Chapter 4
Beyond Transparency: Computational Reliabilism as an Externalist Epistemology of Algorithms

Juan M. Durán

Abstract This chapter examines the epistemology of algorithms, framing the discussion as a question of epistemic justification. Current approaches emphasize algorithmic transparency, which involves elucidating internal mechanisms—such as functions and variables—and demonstrating how (or that) these compute outputs. Thus, the mode of justification through transparency is contingent on what can be shown about the algorithm and, in this sense, is *internal* to the algorithm. In contrast, I propose an *externalist* epistemology of algorithms called *computational reliabilism* (CR). While I have previously developed CR in the context of computer simulations (Durán, Explaining simulated phenomena: A defense of the epistemic power of computer simulations, 2013; Durán, Computer simulations in science and engineering. Concepts - practices - perspectives. Springer, 2018; Durán, Formanek, Minds and Machines 28(4), 645–666, 2018), this chapter extends the framework to a broader range of algorithms used across scientific disciplines, particularly in machine learning and deep neural networks. At its core, CR posits that an algorithm's output is justified if it is generated by a reliable algorithm, where reliability is determined by reliability indicators. These indicators arise from formal methods, algorithmic metrics, expert competencies, research cultures, and other scientific practices. The chapter's primary objectives are to delineate the foundations of CR, explain its operational mechanisms, and outline its potential as an externalist epistemology of algorithms.

4.1 Introduction

The use of algorithms for scientific purposes is delivering remarkable results. A couple of examples will suffice to illustrate this. In molecular biology, AlphaFold can predict protein structures with atomic accuracy in cases where no similar struc-

J. M. Durán (✉)
Department of Values, Technology and Innovation, Faculty of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands
e-mail: j.m.duran@tudelft.nl

J. M. Durán, G. Pozzi (eds.), *Philosophy of Science for Machine Learning*, Synthese Library 527, https://doi.org/10.1007/978-3-032-03083-2_4

tures are known (Jumper et al., 2021). In medicine, BenevolentAI has combined structured and unstructured biomedical data sources to identify rheumatoid arthritis drugs like *baricitinib* as therapeutic for COVID-19 symptoms (Medeiros, 2021). In an increasingly number of cases, algorithms have successfully extended the class of tractable chemistry, biology, physics, and medicine, broadening the range of modeling and experimental capabilities available to researchers.

Yet, unlike other methods, the algorithm's scientific merits cannot be easily determined by association with a body of scientific knowledge, by adequacy to empirical data, or by diverse theoretical constructs—such as explanation and observation. This is for a variety of reasons. Algorithms are epistemically and methodologically opaque,[1] making it difficult to associate a given algorithm and its output with the general scientific canon. Likewise, empirical phenomena are often temporarily, spatially, or cognitively inaccessible for validation of the algorithm, potentially casting doubts over any representational value of these systems.

When confronted with these issues, philosophers and computer scientists gravitate towards *transparency*, an umbrella term capturing diverse methods linking the internal mechanisms and properties of algorithms to their outputs (Creel, 2020; Wachter et al., 2018; Ribeiro et al., 2016). To see how transparency works, consider BenevolentAI. At its core, this algorithm is a search engine that combines structured and unstructured biomedical data sources, drug industry data, and automated retrieval of information from diverse 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 diverse 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). To justify Richardson's belief that the output has scientific value, proponents of transparency would emphasize showing how baricitinib is computed through procedures that integrate biomedical data, instantiate key variables, invoke function calls that identify structural relationships within the algorithm, employ relevant conditional statements, and execute other algorithmic operations. Another way to make BenevolentAI transparent is via a *knowledge graph* (Richardson et al., 2020). This visualizes how baricitinib inhibits AAK1 (associated with interrupting the COVID-19 virus' passage into cells) and JAK 1/2 (critical for signal transduction pathways), and how baricitinib binds with GAK (known 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

[1] The philosophical literature on opacity—particularly in the context of computational systems—is too extensive to cite exhaustively. Key contributions include Humphreys (2004), Humphreys (2009), Burrell (2016), Alvarado and Humphreys (2017), Durán and Formanek (2018), Beisbart (2021), Boge (2022), Beisbart (2025).

