Machine learning, justification, and computational reliabilism

Juan M. Durán

Unpublished manuscript

Abstract

This article asks the question, "what is reliable machine learning?" As I intend to answer it, this is a question about epistemic justification. Reliable machine learning gives justification for believing its output. Current approaches to reliability (e.g., transparency) involve showing the inner workings of an algorithm (functions, variables, etc.) and how they render outputs. We then have justification for believing the output because we know how it was computed. Thus, justification is contingent on what can be shown about the algorithm, its properties, and its behavior. In this paper, I defend *computational reliabilism* (CR). CR is a computationally-inspired off-shoot of process reliabilism that does not require showing the inner workings of an algorithm. CR credits reliability to machine learning by identifying *reliability indicators* external to the algorithm (validation methods, knowledge-based integration, etc.). Thus, we have justification for believing the output of machine learning when we have identified the appropriate reliability indicators. CR is advanced as a more suitable epistemology for machine learning. The main goal of this article is to lay the groundwork for CR, how it works, and what we can expect as a justificatory framework for reliable machine learning.

1 Introduction

The use of Machine Learning (ML) for scientific purposes is delivering remarkable results. A couple of examples will suffice to show this. In molecular biology, AlphaFold can predict protein structures with atomic accuracy for cases in which no similar structure is known (Jumper et al., 2021). In medicine, BenevolentAI has combined structured with unstructured biomedical data sources to identify rheumatoid arthritis drugs like baricitinib as therapeutics for COVID-19 symptoms (Medeiros, 2021). It is clear that ML

can successfully extend the class of tractable chemistry, biology, physics, and medicine, broadening the range of modeling and experimental capabilities of researchers.

Yet, unlike many other methods, ML's scientific value cannot be easily determined by association with a body of scientific knowledge, by means of adequacy to empirical data, or supported by theoretical constructs (e.g., explanation and observation). This for a variety of reasons. ML is epistemically and methodologically opaque (Humphreys, 2009), making it difficult to associate a given algorithm and its output with the general scientific canon; and empirical phenomena are often temporarily, spatially, or cognitively inaccessible for empirical validation of the model, casting doubts over the representational value of these systems. As a consequence, there are significant impediments for making claims about our reliance on these systems and their output.

When confronted with these issues, philosophers and computer scientists gravitate towards *transparency*. Transparency is an umbrella term capturing diverse methods linking the internal mechanisms and properties of algorithms to its outputs (Creel, 2020; Wachter et al., 2018; Ribeiro et al., 2016). To see how transparency works, consider BenevolentAI. At its core, BenevolentAI is a search engine combining structured with unstructured biomedical data sources, drug industry data, and automated retrieval of information from scientific research papers. The data is curated and standardized via data analysis and data fabric. It is then fed into knowledge graphs that structure the data into relationships between diseases, genes, and different drugs (Smith et al., 2021). Richardson led the team that used BenevolentAI to identify rheumatoid arthritis drugs – notably baricitinib – as suitable therapeutics for COVID-19 symptoms (Richardson et al., 2020). In order to justify Richardson's reliance on this output, partisans of transparency make efforts to show how baricitinib obtains from procedures integrating biomedical data, functions identifying new structural relationships that associate possible causal relationships within the system, and other relevant algorithmic operations. In the case of BenevolentAI, one way to make the algorithm transparent is via a *knowledge graph* (Richardson et al., 2020, 30). This visualizes how baricitinib inhibits AAK1 (associated with interrupting the COVID-19 virus' passage into the cells) and JAK 1/2 (critical for signal transduction pathways), and how baricitinib binds with GAK (know to decrease certain viruses' infectiousness). This knowledge graph also provides reasons to consider drugs like fedratinib, sunitinib, and erlotinib as less effective and, depending on the case, unsafe. For instance, it is shown how these drugs only inhibit AAK1, and neither decrease the chances of cell infection (by binding with GAK) nor inhibit cytokine signaling (by inhibiting JAK 1/2) (Richardson et al., 2020, 30).

Are Richardson and his team right in thinking that baricitinib is a scientifically valid outcome given the conditions set up for the ML? What reasons do they have to discard other drugs as either less effective or unsafe? What supports their claim that Benev-

olentAI is a reliable system for the intended purposes? These are questions about the epistemic reliance of ML and the justification of their outputs. To a great extent, transparency provides answers to these questions. This paper, however, is an effort to provide an alternative answer, one that does not depend on methods for transparency of the algorithm. More specifically, this paper lays the groundwork of *computational reliabilism* (CR), an algorithmic-centered reliabilist framework for the justification of ML outputs.

As the name suggests, CR borrows from epistemological reliabilism, notably Alvin Goldman's (2012) process reliabilism. But, unlike process reliabilism, CR does not take as pre-theoretical assumption that ML is a reliable belief-forming method. Instead, a central issue for CR is to find out which *reliability indicators* (RIs) credit reliability to ML. A large portion of this paper is therefore geared towards identifying pertinent methods (formal and otherwise), algorithmic metrics, expert competencies, cultures of research, and the like that make up for our best epistemic and normative efforts for the reliability of ML. Also, unlike process reliabilism, the discussion of what constitutes 'truth' here is rather exiguous. I will not discuss RIs as truth-conductive, nor as truth-makers, fact-checkers, and so on. I will though touch on the issue of when a belief *is* or *counts* as justified (even if I will not discuss whether a belief is 'true'). As will become clear when discussing RIs, my thesis supports the idea that belief, justification, and knowledge are historically situated, incremental, and perspectival (Massimi and McCoy, 2020).

