency in terms of comprehensibility, Turilli and Floridi 2009). I have chosen to treat these terms as rough cognates in this paper first because I hold that these terms are problematic in parallel ways in the ML context, and second because making distinctions among these terms in the ML context requires the prior acceptance that these terms say something meaningful and precise about algorithms.

[2] I would like to thank an anonymous reviewer for raising this point.

property of a model rather than the deflationary notion of a synonym for the possession of some particular parameter value, the term cannot be rendered sufficiently precise to supply (for instance) researchers or public policy makers with a coherent notion of what, exactly, they would be trying to achieve in talking about the desirability of interpretability.

While this paper questions both the importance and the coherence of interpretability and cognates, it does not make a decisive case for the abandonment of the concepts. The aim of this paper is to provide a corrective response to the unclarity that has accompanied their present enthusiastic uptake. It seeks to issue a constructive challenge to the concept of interpretability—a challenge which aims both to encourage the development of alternative frameworks and to motivate better articulations of what interpretability is and why it matters. Skepticism about interpretability is one part in a process of foundational questioning whose goal is to make conversations about the challenges posed by ML algorithms both more pluralistic and more precise.[3]

2 Problems with the "Explicability Problem"

Many reasons have been given to explain why "explicability," "interpretability," and/or "transparency" are important desiderata: If we do not know how ML algorithms work, we cannot be sure that they will not fail, perhaps catastrophically so, when used in real-world environments (Tutt 2016). If we do not know how ML algorithms work, we cannot check or regulate them to ensure that they do not encode discrimination against minorities (Goodman and Flaxman 2016). If we do not know why an ML algorithm produces the output that it does, we will not be able to learn from instances in which it is mistaken. If algorithms lack transparency, domain experts or the public will not trust them (Kim 2015). People have a right to know why an ML algorithm has produced some verdict (such as lack of creditworthiness) about them (Goodman and Flaxman 2016). Algorithms whose inner workings are not interpretable will not enable us to produce causal explanations of the world (Ratti and López-Rubio 2018).

This section argues that the concepts of interpretability, explicability, and transparency have not been given the kind of definition they would need in order to have a role in solving these problems. It then questions whether any particular kind of knowledge of the inner workings of algorithms will be able to play a significant role in addressing these problems.

2.1 "Explanation" and "Interpretation" Are Difficult to Define

Although some degree of ambiguity affects any attempt to define a non-technical concept, "explanation" and "interpretation" are particularly difficult

[3] In the remainder of this paper, references to interpretability alone should be understood as a shorthand for reference to explicability, intelligibility, interpretability, and transparency, and references to interpretation alone should be understood as shorthand for interpretation or explanation, unless otherwise explicitly indicated.

to define.[4,5] In "The Mythos of Model Interpretability" (2018), Zachary Lipton draws attention to both the lack of definitional agreement among ML researchers and the absence of a technical meaning of interpretability. This section builds on Lipton's point by arguing that interpretability and its cognates are unclear notions. This problem is fundamentally a conceptual one. That is, we do not yet have a grasp of what concept(s) any technical definitions are supposed to capture—or indeed, whether there is any concept of interpretation or interpretability to be technically captured at all.

The difficulty of defining interpretability follows from what the concept is supposed to be able to capture. Presumably, the possession of an interpretation requires more than knowing a list of facts about something. There is a special kind of relationship that the knower is supposed to have with what they know in order to count as having an interpretation or explanation of it. But it is hard to spell out precisely what this special relationship consists in without recourse to a term that itself suggests a special kind of relationship to what is known. Parallel points can be advanced about explanation or "rendering intelligible" or "making transparent," and a survey of definitions available in the literature demonstrates this pattern:

> Interpret means *to explain or present in understandable terms*. In the context of ML systems, we define interpretability as the *ability to explain or to present in understandable terms to a human* (Doshi-Velez and Kim 2017).
>
> In the strictest sense, we might call a model *transparent* if a person can *contemplate* the entire model at once (Lipton 2018, emphasis mine).
>
> A second notion of *transparency* might be that each part of the model — each input, parameter, and calculation — admits an *intuitive explanation* (Lipton 2018, emphasis mine).
>
> In the *transparent* type [of representation], we represent the states of a system in a way that is open to *explicit scrutiny, analysis, interpretation, and understanding* by humans, and transitions between those states are represented by rules that have similar properties. A representation that is not transparent is an opaque representation (Alvarado and Humphreys 2017, emphasis mine).