(by binding with GAK) nor inhibit cytokine signaling (by inhibiting JAK 1/2) (Richardson et al., 2020).

Are Richardson and his team justified in believing that baricitinib is a medically valid outcome for the issue at hand? What reasons do they have to discard other drugs as either less effective or unsafe? What supports their claim that BenevolentAI is a reliable system for the intended purposes? These are questions about the epistemic reliance on algorithms 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 aimed at the transparency of the algorithm. More specifically, this paper lays the groundwork for *computational reliabilism* (CR), a reliabilist epistemology centered on algorithms, aimed at justifying their outputs.

As the name suggests, CR borrows bits and pieces from epistemological reliabilism, notably Goldman's process reliabilism (Goldman, 1979, 2012). However, the version of CR I develop here draws from, but also expands on, my previous work on computer reliabilism for computer simulations (Durán, 2013, 2018; Durán & Formanek, 2018; Durán & Jongsma, 2021). With these ideas in mind, this chapter is divided as follows. Section 4.2 presents and discusses two epistemologies of algorithms: one that is internal to the algorithm (e.g., transparency) and one that is external to the algorithm (i.e., CR). As expected, these epistemologies have different modes of justification, which are exemplified in Sect. 4.2.1. The example provided is only intended to motivate CR. In Sect. 4.3, I lay the groundwork for *computational reliabilism* (CR). Here, I present three types of *reliability indicators* (type-RIs) that credit reliability to algorithms. These are (1) type$_1$-RI technical performance of algorithms (Sect. 4.3.1.1), (2) type$_2$-RI computer-based scientific practice (Sect. 4.3.1.2), and (3) type$_3$-RI social construction of reliability (Sect. 4.3.1.3). In Sect. 4.4, I briefly take stock of 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 the algorithms' outputs? My answer is that reliability comes through myriad methods, practices, and processes at diverse stages of specification, coding, use, and maintenance of the algorithms.

4.2 Internalist and Externalist Epistemologies for Algorithms

A central motivation for seeking justification is that algorithms are often epistemically opaque. This concept has two distinct but related interpretations. The first interpretation addresses how algorithms involve multiple complex elements (functions, variables, decisions, data, etc.) in their specification, 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 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

sometimes are from human insight. The second interpretation sets the focus on our limited capacities to say something meaningful about the output of an algorithm (Humphreys, 2009, p. 649). That is, no human being (or group of human beings) can know which functions, variables, decisions, data, etc. are relevant to a given output.

Whereas the first interpretation focuses on the algorithm as an opaque method, the second highlights our cognitive, epistemic, and other limitations in making knowledge claims about the algorithm's output. Both interpretations, I believe, can be cast as human agents lacking proper justification for the algorithm's output. Here justification is taken to be epistemic, and understood in the general sense of a belief being formed in the proper manner. Thus, either because the algorithm is a black-box or because human agents are cognitively limited, there is no basis for claims about the proper formation of a belief about whether the algorithm's output is true, has scientific value, or represents a fact in the world (Humphreys, 2020).[2]

Under this heading, transparency emerges as a promising epistemology of algorithms. It begins with uncovering the algorithm's internal mechanisms and properties—such as its functions, variables, and implemented metrics—and proceeds by linking these to its outputs. Justification, then, is understood as the provision of reasons or supporting evidence by the internal workings of the algorithm for believing that the output is true, has scientific value, or accurately represents a fact about the world (Creel, 2020; Zednik, 2021; Burrell, 2016).[3] Recall from the introduction that BenevolentAI employs *knowledge graphs* to visualize how the algorithm prioritizes baricitinib over alternative drugs. On a transparency-based account, access to the functioning of the knowledge graph justifies the belief that baricitinib is an appropriate candidate for combating COVID-19.