With these ideas in mind, this paper is divided into the following sections. Section 2 of this paper, I discuss the scientific importance of, and the difficulties in, obtaining justification for believing ML algorithm outputs. To do so, I briefly discuss an ML system designed to infer criminality through facial trait analyses. This example is only used to support the claim that, for certain systems, transparency falls shorts in its justification. The same example is used to motivate the justificatory value of CR. In section 3, I lay the groundwork for *computational reliabilism* (CR). I present three modes of *reliability indicators* (RIs) that credit the reliability of MLs. These are (1) technical robustness of algorithms (subsection 3.1), (2) computer-based scientific practice (subsection 3.2), and (3) social construction of reliability (subsection 3.3). In section 4, I take stock on my findings and suggest further lines of investigation that substantiate the merits of CR. In gist, this article invites us to reflect on a crucial, but often overlooked, question: under what conditions are researchers justified in believing ML outputs? My answer is that reliability comes through a myriad methods, practices, and processes tailored to ML algorithms. This article is an attempt to provide a suitable justificatory framework for the reliability of ML.

2 Merchants of mistrust

A central motivation for seeking justification is that ML algorithms are often methodologically and epistemically opaque. This metaphor has two distinct but related interpretations. The first interpretation addresses how ML algorithms involve multiple complex elements (functions, variables, decisions, data, etc.) in their design, coding, execution, and maintenance. This means that little can usually be said about how these algorithms cluster data, which criteria are used for creating categories, and overall why ML algorithms behave the way they do. This interpretation is captured in the epithet 'black-box' algorithms as a way to express how far-removed algorithms are from insight. The second interpretation sets the focus on our limited cognitive capacities to know the relevant elements in the algorithm (Humphreys, 2009, 649). That is, no human being (or group of human beings) can *identify* which functions, variables, decisions, data, etc. are relevant to a given ML output (clusters of data, categories, etc.).

Whereas the first interpretation focuses on the algorithm as an opaque method, the second highlights our cognitive limitations to say something meaningful about it and its output. Both interpretations, however, can be cast as showing a lack of justification for the algorithm's output. Either because the ML is a black-box or because human agents are cognitively limited, there are no basis for having justification for believing that the algorithm's output has any scientific value.

Under this heading, transparency surfaces as a promising solution. It first enables an agent to survey an algorithm by uncovering its inner mechanisms and properties; then, it enables the agent to link these mechanisms to the algorithm's output. As a result, the agent is justified in believing the algorithm's output because she has access to the relevant functions, values, etc. responsible for it. Recall from the introduction that BenevolentAI utilizes knowledge graph to visualize how the algorithm favors baricitinib over other drugs. Another example is LIME: a general algorithm that accounts for the predictions of any classifier by locally learning an interpretable model. Formally, LIME produces a model g $\in$ G, where G is a class of potentially interpretable models (e.g., linear models, decision trees, or falling rule lists). In practice, if an ML system predicts that a patient has the flu, LIME can highlight the symptoms in the patient's history responsible for the prediction. 'Sneeze' and 'headache', for example, are key variables used by the algorithm. They are flagged as net contributors to the flu prediction. In contrast, 'no fatigue' is a variable used as evidence against the prediction (Ribeiro et al., 2016).

CR suggests a different approach. Here justification comes from identifying which RIs are relevant for crediting reliability to the algorithm. In this sense, neither CR or the RIs depend on knowing the internal mechanisms and properties of the algorithm. While I explore RIs in depth in the next section, a preliminary conclusion can be drawn: we

can say that transparency and CR have different justificatory modes. With the former, we have justification by having access to the inner workings of the algorithm. With the latter, we have justification by identifying methods (formal and otherwise), metrics, expert competencies, cultures of research, and the like that make up for our best epistemic and normative efforts that might increase the degree of warrant we have to believe the outputs of ML systems.

I now briefly discuss an example of ML where these two contrasting modes of justification surface. Here, transparency justifies our belief that the output of this ML has scientific value, while CR flags the ML as unreliable and therefore we are lacking justification for such belief (I return to this discussion in section 3). It goes without arguing that the example is only meant to contrast the justificatory mode fostered by transparency – or to be more precise, the interpretation of transparency here advanced – against CR's. No conclusions about their comparative justificatory value can be derived from it.

In 2016, computer scientists Xiaolin Wu and Xi Zhang developed a Convolutional Neural Network (CNN) that analyzed over 1,850 ID photos.[1] About 1,120 of these photos were of people with no criminal convictions; 730 were of people who were either wanted for crimes or convicted of crimes. The CNN's operation was simple. It picks out facial traits (e.g., distance between the eyes, length and curvature of the mouth). It then classifies each photo as 'criminal' or 'non-criminal.' The predictive accuracy was measured using the Area Under the Receiver Operator Characteristic Curve (AUC-ROC). The result was impressive: Wu and Zhang measured 0.9540 accuracy. This means that the CNN was able to successfully classify 'criminal' versus 'non-criminal' faces approximately 95% of the time (Wu and Zhang, 2016, 2).

To further validate the CNN and rule out that such a high predictive accuracy resulted from overfitting, Wu and Zhang retrained their CNN on a dataset where the labels 'criminal' and 'non-criminal' were assigned randomly as negative and positive instances with equal probability. For the retraining case, the CNN failed to distinguish between the two categories, plummeting the average classification's accuracy to 48%, with a false negative rate of about 51%, and the false positive rate close to 50%. Wu and Zhang also accounted for problems related to unbalanced datasets, choice of photos (light, angle, clothing, etc.), and other issues pertaining to accuracy. To most algorithmic standards, these results speak in favor of a reliable CNN capable of consistently classifying the photos in question.

Wu and Zhang naturally defend the scientific merits of their algorithm. To their mind, high predictive accuracy means that the algorithm's output have scientific value. It is no coincidence that they confidently announce the "law of normality for faces of non-

[1] Wu and Zhang's use of photos from actual people contrasts with other approaches that use synthetically generated photos (Turk and Pentland, 1991; Blanz and Vetter, 1999).

criminals" (Wu and Zhang, 2016, 8). But high predictive accuracy is no standard for claims about scientific value. While the Ptolemaic model exhibited high predictive accuracy, its predictions have no scientific value for astronomical explanations and predictions simply because they do not represent planetary motion. Additionally, predictive accuracy can be manufactured by carefully selecting the input data, and calibrating variables and functions in the algorithm to some desired degree. For example, finding optimal values for hyperparameters (number of hidden layers, batch size, choice of activation function, etc.) is fundamental for having faster convergence, high accuracy, and overall better results of ML. Now, algorithms allow multiple optimal hyperparameter configurations depending on datasets, purposes of the algorithm, and tasks (Morales-Hernández et al., 2022). Furthermore, optimal configurations for one algorithm do not typically translate to others, making them incompatible in many different ways (van Rijn and Hutter, 2018). As result, selecting optimal values for hyperparameters, along with the best configuration for a given algorithm, is largely a matter of human decision. Without further provisions in place, such as ensuring compliance with scientific standards, professional integrity, and standardized measurements for the optimality of hyperparameters, predictive accuracy can (relatively easily) be manufactured.