[4] Participants in the interpretability literature make more fine-grained distinctions within these terms, for instance between a "global" explanation which provides some account of the overall pattern of functioning of the algorithm, and a "local" interpretation which provides some account of why the algorithm provided the output it did for a particular input (Doshi-Velez and Kim 2017). Since I raise concerns regarding whether the notion of interpretability has (or can have) any precise meaning with reference to algorithms, these objections apply equally to any qualified notion of interpretability.

[5] There is a substantial literature within philosophy of science concerning the nature of explanation. Hempel and Oppenheim's seminal "Studies in the Logic of Explanation" (1948) initiated a debate that subsequently included diverse contributions from Salmon (1984, 1989), van Fraassen (1977), Kitcher (1976, 1989), and Friedman (1974), among others. However, the proposals within this literature are largely orthogonal to the concerns of those seeking explicability or interpretability of ML algorithms. The literature on scientific explanation is concerned with what it means to scientifically explain the occurrence of events or phenomena, which is a very different task from "explaining" an algorithm in the way interpretability theorists have in mind. Section 1b expands on this point in discussing a difference between causal and justificatory notions of explanation. Symptomatic of the differences in both the desired type of explanation and the nature of the explananda, sample proposals in the literature lack clear relevance to the concerns of the ML literature (e.g., Hempel and Oppenheim's proposal that a scientific explanation consists in an argument that deduces the occurrence of an event from general laws).

A definition of interpretation, explanation, or transparency that relies on terms such as "understanding" or "intuition" cannot specify what makes an interpretation different from a list of facts. This is because the same question can be asked of the very terms that enter into the definition. One interpretability-like word is defined in terms of another interpretability-like word. These definitions move the bump under the rug.

This need not be a problem for these definitions in all possible scenarios. It is not necessarily required that all definitions be reductive, or that they omit any reference to terms cognate with the term to be defined. Furthermore, when there is tacit agreement about what would constitute an interpretation or explanation in some particular context (e.g., a teacher asks their student to "explain" why they gave the answer that they did), the practice of asking for an interpretation or explanation can productively proceed in the absence of critique.

However, in the contemporary discussion of ML algorithms, the tendency to define one interpretability-like word in terms of another is a significant problem. The designer of a tool that seeks to promote interpretability will have to judge the success of the various prototypes of their tools, and a regulator trying to enforce legislation that requires interpretability of an algorithm will have to determine whether an algorithm conforms to the law. Since there are not already well-functioning practices of providing something called an interpretation of an algorithm, it is hard to determine whether one has succeeded in providing an interpretation. If a tool teaches users some particular list of facts about an algorithm, or a company provides a list of facts about an algorithm, what would make it the case that this list of facts adds up to an interpretation, or that someone possesses understanding? In most cases in which there is an open question about whether (e.g.) an interpretation of an algorithm has been achieved, the same question is likely to arise for (e.g.) the understanding of an algorithm. Insofar as definitions of interpretation and cognates are part of an effort to establish practices of providing interpretations of algorithms, they ought to move beyond definitions that presuppose knowledge of how the cognate terms apply.

The absence of a clear articulation of what makes an interpretation different from a list of facts also leaves space for more extreme skepticism about interpretability-talk. The skeptic can charge that the reason an adequate definition has not yet been proposed is because there is not, in fact, anything that it is to *be* an interpretation or an explanation of an algorithm. There are only, by this account, lists of facts that are more or less useful for particular purposes. Sometimes, the skeptic might say, we call such lists interpretations, but this tracks only the subject's sense of confidence, which is an elusive and ill-defined feeling rather than a distinctive kind of knowledge or way of knowing. The concept of interpretability is the concept of a mysterious "I-know-not-what" which is not really there.

A different approach to defining interpretability would be to sidestep these issues by appealing to pragmatic criteria. Ratti and López-Rubio (2018) define "intelligibility" in the context of a discussion of intelligible models in molecular biology as "the ability to perform precise and successful material manipulations on the basis of the information provided by the model about its components" (p. 2). However, pragmatic criteria are insufficient to specify the concept. The ability to perform precise and successful material manipulations is more plausibly a consequence of interpretability than it is a statement of that in which interpretability itself consists.

Another way of trying to define interpretability is to identify its causes or grounds, such as a mismatch between how people think and the requirements of mathematical optimization (Burrell 2016). But it is hard to specify the cause in a way that only picks out the intended consequences. Regarding Burell's definition, there are many consequences of a mismatch between human thought and optimization, and not all of them are perspicuously identified as cases of "uninterpretability." For instance, such a mismatch can result in software being more difficult to maintain, but "being uninterpretable" is not the same thing as "being difficult to maintain," although one state might bring about the other.