Another example of a transparency-conferring mechanism is LIME: a general algorithm that explains the predictions of any classifier by locally approximating

[2] For simplicity and continuity with the literature on the tripartite view of knowledge, I distinguish between the belief that the algorithm's output ō is true—or false—scientifically valid—or sound—or that it represents—or misrepresents—from *S* having a propositional attitude toward that belief. In the former case, I do not intend to defend either a realist or an anti-realist position. I take this to be an important debate that philosophers interested in algorithms must eventually confront but this is not the place for it—see Chap. 3 by Casey in this volume. Thus, for *S* to believe that the algorithm's output ō is "true" is for *S* to have a propositional attitude toward the claim that ō is true, has scientific value, represents a fact in the world. Following Elgin (2017), the latter sense of truth is not understood as (absolute) correspondence with reality, but rather as sufficient adequacy for the purposes at hand—an adequacy that enables practical engagement, scientific progress, and understanding (see also Parker, 2020). I am grateful to Jack Casey for his close reading on this attempt to clarify the issue. Likewise, the claim that a human agent *S* holds the belief that ō is "true" constitutes a distinct and higher-order belief. In this respect, it is important to note that beliefs, in this sense, can be occurrent or dispositional, subject to revision or revocation, and that their acceptance may vary in degrees of strength and confidence. Beliefs are also taken to be historically situated, incremental, and perspectival (Massimi & McCoy, 2020).

[3] Like others–for example, Räz and Beisbart (2022)—I take at least some of Sullivan's work to be internalist, despite its strong references to externalist approaches (e.g., Sullivan, 2022). It is less controversial, however, that Sullivan endorses an evidence-based epistemology.

it with an interpretable model. Formally, LIME computes a model $g \in G$, where G is a class of potentially interpretable models (e.g., linear models, decision trees, falling rule lists). In practice, if an algorithm 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 LIME. Indeed, 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). Let us note in passing that many forms of explanatory AI (XAI) provide a rich source for transparency, as they often involve tracking back the *path-dependency* of the algorithm that relates a given function (or set of functions), variables, etc., to its output (Durán, 2021).

Thus understood, transparency purports justification as an *internal to the algorithm* matter. That is, the justification of our beliefs that the output is true depends exclusively on some form of surveying the inner workings of the algorithm. To put this idea more or less formally,

Definition 1 A human agent S is justified in believing the algorithm's output $\bar{o}$ just in case: (a) it is shown, directly or indirectly, the algorithmic path-dependency to $\bar{o}$; and (b) S has reasons or supporting to believe that the path-dependency to $\bar{o}$ are the case.[4]

Opposing this view is computational reliabilism (CR), presented here as an *external to the algorithm* epistemology. It consists in identifying formal methods, algorithmic metrics, expert competencies, and research cultures that constitute our best epistemic and normative efforts to specify, code, use, and maintain reliable algorithms. I refer to these as *reliability indicators* (RIs), which, as I will explain in Sect. 4.3, can be divided into *types* and *tokens*.

By construction, then, CR does not depend on revealing the internal mechanisms and properties of the algorithm. Instead, it relies on reliability indicators that are external to it, including the algorithm's socio-technical dimensions. For this reason, CR aspires to be an epistemology of *algorithmic systems*, rather than one limited to program-scripts or program-processes—that is, to algorithms in the narrow sense (Eden, 2007).[5] For simplicity, however, I will occasionally speak of algorithms. A provisional working definition might be stated as follows:

Definition 2 A human agent S is justified in believing the algorithm's output $\bar{o}$ if and only if $\bar{o}$ was generated by a reliable algorithmic system. A reliable algorithmic

[4] There is a more nuanced—and arguably more scientifically accurate—approach to transparency, namely, *contextual transparency*. The underlying idea is that transparency does not arise solely from the inner workings of the algorithm, but also from "knowledge of the environmental patterns and regularities that are being tracked and of the abstract representational structures that are tracking them" (Zednik, 2021, p. 268). A similar view is expressed in Burrell (2016). In this form of contextual transparency, reasons or supporting evidence are both internal and external to the algorithm.