Now, Wu and Zhang believe that high predictive accuracy grants scientific value to their CNN's outputs. To further defend this, they retrain the parameters of every layer in the CNN while retaining its architecture (Wu and Zhang, 2017, 3). As a result, the high accuracy in the output remains and thus the algorithm is technically robust. That is, the CNN correctly picks out specific facial attributes from photos, and then classifies them into the appropriate category 95% of the time.[2] Unfortunately, taking high predictive accuracy as an indication of the reliability of automated inference on criminality algorithms is problematic. It confounds justification of a technically correct and well-executed procedure with the justification required for belief. Wu and Zhang have no justification for believing that someone is a criminal based on facial traits alone. This is the case regardless of how accurate their algorithm is at picking out and classifying photos.

In this context, transparency does not seem to be of much help for justificatory purposes. When Wu and Zhang try to justify their outputs on high predictive accuracy, they look at what their AUC-ROC values are telling them. This means that specific functions and properties of the CNN are responsible for the output. But justifying the CNN's output using the same functions used to render them is epistemically circular and inadmissible as justification. Thus understood, transparency does not evaluate the

[2]Despite these efforts, the system's high accuracy remains questionable. While there is no evidence of output manipulation, one can't help but wonder whether the system would maintain the same level of accuracy when faced with a larger and more diverse dataset.

scientific value of algorithms' output, it only speak of their performance. It follows that Wu and Zhang can pin down the functions and properties of their CNN that account for the high predictive accuracy, but at no point can they use those functions and properties alone for claims about justification. Under conditions of high predictive accuracy and transparency, Wu and Zhang are as justified in believing that a photo is of a 'criminal' as they are in believing that it is of a 'non-criminal.' It follows from this case that this form of transparency does not help to distinguish what we are compelled to believe from what cements that belief.[3]

3 Modes of computational reliability

Claims about justification find home in CR, a branch-out of process reliabilism (Goldman, 2012). At is core, CR is a frequentest approach in which beliefs formed by reliable algorithms are better justified than beliefs formed by unreliable ones. As a reliable belief-forming method, ML tends to produce more outputs with scientific value (i.e., outputs that correspond to accepted scientific results, satisfy standards of scientific research, are scientifically sound, and so on), than outputs without scientific value. To credit reliability, CR utilizes *reliability indicators* (RIs) as markers of methodological and epistemological competence of the computer, algorithm, and social processes involved in the formation of beliefs. RIs can be understood as algorithmic-related methods and practices with a reliability-conferring property. Although I will not discuss the nature of reliability-conferring properties, it should not be taken as a 'spooky' property of said methods and practices. Claims about the reliability of an observation, for example, presuppose properties of the implemented method that concede its reliability (good lighting, the existence of the entity to be observed, etc). Similarly, the reliability-conferring property of RIs will become clearer once I analyze each indicator in some detail. For now, I identify three RIs that form the basis of CR:

- RI_1 *Technical robustness of algorithms* focuses on the design, coding, execution, maintenance, and other technical features that makes an ML system robust. This includes the collection, curation, storage, distribution, and analysis of data; parametrizations; modularity; kludges (Clark, 1987); and other practices pertaining to algorithms.

- RI_2 *Computer-based scientific practice* focuses on ML-based scientific research. It

[3] As suggested earlier, transparency is a broad concept that admits different interpretations. A partisan of post-hoc explanatory AI, for instance, could argue that the algorithm was not taking into account scientifically salient aspects of the pictures. By having a XAI tool pointing to high-level features that refer to domain knowledge (e.g., list of criminality-based characteristics), then the XAI can be used to evaluate the scientific soundness of the algorithm.

results from the algorithmic implementation of scientific theories, principles, hypothesis, and other relevant units of scientific analysis. It also accounts for scientific interactions, debates, and other ways of engaging in scientific research that involve ML;

- RI_3 *Social construction of reliability* focuses on broader goals related to accepting ML and its outputs in diverse communities (e.g., scientific, academic, public communities). This occurs through debate and similar forms of intellectual exchange.[4]

Under this heading, CR is understood as a family of reliability-eliciting algorithmic-related indicators capable of crediting ML as a reliable belief-forming method. Accepting a frequentist likelihood acknowledge that ML systems are, on occasion, inefficient, contain errors, are unsuitable for specific purposes, and compute the wrong data. If failures perpetuate over time, the relative frequency governing CR will flag an ML as unreliable. Furthermore, CR takes note of human cognitive limitations to access some RIs. It also accommodates the fact that RIs are neither absolute nor universally applicable. Not all RIs can be credited as reliable under the same criteria, nor does the same RI apply equally to all ML. CR should thus be understood as perspectival, provisional, and subject to corrections. No RI should be taken as having an all-or-nothing property, or whether the indicator applies.

Thus understood, RIs come in degrees. The degree to which one RI is more relevant than another will depend on the context in which ML is applied. It will depend on the (scientific or social) values and goals at stake, and on the community designing, coding, and using the system. No individual RI guarantees the reliability of ML. Even if some RI is currently suitable, this does not ensure that we got it right, that we applied the RI correctly, or that our reliability claims are eternally warranted. Science changes fast, and technology changes even faster. Old RIs can lose their appeal as new ones come to the fore. In this respect, this paper holds no pretensions to claim the completeness of the various RIs presented here. Further arguments could be given on the need for additional indicators not discussed here, on misplaced RIs, and on RIs in need of replacement. This, however, is not to say that there are no stable RIs that apply across ML systems. In fact, most of the RI discussed here have a kind of permanence in time despite changes and fine-tunings that occur with new developments.