Zachary Lipton executes a different version of the strategy of defining interpretability in terms of its grounds by exploring different technical features that are thought to increase interpretability, such as low dimensionality or decomposability. The different options that Lipton explores all have the important virtue of serving as concrete criteria for judging an algorithm.

But as crucial as the task of specifying such features may be, the different ways of supplying what Lipton calls a "formal technical meaning" to the concept of interpretability remain incomplete. This is because there is currently no clear non-technical meaning or definition of the concept. In the absence of such a non-technical meaning or definition, it is not possible to determine whether any list of technical features is supplying the technical meaning of interpretability, rather than the technical meaning of another concept. Indeed, such a list might be merely a collection of useful features which share a family resemblance but do not constitute any concept (as would be the case if, for instance, the extreme skeptic were correct). This issue reflects the more general point that it is hard to evaluate the success of an effort to articulate the grounds of a phenomenon without a corresponding specification of the phenomenon itself. The identification of technical features can still be useful in the process of arriving at such a specification, insofar as the examination of similarities between uncontroversially relevant technical features can stimulate reflection on what higher level concept the features are supposed to capture.[6] But the process of engaging in such reflection remains an open task, and particular proposals of technical specifications cannot in principle receive confirmation at the present stage in the dialectic.

One might argue that the specific technical criteria should replace a non-technical definition of interpretability, which is to say that they should serve as stipulative definitions of interpretability. But this approach would save only the word. Interpretability, explanation, and cognates are first and foremost epistemological concepts. A law that would require certain kinds of algorithms to be interpretable would not reflect lawmakers' desire that algorithms exhibit any particular technical feature. The law would rather reflect a concern to ensure that users of the algorithm have the right sort of knowledge about the algorithm which in turn enables them to have a very particular sort of relationship to that algorithm (although as this section has argued, it is hard to say what this amounts to). There is a putative epistemological concept for which the technical criteria Lipton explores would supply a corresponding technical meaning.

Having a more precise non-technical specification of the epistemological concept is important because, if we do not know what we mean by interpretability from an epistemological point of view, it is difficult to assess whether and to what extent any given set of technical features successfully resolves the black box problem in any given

[6] I would like to thank an anonymous reviewer for raising this point.

circumstance. A regulator who is trying to determine whether an algorithm conforms to legislation requiring interpretability cannot merely cite an algorithm's dimensionality and degree of decomposability. The regulator must also assess how the dimensionality and (non-)decomposability of the algorithm affect the capacities that human users can exercise. Likewise, legislators who might themselves try to provide a precise account of what an interpretable algorithm consists in cannot merely appeal to technical features of dimensionality and degree of decomposability. Choosing the right parameters and establishing the appropriate thresholds presuppose the availability of a general yet precise assessment of how particular technical features of algorithms affect the capacities of their users. But talk of "intuitive explanation," "contemplation," or "understanding" remains too abstract and vague to ground an assessment of how the technical features amalgamate into either the success or failure of an algorithm to meet higher level epistemological goals. What principles might one use to assess whether an algorithm's dimensionality or non-decomposability has led to (e.g.) the absence of an "intuitive explanation"?

An appropriate non-technical definition would serve as a bridge between abstract talk of interpretation and more technical desiderata such as the ones that Lipton explores. What is lacking is the right kind of conceptual work. Concepts such as interpretability need both a clear epistemological definition and a clear technical definition—or rather, an epistemological definition which has corresponding technical criteria.

It is worrying that so much importance has been afforded to interpretation in the absence of an adequate grasp of what the concept means when applied to algorithms. This is not to cast doubt upon the usefulness of recent work on techniques for providing information about the operation of ML algorithms (see Lipton 2018, Hall et al. 2017 for overviews). But in the absence of a definition of interpretability that enables the articulation of epistemological goals with a clear meaning in concrete circumstances, interpretability seems to be, at most, a useful way of labelling some family resemblances among a diverse body of technical work. The present lack of a suitable epistemological definition of a more substantial notion of interpretability renders that putative concept vulnerable to deflationary skepticism.

2.2 What Goes by the Name of an "Explanation" May Not Be as Useful as Expected

An interpretation or an explanation of an algorithm is supposed to provide insight into how it works. If we have an interpretation of an algorithm, we should be in a position to know why, when that algorithm generates a particular output, it produces the output that it does. While the last section highlighted the difficulty of determining what exactly an interpretation is, this section questions how the desired function of an interpretation—to provide an account of why the algorithm provides the outputs that it does—can be made relevant to ultimate ends of interpretation. It is not clear how or why knowing about the process that leads an ML algorithm to produce particular outputs leads to knowledge about the basis or justification for the output itself.