[5] Thanks are due to Emma-Jane Spencer for pressing me to clarify the scope of CR.

system is one that produces outputs ō that are true—scientifically valid, represent—most of the time. To achieve this, the algorithm must have been specified, coded, implemented, and maintained in accordance with a range of reliability indicators.

While I will dedicate a large portion of this chapter to the characterization of type and token reliability indicators, a preliminary conclusion can be drawn now: we can say that transparency and CR have different justificatory modes. According to the former, we have justification by having access to the inner workings of the algorithm. According to the latter, we have justification by identifying methods (formal and otherwise), metrics, expert competencies, cultures of research, and the like external to the algorithm that make up our best epistemic and normative efforts to increase the algorithm's reliability.

As a final attempt to illustrate these two justificatory modes, let me briefly present and discuss an example of an algorithm that classifies individual suspects as {criminal; non-criminal} based on their facial traits. The idea is to show that transparency may justify the belief that a given suspect is a criminal—or not—whereas CR flags the algorithm as unreliable and therefore lacking the epistemic merit required for justification. It goes without saying that this example is only meant to contrast these two justificatory modes. No conclusions regarding the relative value of these two epistemologies are intended to be drawn from this.

4.2.1 Merchants of Unreliability

In 2016, computer scientists Wu and Zhang developed a Convolutional Neural Network (CNN) that analyzed over 1850 ID photos and classified them as {criminal; non-criminal}.[6] About 1120 of these photos were of people with no criminal convictions, and the remaining 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) and classifies each photo as {criminal} or {non-criminal.}. No other concept or category was operational. Despite this—or perhaps because of this—the predictive accuracy measured using the Area Under the Receiver Operator Characteristic Curve (AUC-ROC) was very impressive: Wu and Zhang measured 0.9540 accuracy in the classifications. This means that the CNN was able to successfully classify faces of individuals as being {criminal} or {non-criminal} approximately 95% of the time (Durán et al., 2024; Wu and Zhang, 2016).

To further validate their algorithm and rule out that such a high predictive accuracy resulted from overfitting, Wu and Zhang retrained the 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

[6] It is worth noticing that Wu and Zhang's use of photos from actual people, in contrasts with other approaches that use synthetically generated photos (Turk & Pentland, 1991; Blanz & Vetter, 1999).

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 unbalanced datasets, choice of photos (light, angle, over and under exposure, 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, as to many, high predictive accuracy means that the algorithm's outputs represent, have scientific value, are true, etc. It is thus no coincidence that they confidently announce the "law of normality for faces of non-criminals" (Alston, 1995; Wu & Zhang, 2016). But high predictive accuracy is no standard for claims about scientific value or truth. One could argue that while the Ptolemaic model exhibited high predictive accuracy in its measurements, the model fundamentally misconceived and misconstrued planetary motion. Additionally, it is well known that high accuracy can be manufactured by carefully selecting input data and calibrating the algorithm's variables and functions to achieve a desired degree of predictability. 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. Now, algorithms allow for multiple optimal hyperparameter configurations, depending on the dataset, the algorithm's intended purpose, and the specific task (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 & Hutter, 2018). As a 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.

Against all epistemic and moral odds, Wu and Zhang insist that high predictive accuracy confers scientific value on their CNN and their outputs. To further defend this, they retrained the parameters of every layer in the CNN while also modifying the architecture (Hao & Stray, 2019; Wu & Zhang, 2017). As a result, the output maintained the same level of accuracy relative to previous executions of the algorithm. In fact, the CNN correctly picks out specific facial attributes from photos, and then classifies them into the appropriate category ca. 95% of the time.[7] But again, 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 output with the justification required for believing that output.[8]

[7] 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 datasets.