Let us note that CR raises important issues, issues that will not find a complete answer here. I will not discuss issues around the relevance and availability of RIs, for example. I will also not discuss conflicts emerging between RIs, nor the precedence, order, and

[4]Heather Douglas (2004) has argued persuasively that scientific practice and social processes are neither reducible one to the other nor completely uncoupled. This sentiment applies to all three modes of reliability.

weight of each RI. It is the richness and urgency of this problem that requires putting into practice demands for an account at least as complex as the one presented here. In this sense, CR does not provide, nor intend to provide, absolute assurances. Instead, CR aims to ensure that our best epistemic efforts are geared towards encouraging the reliability of ML. Little more can be expected given the fallibility and limitations of human cognition. This is particularly noticeable when it comes to complex socio-technological systems like ML. Taking note of these caveats, I now address RI_1, RI_2, and RI_3 in turn.

3.1 RI_1: Technical robustness of algorithms

The first three reliability indicators discussed by Durán and Formanek focus on the design, implementation, and tractability of computer systems. All three put forward defining criteria for assessing the utility value of computer systems and their outputs. Consider *validation* procedures. In automated diagnosis that utilizes ML, data is used to predict patient prognosis. This is done by comparing disease progression with clinical data from prior patients sharing the same endotype or phenotype (Myszczynska et al., 2020). This practice validates the synthetic data rendered by ML with empirical data collected via scientific methods (observation, experimentation, intervention, measurement, and others). The utility value of the ML system is then considered appropriate if validation standard are passed. On CR, validation increases our confidence in the reliability of an algorithm because it increases the algorithm's accuracy and reduces error. Understood as an RI, validation furnishes reason to believe in the reliability of ML.

Note that validation methods encompass different techniques and methods. They are not all appropriate for the same goals. For instance, validation techniques cannot be transferred to and from other disciplines without prior critical discussion. There must be agreement on how suitable the code and data are for the task at hand (Lorscheid et al., 2012; Fagiolo et al., 2007). This is part of the social practice of designing algorithms with specific scientific purposes in mind. Durán and Formanek's third indicator (i.e., a *history of (un)successful implementations*) affords and extends this interpretation. According to the authors, researchers utilize techniques that are "improved upon, reconfigured, and radically revised when the technology changes or a new method is envisaged" (2018, 661).

Although there is no standard procedure for choosing the best strategy, one can repurpose algorithms that have worked in the past. BenevolentAI, for example, makes use of *Best First Search*, an heuristic algorithm that guides BenevolentAI in searching for suitable drugs (Segler et al., 2018, 604). The past success of this heuristic algorithm for similar projects is further reason to believe in the reliability of BenevolentAI. Likewise, past failures must be, and typically are, avoided by competent programmers. Having said that, the history of computing is littered with cases of failed software that should have,

but was not, validated or verified. Therac-25 is one tragic case (Leveson and Turner, 1993). Verification and verification of algorithms, whether totally or partially, are always considered to support an algorithm's reliability.

3.2 RI_2: Computer-based scientific practice

Technical robustness facilitates justification in terms of increasing accuracy, predictive power, and errors minimizations. But it is silent on the adequacy of ML systems for scientific purposes. An account of reliable ML should be able to offer standards according to which we understand why ML use in a scientific context is warranted.

Wu and Zhang's automatic facial recognition system illustrates how accuracy does not exhaust reliability. As mentioned in section 2, the AUC-ROC measured 0.9540 predictive accuracy. Such a high predictive accuracy was enough to cement the researchers' trust in the reliability of the CNN. But there is no basis for so much optimism. Criminality is a socially construed concept. It depends on diverse and sometimes contradictory interpretations of the socio-economic basis of criminality, psychological studies of criminals, and laws that determine when and to what degree someone is considered a criminal. Without reference to some of these frameworks, the prediction – however accurate – lacks the grounds for legitimate scientific claims.

On CR, the adequacy of ML systems for scientific purposes cannot be exclusively assessed on high predictive accuracy. Science affords more than just measuring and classifying outputs. ML systems do not and cannot operate in isolation from the broader context of scientific undertakings. Yet, we need not only delve into traditional scientific practice but into scientific practice that grows with and depends on computer models. RI_2 is an attempt to capture the RIs connected to a larger body of scientific theories, beliefs, and practices within which ML is embedded.

3.2.1 Expert knowledge

Expert knowledge is an umbrella term that covers the myriad of background education, knowledge, activities, training, virtues, and skills of researchers that bring to bear a broad range of talents to the development of ML systems in scientific contexts. Understood as an RI, *expert knowledge* reports on the many ways in which scientific expertise, technical expertise, and general competencies can be implemented into ML.

To best understand this RI, we must look at its various functions. For starters, it puts forward an ML system's competencies, requirements, and theoretical assumptions as conceptualized by the researchers involved in the design, coding, and execution of the algorithm. It also accounts for the ability to describe a target system and its conditions for adequacy (e.g., to be applicable in a specific domain, to be representative of a particular

condition, to be context-sensitive, to be repurposed). Expert knowledge covers social practices tailored to the development of ML, aptitudes to anticipate its merits intelligibly, and abilities of agents to manipulate ML systems. Manipulability involves varying an ML's set of parameters, initial conditions, datasets, and parameters (epochs, batch size, number of neurons, etc.), all of which are complex yet critical for the ML performance and scientific merits. Consider determining which parameter to prioritize. Their selection and optimization is not trivial and yet fundamental for the general performance of the ML (convergence of results, accuracy, overall performance) (Hutter et al., 2014). van Rijn and Hutter (2018) have conducted an informative experiment to show that the final performance metrics for deep learning models varies according to how different researchers select and optimize algorithmic parameters and instantiations. Thus, experts contribute to the reliability of ML by designing relevant internal data-types, structures, relations, and operations. They also credit reliability (or might identify instances of unreliability) by selecting datasets, parameters, and other performance-related variables.