When a process leads to a particular answer to a question and we ask "why" that answer was produced, there are (at least) two ways of understanding that question. On the one hand, we could be asking how it came about that the process generated the output that it did. This is the causal sense of why. On the other hand, we could be asking what reasons speak in favor of giving that answer to that question. This is the justificatory sense of why.

Often, these two senses of why coincide, or are at least closely related. Suppose that a computer is following a simple procedure to predict the success of job applicants in getting to the interview round, and that it implements the instruction to predict that applicants with no experience will be rejected, while all others make it to the next round. In this case, the causal story and the justificatory story coincide. If you know what caused the computer to predict rejection as the outcome, you can also supply the reason that justifies believing that an applicant will be rejected. To the extent that an algorithm follows some version (albeit a more complicated one) of this template, information about the causal workings of the algorithm's categorization process has clear relevance to the justification for any given categorization.

However, there is a risk that the two senses of why come further apart in many ML contexts. This point is particularly apparent in the case of neural networks. The causal process by which some input triggers a certain pathway within the network does not straightforwardly map on to justificatory considerations. For instance, if you ask a person why they have given an answer to a particular question and they respond with an account of how their neurons are firing, they have given you information about the causal process that subserves the generation of their answer, without telling you anything that has clear significance to the justification of their answer. These two might ultimately have some relationship to one another, but in this case there is a difficult task of translating information about causal processes into considerations relevant to the justification of a categorization.

One common way to frame the black box problem in relation to ML algorithms is to discuss not knowing "how the algorithms work" or "why you get the output you do." But ML algorithms may be even more terra incognita than these formulations suggest. To the extent that causal process and justificatory considerations are likelier to come apart in ML than non-ML contexts, it is not clear that obtaining an interpretation of the process (or information about it) will prove as helpful as it does in non-ML contexts. While it is plausible that information about causal processes will ultimately prove relevant to knowing the reasons for outcomes, talk of a black box problem or interpretability problem is misleading insofar as it presupposes that an interpretation will satisfy the demands that lead to the articulation of the problem. There remains an important open problem of accounting for how to transition from knowledge of "how the algorithms work" to the justification for any given output.

2.3 Summing Up

There has been a recent explosion of interest in interpretability and cognates. Doshi-Velez and Kim (2017) report finding, via Google Scholar, over 20,000 publications about interpretability in ML from 2012 through 2017. This section has advanced doubts about the conceptual foundations of such research. The purpose of doing so is not to recommend an end to the interpretability literature, but rather to point out problems with its conceptual foundations in a manner that can motivate both the improvement of that literature and a clearer view of the boundaries of its usefulness. The concept of interpretability has enjoyed centrality in recent debates over the philosophical and social implications of ML algorithms. It is time to see if it can withstand skeptical scrutiny.

3 Reframing the Debate

Interpretability is often important not for its own sake, but because it might contribute to achieving some further goal. The purpose of this section is to argue that talk of an "interpretability problem" can obscure this well-known point, and to demonstrate that there are alternative ways of framing problems that would facilitate a more pluralistic approach to problem-solving. A number of motivations one might have in seeking explicability were cited earlier, such as safety, assurance of accuracy, and non-discrimination. To the extent that there can be ways of achieving these goals without anything that might have a claim to be called interpretability, it would be preferable to organize discussion in terms of those fundamental goals, such as a "(non)discrimination problem," rather than one particular means.

Because black boxes and interpretability dominate the conversation surrounding ML algorithms, it is difficult for academics, end users, and policy makers to achieve a clear view of the overall problem space and range of potential solutions. Reframing discussion in terms of the ends of interpretability would help avoid prematurely shutting down investigation via a tacit implication that there is only one way of achieving some given goal. It would also help prevent an overconcentration of attention and resources on one particular type of solution at the expense of other strategies. Such a reframing would be needed even if the definitional problems with the notions of interpretability were to achieve resolution.

3.1 The justification problem

Concerns about having "oversight" over ML algorithms, avoiding failures, and reassuring oneself about the accuracy of one's program all center around the epistemological notion of justification. In the most general terms, justification is the property that beliefs have when they are epistemically good, or when they are formed and maintained in the right kind of way.[7]

Interpretability is often thought to play an important role in justification in an ML context. It can seem outright irresponsible to believe algorithmic outputs regarding unseen real-world data in the absence of detailed knowledge of the algorithm's inner workings.

However, in some contexts, there are ways of achieving the desired assurance in the absence of knowledge about inner workings of tools. People were both responsible and justified in relying on the deliverances of their eyes prior to the development of a sophisticated and accurate science of optics. We are (at least sometimes) justified in relying on the outputs of our brains in the absence of a comprehensive picture of how all of the individual electrical and chemical signals combine in order to produce higher level cognitive phenomena. This is not simply because of a lack of a more familiar alternative.[8] If prior to the development of a science of optics, people had an "explicable" but unreliable device that could stand in for sight, the unreliable device would not be preferable in virtue of its familiarity. Reliability can sometimes beat out familiarity.

[7] Some epistemologists eschew talk of "justification" in favor of talk of what confers the status of knowledge (rather than mere belief). Here, I use the terminology of justification because it sounds less jargon-laden than would a more general term such as "epistemic goodness" or "epistemic responsibility," although this section could be restated using only these more general terms.

[8] I would like to thank Emma Pierson for raising this point.

The point of these examples is not to suggest that eyesight and brain use are precise parallels to the use of ML algorithms. The point is rather to demonstrate that there are ways of being justifiably confident in outputs that circumvent the call for interpretation and cognates. For instance, reliabilism is an approach to thinking about justification or knowledge that many analytic epistemologists have endorsed in the last several decades (Goldman 1967; Armstrong 1973; Sosa 1999, 2007; Williamson 2000). Roughly speaking, reliabilists hold that a belief is justified or that we count as knowing (rather than merely believing) just in case a belief is formed by a reliable process. The reliabilist position provides a plausible account of why relying on the outputs of one's visual system or brain can be responsible even in the absence of a detailed scientific theory: because the system reliably produces the right output. In some cases, a record of success can stand in for a precise account of the inner workings of an instrument.[9]

The in-principle availability of reliability as a way of achieving justification suggests the need for a pluralistic discussion regarding the acceptance of outputs from ML algorithms. Yet both the position that one needs to "oversee" epistemic processes in order to be responsible in relying on their results (Mittelstadt et al. 2016; Tutt 2016) and the position that "oversight" requires an interpretation of one's instruments leave no room for alternative ways of thinking about justification. To the extent that the appropriate form of justification can vary depending on the particularities of algorithms, training environments, and contexts of use, the language used to discuss problems surrounding ML algorithms should not imply the superiority of one solution.

3.2 The anti-discrimination problem

ML-based algorithms raise various concerns about discrimination, and one might think that interpreting the process by which the algorithm reached its results is important to combatting this discriminatory potential. But an examination of real case studies shows that the importance of interpretability can easily be overstated in this domain.

Consider the infamous case of the Google AI which, when given the task of labelling different photographs, identified Black people as gorillas. Upon an examination of the training set, engineers realized that the training data did not contain any images of Black people, and that the classifier therefore categorized the images of Black people along with the only other figures with dark complexions that it had been given in the training set.

The problem is diagnosable without detailed knowledge of the steps of the classification procedure performed on inputs. Upon seeing the absence of images of Black people in the training set, and knowing the bad outcome, it is not difficult to reconstruct what must have happened. Here, the underlying problem had to do with the training set, and this problem manifested itself both at the level of the algorithm's classification procedure and at the level of the final output. If one were not able to examine the classification procedure, it would still be a viable option to develop better procedures for reviewing the construction of training sets, or for testing classifiers on sample data sets which are designed to test how the algorithm treats people of different identity categories.

[9] Recently, Alvarado and Humphreys (2017) have suggested that reliabilism provides an appropriate theory for making sense of how users can responsibly interact with ML algorithms in the absence of significant knowledge of the inner workings of the algorithm.

A second type of problematic case consists of instances in which recommended outcomes might end up tracking identity categories and thereby lead to discrimination. Credit rating algorithms and algorithms that recommend whether an imprisoned person should be granted parole are examples of use cases in which the potential for identity-based discrimination is clear. Here, as in the case of Google's categorization problem, better ways of vetting training data sets and better ways of testing the resulting classifiers (e.g., with fake data sets of people who differ only with respect to race) can be ways of ensuring that algorithms do not discriminate. These testing strategies would mirror the social-scientific experiments that ask two groups of reviews to rate job applications, writing samples, and so on, changing only some specific identity category between groups (see, e.g., Steinpreis et al. 1999, Moss-Racusin et al. 2012).

While this type of approach might not work for every possible scenario, there is considerable potential for combatting discrimination even if ML algorithms cannot "give us their reasons" like a person does. Moreover, people are not always forthcoming in reporting the actual grounds for their decisions, especially in cases where discrimination is at issue. The inability of an ML algorithm to "give reasons" should not be seen as an overwhelming problem from the point of view of preventing discrimination, and may even prove a benefit, insofar as this feature decreases the amount of time that is spent puzzling over the meaning of inaccurate self-reports.