As a RI, expert knowledge also attempts to accommodate the complexities of ML through the division of cognitive labor. Rather than being developed in isolation, ML involves in the development stages a myriad of direct and indirect stakeholders (e.g., software engineers, physicians and chemists – in the case of BenevolentAI –, biologists – in the case of AlphaFold –, and psychologists and legal officers – in what should have been the case for Wu and Zhang –, just to mention a few). A core team designs algorithms utilizing ready-made computer modules others have coded. They employ measuring techniques others have designed, constructed, and calibrated. They analyze data using mathematical and statistical techniques others have validated. They use mathematical and computational methods others have devised and tested. There is no development of ML systems in solitude. Teams with diverse cognitive strengths and talents collectively collaborate in a variety of ways. The success or failure of ML is tailored to this collective knowledge, just as much as it depends on individual competencies. What one team member overlooks, another might notice. What one team member forgets, another might foresee. What one team member does not know how to solve, another might be able to teach. Thus diversified, the range of achievable solutions is far greater than what is available in atomized practices.

The role and value of experts is being increasingly recognized in philosophical studies on ML. Ratti and Graves (2022) argued that documenting developer's motives and the code and design of an ML are indicators of the reliability of the system. Newman (2016) has argued along similar lines. Newman considers the entire practice of software engineering to be at stake, from test plans to selecting programming languages and modeling tools, including configuration management.

While I am sympathetic to these ideas, my interpretation of expert knowledge is some-

what broader. It includes technical personnel with no training in software development, practices that exceed software engineering standards, and accommodates the possibility that complete documentation of an algorithm is not always available.

In practice, non-technical personnel are intimately involved in ML development (e.g., physicians and chemists in the case of BenevolentAI, and biologists and chemists in the case of AlphaFold), despite having little to no idea how key features of the system are designed and implemented. Their expertise is, however, crucial for the assessment of the reliability of the ML, and thus must be considered. Typically, their role is to inform, supervise, and test the design and coding of algorithms. But of course, the roles and interaction varies (Sundberg, 2010).

In connection with this, local practices and vernacular terminology often exceed what is captured by software engineering standards. Consider for instance how ML naturalizes or 'fossilizes' concepts. Once a concept is coded into the system, it is universally and indistinguishably applied across large and heterogeneous databases with varying degrees of success. Take the concept of 'health' as a case in point. One interpretation takes statistical measures of and standards of normal biological measurements of someone's body as the baseline for whether they are healthy. This concept of 'health' can be relatively straightforwardly implemented on an algorithm. However, the concept 'health' also allows interpretations tailored to the diverse values of an individual or a community (Richman, 2004). If a community considers blood transfusion to be harmful, they will treat any members of the community who have received a blood transfusion as unhealthy (Richman and Budson, 2000). Implementing a cogent definition of health is no trivial matter. However, software engineering prefers to adopt the technical perspective, one that favors the former definition over the latter because it is easier to implement and will lead to more accurate results.

Lastly, anyone who has written a piece of code knows that not everything is documented. And even so, there is no guarantee of understanding the code and its various functions. Thorough documentation – when it happens – and well-intended software engineering still falls short of capturing the methodological and epistemological competencies of ML. We need to highlight the subtle interpretations, gentle disagreements, and non-verbal practices pervading computational and scientific practice and which make their way into the algorithm.

Let me finish by noticing that this RI brings about another important aspect of ML systems: they might only be *locally* reliable. The idiosyncrasies attached to documenting, designing, coding, executing, and maintaining ML systems makes them reliable in one context but not necessarily in another. This is, I believe, at the root of IBM's Watson for Oncology's difficulties of implementation in South Korea and Denmark, despite its success in the US market (Vulsteke et al., 2018; Emani et al., 2022). Notoriously, Watson

for Oncology was capable of analysing large amounts of data and multiple variables, rendering accurate diagnoses and treatments for cancer patients in the US. But while IBM presents Watson for Oncology as offering more objective medical decisions and more accurate diagnoses than actual oncologists (Swetlitz, 2016), it has been reported that many of these claims have been aggrandized (Ross and Swetlitz, 2017, 2018). When implemented in South Korea and Denmark, only a fraction of the outputs rendered by the algorithm matched – or closely matched – the local clinician's best diagnosis (Hamilton et al., 2019).

3.2.2 Knowledge-based integration

Scientific results do not come in discrete bits, nor are the objects of scientific inquiry individual facts independently sanctioned. Instead, scientific practice constitutes a web of mutually supportive claims and commitments that are reached after complex negotiations in complex socio-economic and political environments. However, many studies utilizing ML portray a sanitized image of scientific research, where there is privileged access to structured data, undisputed model implementation, and meaningful representations of the world.

Wu and Zhang, for example, state that the quality of their databases and the methods implemented for data analysis prevent "the garbage of human biases from creeping in" (2017, 2). They state that "like most technologies, machine learning is neutral," (Wu and Zhang, 2017, 2). Moreover, given "race, gender and age, the faces of [the] general law-biding public have a greater degree of resemblance compared with the faces of criminals" (Wu and Zhang, 2017, 2). It is however doubtful whether Wu and Zhang's CNN has any scientific value. One reason (to add to those from previous sections) is that Wu and Zhang's CNN is largely disconnected from accepted bodies of scientific knowledge. More precisely, the properties their CNN purports to refer to – 'criminal' and 'non-criminal' – are posited in isolation from established evidence, models of criminal psychology, social studies on crime, and the relevant theories on criminality. Wu and Zhang's conclusion is premised solely on the data analysis methods implemented in the CNN and the databases used.