Although discussion of interpretability has been motivated by the need to combat discrimination (Goodman and Flaxman 2016), there is no obvious link between fighting discrimination and having knowledge of the inner workings of algorithms. Interpretability is more plausibly construed as one tool among many which may prove useful against discrimination.

3.3 The reconciliation problem

A third issue which might motivate a call for interpretability is the question of how to reconcile human judgments and ML outputs. I use the term reconciliation to mean the process of determining how to synthesize the outputs of different sources into an overall decision and corresponding level of confidence in that decision. This is a problem that arises in any real-world context in which an ML algorithm is used not as the single determinant of a decision, but rather as one source among many. For instance, in making decisions in a clinical context, doctors might consider each of their knowledge of relevant studies, their clinical experience, and the output of an ML algorithm in order to arrive at a single diagnosis (Montgomery 2006; van Baalen and Boon 2014). Strictly speaking, the outputs of ML algorithms are more analogous to the beliefs of a reasoner than a source of data. ML algorithms respond to diverse indicators in order to arrive at an overall verdict, rather than putting forward evidence or generalizations (e.g., that a certain drug is effective in 60% of a certain category of cases) that are subsequently translated into the overall verdict. The particular issue they raise is therefore most precisely analogous to the problem of how to reconcile the beliefs of different reasoners. One reason to seek an interpretation of how ML algorithms work is to facilitate this process. The problem of what can be called positive reconciliation arises when the outputs of different sources coincide, which raises the question of how much one ought to increase one's confidence in the solution in virtue of that coincidence. The problem of what can be called negative reconciliation arises when outputs of different sources diverge, which raises the problem of what to believe and how much confidence to assign to that belief.

Although these problems have not received explicit mention in the interpretability literature, they might nonetheless reasonably motivate a call for something like interpretability, insofar as supplying one's reasoning is frequently central to the reconciliation procedures that people use when dealing with one another. Likewise, when considering (e.g.) how to synthesize the results of a medical study with clinical experience, the data and methodologies of each distinct source are in principle available to the ultimate decision maker, which means that the grounds of potential disagreement can be rendered explicit (at least to some non-trivial degree). Many strategies developed in the interpretability literature facilitate an approach to reconciliation that would seek to make ML "reasoning" resemble human reasoning to the extent possible. For instance, algorithms that translate neural networks into decision trees identify in the workings of ML algorithms the same kind of general categorization principles that humans might use. Here, reconciliation requires homogeneity or at least similarity between reasoners. Correspondingly, the most recent philosophical work on the "epistemology of disagreement" focuses on what is called "peer disagreement," or disagreement between people with relatively similar capacities and evidence (Kelly 2005, 2010; Frances 2014). That work which does address disagreement between non-peers focuses on disagreement between people with different degrees of competence or experience, rather than disagreement between systems of fundamentally different kinds (Goldman 2001; Frances 2014).

However, there can be different approaches to reconciliation problems that allow human and ML processes to remain fundamentally distinct. For instance, one crucial piece of information is what factors or indicators each process tracks. When human processes and ML processes track different indicators, they are independent sources whose coincidence should strengthen confidence in the accuracy of a shared conclusion, whereas tracking the same indicators (and especially using the same indicators in a different way) can strengthen confidence in the aptness of the way that human reasoners are processing a given data set, without providing fully independent evidence in favor of the accuracy of that conclusion. Both scrutiny of the content of training data sets and ways of testing classifiers to see what features they actually track are viable ways of extracting this information without scrutiny of the steps that the algorithm performs in arriving at categorizations. This point therefore parallels the alternative suggestions made previously in the case of anti-discrimination measures.

This brief discussion of reconciliation problems takes place at a very high level of abstraction. It is likely that in practice, the development of useful strategies for reconciliation (and therefore the question of the importance of interpretability) will be domain-specific and perhaps even case-specific (see, e.g., Tonelli 2006). Nonetheless, the near-ubiquity of these problems in real-world decision-making contexts, and especially contexts such as finance, medicine, and public policy in which ML algorithms can have or are having a significant impact, motivates the explicit mention of these problems and suggestion of the breadth of the solution space.

3.4 Where Interpretability Counts the Most

The second part of this paper has thus far examined alternatives to interpretability-based solutions. While interpretability at minimum requires some kind of information about the inner workings of ML algorithms, the proposed alternatives circumvent interpretability by appealing only to "external" factors such as information about reliability and features of input and training data sets. However, there are certain problems for which interpretability, or at minimum information about the inner workings of the classification procedures of ML algorithms, seems harder to replace.

The use of ML algorithms by scientists for the purpose of generating causal explanations is one such case. Roughly speaking, causal explanations are explanations that appeal to the interrelationship of the factors that produce or maintain a phenomenon, and they are plausibly central to the aims of many scientific disciplines (Ratti 2018). Ratti and López-Rubio (2018) have emphasized a connection between interpretability and causal explanation in molecular biology. They argue that models facilitate causal explanation by helping scientists identify particular causally relevant factors and their interactions. Ratti and López-Rubio appeal to the arguments developed in relation to cognitive and systems neuroscience by Kaplan and Craver (2011), who hold that models in these areas of neuroscience must meet a set of requirements according to which "(a) the variables in the model correspond to components, activities, properties, and organisational features of the target mechanism… and (b) the (perhaps mathematical) dependencies posited among these variables in the model correspond to the (perhaps quantifiable) causal relations among the components of the target mechanism" (p. 611). To the extent that the process of discovering causal explanations requires that humans examine models in order to identify causally relevant components and their interactions, it is highly plausible that something like interpretability, or at least certain information about internal processes, must be available to scientists. Furthermore, insofar as the goal of scientists is not only to develop causal explanations but to develop causal explanations that are amenable to human cognitive capacities, models that identify smaller numbers of relevant factors and simpler forms of dependencies are preferable (Craver 2006; Ratti and López-Rubio 2018). It is hard to see an easy replacement for something like interpretability for models that are to be used this way in the context of scientific discovery.

Likewise, there is the case of public trust. To the extent that public acceptance of ML algorithms requires that end users have some grasp of the inner workings of what they are relying upon, the notion of interpretation acquires heightened importance. However, this case differs from that of scientific explanation insofar as the problem of public trust is not independent from the interpretability literature on ML algorithms itself. What the public requires in order to trust ML algorithms does not form in a vacuum, but is rather informed by the discussions of experts and academics, especially as filtered through popular scientific media outlets. To the extent that those working on interpretability emphasize its indispensability, they can contribute to heightened public mistrust of ML algorithms. Addressing public trust in a review of the motivations for interpretability therefore requires particular caution. It is possible that many public trust issues could be to some extent addressed with robust procedures for testing the accuracy of ML algorithms, in conjunction with the development of regulations and best practices surrounding the application of such tests to ML algorithms. While this point does not by itself undermine the idea that interpretability could facilitate public acceptance of ML algorithms, it does recommend heightened attention to the porous boundaries between public and academic spheres when appealing to the desires of the public.

3.5 Summing Up

When formulating a problem, it is preferable to minimize the number of non-essential philosophical commitments that are built into the statement of that problem. Focus on an interpretability problem can undermine the kind of modularity that would make it possible to draw on, for example, the full range of literature about justification in contemporary epistemology. Where feasible, talk of interpretability should be eschewed in favor of talk of the ends of interpretability.

4 Conclusion

The development of the vocabulary of black boxes and interpretability responds to an urgent need to facilitate social, political, and philosophical discussions of increasingly ubiquitous ML algorithms. This vocabulary is the first and primary resource of its kind. The widespread and rapid adoption of ML algorithms has led to the equally widespread and rapid adoption of the only conceptual framework currently available for talking about what makes the algorithms seem so strange and different. Discussion of the black box problem and the "uninterpretability" of ML algorithms serves as a useful check on the enthusiasm that has accompanied the practical successes of those algorithms. However, the interpretability framework has itself experienced an enthusiastic uptake which in turn generates a need for critique.

This paper has interrogated the concept of interpretability by questioning both the coherence of the concept itself and the place of that concept within the broader discourse on ML algorithms. The first section argued that the concepts of interpretability and cognates lack the kind of definition that would render them adequate to the kind of work that their proponents want them to do, and also suggested that interpretability may end up being of more limited use than is often thought. The second section argued that since interpretability is most often proposed as a means to further ends rather than an end in itself, it would be more perspicuous to organize discussion around the fundamental problems rather than one putative solution. The language of interpretability is unhelpful when it dominates the academic and the public imagination.

After setting aside the hype of the black box problem, there remains the point that ML algorithms have generated an urgent need for more conceptual work concerning the nature of ML algorithms and how humans interact with them. This paper has advanced a critique of the notions of "interpretability," "intelligibility," "explicability," and "transparency" in the hope that the conversation surrounding ML algorithms might ultimately become as far-reaching and fundamental as the changes that the algorithms have brought about.[10]

[10] This work was initiated with the support of Bridgewater Associates. I would like to thank in particular Jacob Kline. For extremely helpful suggestions and corrections, I owe thanks to Logan Graham, Emma Pierson, and Colton Valentine, as well as three anonymous reviewers.