I will call approaches that conceive of ML as disconnected from the larger body of scientific knowledge *just a bunch of data analysis* (JBDA). My aim is not to target practices utilizing ML for plain data analysis. Some forms of research may find their value in doing so. Much good scientific work even depends on classifications and clustering emerging from data analysis. My aim is rather to highlight the fact that, in some cases, outputs of ML are mistakenly taken at face value. This ignores the 'bigger picture' of knowledge integration, interpretation, and application. JBDA approaches misleadingly portray ML

as an objective, unambiguous, and scientifically-grounded examination of the data that produces outputs whose contents meaningfully represent the world. Wu and Zhang's CNN approach is an archetypal JBDA, as it depicts scientific practice as granular, consisting of discrete pieces of information, separately secured and individually sanctioned. JBDA approaches put forward a form of scientific practice that is non-perspectival, socially disinterested, impartial (i.e., epistemically and normatively neutral), and disembodied from larger corpus of scientific knowledge.

Are there instances where JBDA approaches are scientifically intelligible? I believe so. As suggested, there are indeed cases where mere data analysis renders valuable scientific information about a subject matter. But for such cases, one needs to provide further justification that relates JBDA with a larger corpus of knowledge. A plausible interpretation of an account of scientific practice with ML capable of accommodating cases of JBDA takes the bulk of scientific knowledge and practices in the field under study as background knowledge and as affording sufficient grounds to underwrite particular claims made with the ML. To illustrate this idea, consider BenevolentAI*, an ML system whose working principles are JBDA. Suppose that BenevolentAI* puts forward baricitinib* as a drug with high chances of combating COVID-19 symptoms. Would a team of researchers endorse baricitinib* when it was the outcome of implementing JBDA? Surely not. Would they have reasons to accept it at face value? To my mind, this would be epistemically irresponsible. BenevolentAI*'s output (baricitinib*) must be embedded in a larger body of knowledge about COVID-19 before anyone is justified in accepting or rejecting its output.

What the examples of BenevolentAI and BenevolentAI* show is that we have justification for believing that either output has scientific value if (a) the algorithm – as a whole or as constituent parts – implements scientific models, theories, principles, categories, and/or other elements purposed in our corpus of scientific knowledge, or (b) the outputs are assessed within our corpus of scientific knowledge.

BenevolentAI utilizes information gathered from scientific research papers, structured and unstructured biomedical data, and drug and pharmaceutical industry data. It also implements knowledge graphs that structure data into relationships between known diseases, genes, and approved drugs (Smith et al., 2021). BenevolentAI is also in agreement with auxiliary theories and structures of drug molecular profiles, and the mechanisms involved in damaging healthy cells and tissues. BenevolentAI also implements theories about genetics, medical studies of disease, and biological models relating genes to drug effects. Conversely, BenevolentAI* does not implement any accepted model or theory into its algorithm. It nonetheless classifies baricitinib* as having high probabilities of successfully combating COVID-19 symptoms. It can later be shown that this output is in agreement with background knowledge about inhibiting AAK1 and JAK 1/2 signaling pathways essential for combating the symptoms of COVID-19.

Embedding scientific knowledge into algorithms and embedding outputs within the larger corpus of knowledge are RIs that provide different modes of justification. The former observes that the relevant scientific structures and processes, commitment and categories, entities and concepts are implemented into the algorithm; the latter, that the output is in agreement with accepted scientific commitments, standards of quality, evidence, and relevance, among other scientific qualifications. Consider Wu and Zhang's CNN once more. One could claim justification that a photo is of a 'criminal' when, for instance, either (a) the algorithm implements specific structures and categories (e.g., definitions of criminality, instances of conviction, lawful classifications of crime), or (b) the output can be assessed with current theories of criminality (social theories, psychological theories). Unfortunately, neither strategy is applied by Wu and Zhang, nor can it be applied. The way in which the algorithm classifies a 'criminal' does not involve considerations of any accepted theory or model of criminality (if anything, the algorithm implements a discredited theory like phrenology), nor is the output reviewed in light of scientific evidence about criminality. Under scrutiny, Wu and Zhang's CNN is a non-starter for scientific purposes.

Note that, if an ML output is embedded in larger corpus of scientific knowledge, this does not mean it is free from further concerns and assessments. Favalli and colleagues (2020), for example, have reported potential harms involved in administering baricitinib to some patients. They found that the mechanisms of action of baricitinib block JAK-STAT signaling pathways (mainly mediated by JAK1 and JAK2). This impairs interferon-mediated antiviral responses (Favalli et al., 2020, 1013). Interferon is one of the most powerful immune responses to counteracting viral replication (especially during the early phases of the infection). Blocking interferon with baricitinib allows attack by other viruses (e.g., herpes zoster and herpes simplex) which may actually be more harmful than COVID-19.

There are ongoing investigations into the implications of administering baricitinib to a wide range of patients. Peter Richardson, in fact, does "not recommend that baricitinib or other JAK inhibitors be given to these individuals [immunocompromised patients]" (Favalli et al., 2020, 1013). The findings of BenevolentAI and further laboratory testing, however, "suggests that when hospital care is required for patients with a pathogenic SARSCoV2 infection, JAK–STAT pathway inhibition might be a potential strategy" (Favalli et al., 2020, 1014). Perhaps a future patched version of BenevolentAI will include considerations of JAK inhibitors. In any event, for now, BenevolentAI seems to implement suitable models, metrics, and data. It gives Favalli, Richardson, and the rest of the scientific community justification for believing that baricitinib is a scientifically legitimate output for the intended purposes of combating COVID-19 symptoms.

3.3 RI_3: Social construction of reliability

On RI_1 and RI_2, technical, theoretical, and empirical indicators provide justifications for belief. Wu and Zhang's system cashes in on a subset of RI_1 (i.e., validation). It puts forward claims about a supposed law of facial features as a scientific product. When assessed against RI_2, in particular knowledge-based integration, Wu and Zhang's CNN is flagged as unreliable. Conversely, baricitinib cannot be justified solely on assessing BenevolentAI's validity and knowledge integration. That would expose us to accepting the outputs of algorithms without further debate. For this, baricitinib must be exposed to general and rigorous scientific debate to determine its acceptance as a scientific product. Thus, reliable ML also depend on social processes that aim to achieve standards of scientific evidence and thresholds for acceptance.