References

Alvarado, R., & Humphreys, P. (2017). Big data, thick mediation, and representational opacity. *New Literary History, 48*(4).

Armstrong, D. (1973). *Belief, truth, and knowledge*. London: Cambridge University Press.

Burrell, J. (2016). How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1).

Craver, C. (2006). When mechanistic models explain. *Synthese, 153*, 355–376.

Doshi-Velez, F. & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. Preprint from arXiv. arXiv:1702.09608v2 [stat.ML].

Frances, B. (2014). *Disagreement*. Cambridge: Polity Press.

Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy, 71*(1), 5–19.

Goldman, A. (1967). A Causal Theory of Knowing. The Journal of Philosophy. 64(12), 357-372.

Goldman, A. (2001). Experts: which ones should you trust? *Philosophy and Phenomenological Research, 63*, 85–110.

Goodman, B., & Flaxman, S. (2016). European Union regulations on algorithmic decision-making and a 'right to explanation'. Preprint from arXiv. arXiv:1606.08813 [stat.ML].

Hall, P., Phan, W., & Ambati, S. (2017). Ideas on interpreting machine learning. *O'Reilly,* https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning. Accessed 1 August 2018.

Hempel, C., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*(2), 135–175.

Kaplan, D., & Craver, C. (2011). The explanatory force of dynamic and mathematical models in neuroscience: a mechanistic perspective. *Philosophy of Science, 78*(4), 601–627.

Kelly, T. (2005). The epistemic significance of disagreement. In T. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology* (Vol. 1). Oxford: Oxford University Press.

Kelly, T. (2010). Peer disagreement and higher order evidence. In R. Feldman & T. Warfield (Eds.), *Disagreement*. New York: Oxford University Press.

Kim, B. (2015). Interactive and interpretable machine learning models for human machine collaboration. PhD thesis, MIT.

Kitcher, P. (1976). Explanation, conjunction, and unification. *Journal of Philosophy, 72*(8), 207–212.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (Eds.), *Scientific explanation*. Minneapolis: University of Minnesota Press.

Leese, M. (2014). The new profiling: algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union. *Security Dialogue., 45*(5), 494–511.

Lipton, Z. (2018). The mythos of model interpretability. *Queue. 16*(3), 31–57.

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: mapping the debate. *Big Data & Society, 3*(2).

Montgomery, K. (2006). *How doctors think: clinical judgement and the practice of medicine*. Oxford: Oxford University Press.

Moss-Racusin, C., Dovidio, J., Brescoll, V., Graham, M., & Handelsman. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences USA, 109*(41).

Ratti, E. (2018). 'Models of' and 'models for': on the relation between mechanistic models and experimental strategies in molecular biology. *British Journal for the Philosophy of Science*, axy018.

Ratti, E. & E. López-Rubio. (2018). Mechanistic models and the explanatory limits of machine learning. Draft paper for the symposium *Mechanism Meets Big Data: Different Strategies for Machine Learning in Cancer Research*, Seattle WA.

Ribeiro, M., S. Singh, & C. Guestrin. (2016). *"Why should I trust you?" Explaining the predictions of any classifier."* Preprint from arXiv. arXiv:1602.04938v3 [cs.LG].

Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

Salmon, W. (1989). *Four decades of scientific explanation*. Minneapolis: University of Minnesota Press.

Sosa, E. (1999). How to defeat opposition to Moore. *Philosophical Perspectives 13, 141–153.*

Sosa, E. (2007). *A virtue epistemology: apt belief and reflective knowledge, volume I*. Oxford: Oxford University Press.

Steinpreis, R., Anders, K., & Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants. *Sex Roles, 41*(7–8), 509–528.

Tonelli, M. (2006). Integrating evidence into clinical practice: an alternative to evidence-based approaches. *Journal of Evaluation in Clinical Practice, 12*(3), 248–256.

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology, 11*(2).

Tutt, A. (2016). An FDA for algorithms. 69 Admin. L. Rev. 83, Social Science Research Network,https://ssrn.com/abstract=2747994. Accessed 1 August 2018.

van Baalen, S., & Boon, M. (2014). An epistemological shift: from evidence-based medicine to epistemological responsibility. *Journal of Evaluation in Clinical Practice., 21*, 433–439.

van Fraassen, B. (1977). The pragmatics of explanation. *American Philosophical Quarterly, 14*(2), 143–150.

Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.