I mentioned the debate within the scientific community that followed the announcement of BenevolentAI's output. Favalli et al. (2020) reported on the potential harms of administering baricitinib to some patients (e.g., herpes zoster and herpes simplex infection). Exchanging views with Richardson's team led Favalli et al. to reconsider their view. As an RI, the debate determined how much evidence was required to accept BenevolentAI's outputs. It determined which errors and artefacts can be tolerated, and to what extent. It also determined which assumptions are fit for purpose. Commitments to reliable ML are commitments to a network of scientific methodologies, standards, and traditions expressed through scientific debate. As Catherine Elgin points out, this network enables scientists to build on each another's work. They can be confident that justified outputs have the epistemic value their discipline prescribes (Elgin, 1996, 77). Of course, disputes and disagreements among community members are to be expected. There may be conflicts over values, methods, and what constitutes acceptable evidence. Take again Favalli's concerns about administrating baricitinib to a specific group of patients. Whereas Richardson largely agreed with Favalli's concerns, research on the drug continued. Further laboratory testing confirmed Richardson's beliefs.

The social formation and justification of beliefs is a complex enterprise that is not always successful. There are situated background assumptions and perspectives. Scientific inquiry is permeated by contextual values and interests. Many concepts built into ML are socially constructed, generational, and discipline-idiosyncratic. Take again variations in the definition of 'health', and 'disease' (Boorse, 2011; Sisti and Caplan, 2017). Each concept operates under a myriad of cultural, political, economic, and moral values. Caruana and colleagues (2015) discuss a neural network designed to predict pneumonia risk scores in patients and their readmission to hospital. Caruana et al. find that asthmatic patients are at low risk, and thus less likely to require hospitalization than other patients (with chronic lung disease, for instance). Caruana et al.'s finding are statistically accurate (RI_1

and RI_2 indicators are present). However, the system is perceived as unreliable given that physicians have different starting assumptions about what a predictive algorithm should provide. According to Theunissen and Browning (2022) physicians assume that ML is predicting outcomes according to a shared baseline of care rather than differential care relative to the background of the patient. The ML outputs can be called into question because of oversimplifications in one or more assumptions in the system. This further shows that neither RI_1 nor RI_2 are individually – or jointly – sufficient for the reliability of ML.

CR makes an effort to foster belief-forming methods that accommodate social interventions, scientific scrutiny, and inter-domain justifications. Beliefs are justified in relation to a network of interconnected scientific beliefs. Yet, we cannot expect it to be recurrently successful. Contingent values and interests within the scientific community find their way into the design, use, and maintenance of ML. As with other scientific methodologies, ML reliability requires a delicate balancing act. We bring together our best technical knowledge, theories, methodologies, and social practices. We do so in an attempt to justify believing that the output of a given ML system can be scientifically valuable.

4 Final remarks

I have presented CR as a justificatory framework for ML outputs. I also presented three families of RIs that, to my mind, are capable of crediting reliability to ML. In sum, I set out to defend the following claim: we have – or increasingly have – justification for believing the outputs of ML as having scientific value if and when they are produced by reliable belief-forming methods.

Admittedly, CR construes justification as inherently provisional. As discussed, reliability indicators might change over time and even be replaced. But this is part of self-critical and self-correcting scientific endeavors, now deeply substantiated with ML. It is also the best we can do in a given context with limited resources. These activities constitute our best knowledge and best practices, even if subject to further scrutiny and revision.

In the introduction, I listed a few issues that this paper is unable to address. Nonetheless, I think that CR is a step in the right direction, and in that regard this paper has then achieved its goal. Let me finalize by briefly addressing one further concern about CR, that is, being an all-encompassing epistemology. Whereas the RIs might seem that there is nothing left to consider for the reliability of ML, we need to note that there are still plenty of practices in connection with designing and coding algorithms that are irrelevant to CR. For instance, the portability of algorithms understood as adapting the algorithm for various hardware-operating system combinations has no bearing on the algorithm's

reliability. Another example is some forms of testing, such as metrics for walkthroughs of the algorithm (Schach, 2007, 152-154). Walkthroughs consist of four to six individuals documenting any faults of the system (e.g., failure to follow specifications, rounding-off errors, etc.) without correcting them. While documenting and correcting algorithm faults do contribute to their robustness – and therefore reliability –, the specific methodology of the walkthrough is irrelevant for CR.

5 Acknowledgments

Many people have helped me, directly or indirectly, shape these ideas. I would like to thank Axel Barceló, Atocha Aliseda, Karen González Fernández, Charles Rathkopf, Emanuelle Ratti, and Nico Formanek for their comments. Special thanks go to Federica Russo and Giorgia Pozzi. Both have been exceptionally supportive of these ideas. Finally, thanks go to the participants of the Workshop series issues in XAI §4 and §5.

References

Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, 187–194. ACM Press/Addison-Wesley Publishing Co.

Boorse, C. (2011). Concepts of Health and Disease. In F. Gifford (ed.), *Handbook of thePhilosophy of Science*, 13–64. Elsevier.

Caruana, R., Lou, Y. et al. (2015). Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 1721–1730. Association for Computing Machinery. doi:10.1145/2783258.2788613.

Clark, A. (1987). The kludge in the machine. *Mind and Language*, 2(4):277–300.

Creel, K. A. (2020). Transparency in Complex Computational Systems. *Philosophy of Science*, 87(4):568–589. doi:10.1086/709729.

Douglas, H. (2004). The irreducible complexity of objectivity. *Synthese*, 138:453–473.

Durán, J. M. and Formanek, N. (2018). Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism. *Minds and Machines*, 28(4):645–666.

Elgin, C. Z. (1996). *Considered Judgement*. Princeton University Press.

Emani, S., Rui, A. et al. (2022). Physicians' Perceptions of and Satisfaction With Artificial Intelligence in Cancer Treatment: A Clinical Decision Support System Experience and Implications for Low-Middle-Income Countries. *JMIR cancer*, 8(2):e31461–e31461. doi: 10.2196/31461.
URL: *https://pubmed.ncbi.nlm.nih.gov/35389353*

Fagiolo, G., Moneta, A. et al. (2007). A Critical Guide to Empirical Validation of Agent-Based Models in Economics: Methodologies, Procedures, and Open Problems. *Computational Economics*, 30:195–226.

Favalli, E. G., Biggioggero, M. et al. (2020). Baricitinib for COVID-19: a suitable treatment? *The Lancet*, 20:1012–1013.

Goldman, A. I. (2012). *Reliabilism and Contemporary Epistemology*. Oxford University-Press.

Hamilton, J. G., Genoff Garzon, M. et al. (2019). "A tool, not a crutch": patient perspectives about IBM Watson for oncology trained by Memorial Sloan Kettering. *Journal of Oncology Practice*, 15(4):e277–e288.

Humphreys, P. W. (2009). The Philosophical Novelty of Computer Simulation Methods. *Synthese*, 169(3):615–626.

Hutter, F., Hoos, H. et al. (2014). An efficient approach for assessing hyperparameter importance. In E. P. Xing and T. Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, vol. 32(1) of *PMLR*, 754–762.

Jumper, J., Evans, R. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589.

Leveson, N. and Turner, C. (1993). An investigation of the Therac-25 accidents. *Computer*, 26(7):18–41.

Lorscheid, I., Heine, B.-O. et al. (2012). Opening the 'black box' of simulations: increased transparency and effective communication through the systematic design of experiments. *Computational and Mathematical Organization Theory*, 18:22–62.

Massimi, M. and McCoy, C. D. (eds.) (2020). *Understanding Perspectivism. Scientific Challenges and Methodological Prospects*. Routledge.

Medeiros, J. (2021). How tech is changing healthcare. From rapid development and rollout of the Covid-19 vaccines to the science of isolation, machine-learning-enabled gene editing and digitised medicine. *Wired*.
URL: *https://www.wired.co.uk/article/future-health-trends*

Morales-Hernández, A., Van Nieuwenhuyse, I. et al. (2022). A survey on multi-objective hyperparameter optimization algorithms for machine learning. *Artificial Intelligence Review*. doi:10.1007/s10462-022-10359-2.
URL: *https://doi.org/10.1007/s10462-022-10359-2*

Myszczynska, M., Ojamies, P. et al. (2020). Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature Reviews Neurology*, 16:440–456. doi:10.1038/s41582-020-0377-8.

Newman, J. (2016). Epistemic Opacity, Confirmation Holism and Technical Debt: Computer Simulation in the Light of Empirical Software Engineering. In F. Gadducci and M. Tavosanis (eds.), *History and Philosophy of Computing. HaPoC 2015. IFIP Advances in Information and Communication Technology*, vol. 487, 256–272. Springer. doi:10.1007/978-3-319-47286-7_18.

Ratti, E. and Graves, M. (2022). Explainable machine learning practices: opening another black box for reliable medical AI. *AI and Ethics*. doi:10.1007/s43681-022-00141-z.

Ribeiro, M. T., Singh, S. et al. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Richardson, P., Griffin, I. et al. (2020). Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. *The Lancet*, 395(10223):e30–e31. doi:10.1016/S0140-6736(20)30304-4.

Richman, K. A. (2004). *Ethics and the Metaphysics of Medicine. Reflections on Health and Beneficence*. MIT Press.

Richman, K. A. and Budson, A. E. (2000). Health of organisms and health of persons: Anembedded instrumentalist approach. *Theoretical Medicine and Bioethics*, 21(4):339–352.

Ross, C. and Swetlitz, I. (2017). IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. *Statnews*.
URL: *https://www.statnews.com/2017/09/05/watson-ibm-cancer/. (Accessed 25 April 2020).*

——— (2018). IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. *Statnews (2018)*.
URL: *ttps://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf. (Accessed 25April 2020).*

Schach, S. R. (2007). *Object-Oriented & Classical Software Engineering*. McGraw-Hill.

Segler, M. H. S., Preuss, M. et al. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555(7698):604–610. doi:10.1038/nature25978.
URL: *https://doi.org/10.1038/nature25978*

Sisti, D. and Caplan, A. L. (2017). The concept of disease. In M. Solomon, J. R. Simon and H. Kincaid (eds.), *The Routledge Comanion to Philosophy of Medicine*, 5–15. Routledge.

Smith, D. P., Oechsle, O. et al. (2021). Expert-Augmented Computational Drug Repurposing Identified Baricitinib as a Treatment for COVID-19. *Frontiers in Pharmacology*, 12. doi:10.3389/fphar.2021.709856.

Sundberg, M. (2010). Cultures of simulations vs. cultures of calculations? The development of simulation practices in meteorology and astrophysics. *Studies in History and Philosophy of Modern Physics*, 273–281.

Swetlitz, I. (2016). atson goes to Asia: hospitals use supercomputer for cancer treatment. *Statnews*.
URL: *https://www.statnews.com/2016/08/19/ibm-watson-cancer-asia/*

Theunissen, M. and Browning, J. (2022). Putting explainable AI in context: institutional explanations for medical AI. *Ethics and Information Technology*, 24(2):23. doi:10.1007/s10676-022-09649-8.

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuro-science*, 3(1):71–86.

van Rijn, J. N. and Hutter, F. (2018). Hyperparameter Importance Across Datasets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2367–2376. ACM.

Vulsteke, C., Arevalo, M. O. et al. (2018). Artificial intelligence for the oncologist:hype, hubris, or reality? *Belgian Journal of Medical Oncology*, 12(7):330–333.

Wachter, S., Mittelstadt, B. et al. (2018). Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal of Law and Technology*, 31(2):841–887.

Wu, X. and Zhang, X. (2016). Automated Inference on Criminality using Face Images. doi:https://arxiv.org/pdf/1611.04135v1.pdf.

——— (2017). Responses to Critiques on Machine Learning of Criminality Perceptions (Addendum of arXiv:1611.04135). doi:10.48550/arXiv.1611.04135.