[8] What is operating here is the distinction between *output accuracy*, which is concerned with the correctness of the final computed outputs, and *procedural accuracy*, which is concerned with the execution of steps and adherence to methods.

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 justification. 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 inner functions and properties of the CNN responsible for the output will support the justification. But justifying the CNN's output using the same functions used to compute them is epistemically circular and inadmissible for the proper formation of beliefs. Rather, these functions only speaks of the algorithm's robustness, and only in a very limited way. 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. In other words, transparency here does not help to distinguish what we are compelled to believe from what cements that belief (Durán & Jongsma, 2021).[9]

4.3 Computational Reliabilism (CR)

Claims about justification find a home in CR, a branch of process reliabilism in which a subject *S* is justified in believing output ō if the algorithmic system is reliable (see Definition 4.2 on page 60) (Goldman, 2012; Durán et al., 2024). An algorithmic system is reliable when it generates ō that are true—have scientific merit, represent— rather than false—lack scientific merit, misrepresent—in most cases. Importantly, reliability here is not merely a matter of track record, but concerns the algorithm's propensity to generate true ō across an appropriate range of cases. The debate in epistemology over the most suitable form of reliabilism is extensive and cannot be addressed here. Suffice it to say that I favor *propensity reliabilism* over Goldman's *frequentist reliabilism*, aligning with Alston, who writes: "A reliable [method] is one that *would* usually deliver favorable results over an appropriate range of cases *if and when* they occur" (Alston, 1995).[10]

I will not pursue this point further, as the distinction between relative frequency and propensity reliabilism is unproblematic for the purposes of this chapter. The question, rather, is how to confer reliability to an algorithmic system. To this end, CR employs *reliability indicators* (RIs) as markers of methodological,

[9] 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 means of this, one could in principle identify high-level features that refer to domain knowledge (e.g., a list of criminality-based characteristics) and thus have reasons or supporting evidence for claims about justification.

[10] Humphreys defended a version of Goldman's process reliabilism (Humphreys, xxxx, 2020). For a recent critique of reliabilism in the context of algorithms, see Duede (2022) and Chap. 5 by Alvarado in this volume.

cognitive, social, normative, and epistemological competence (see also Russo et al., 2024; Pozzi & Durán, 2025). RIs are algorithm-related methods, metrics, practices, domain-specific knowledge, and related competencies that either possess a reliability-conferring property or function as means for conferring reliability—I do not wish to be taken as endorsing a strong metaphysical commitment on this point. Although I will not discuss the nature of these or the mechanisms by which they operate, a simple example should suffice to illustrate that it is not of a 'spooky' kind. Rather, it is quite familiar to philosophers of science (Kitcher, 1993). Consider the microscope. Its reliability is typically attributed to several interdependent factors: the application of optical laws in its design and calibration; the effective observation of entities, which also depends on the observer's prior knowledge; and the presence of a scientific community equipped with the background expertise to assess and adjudicate observational claims. The instrument's proper functioning, the researcher's methodological competence, and the community's epistemic norms together confer reliability on the use of the microscope.

The RIs I now discuss play analogous reliability-conferring roles, insofar as they consist in accessible scientific practices, methodological standards, research cultures, disciplinary debates, and other activities that are, to varying degrees, epistemically grounded and scientifically meritorious. Nothing spooky about that. The real challenge, rather, is to be as precise as possible in identifying RIs for specific cases. Here is where this chapter falls short. However, this is for a good reason. Recall that my only pretense with this chapter is to lay down the groundwork for an externalist epistemology of algorithms, and therefore my treatment of CR will be very general. The reader interested in concrete applications of CR to different domains is cordially invited to read Durán and Jongsma (2021) for cases on medicine and healthcare, Durán et al. (2024) for forensic science, and Humphreys (2020) for computer vision.

4.3.1 Reliability Indicators

For conceptual clarity, I distinguish between *type*-RIs and *token*-RIs. While the former refers to a unique category of indicators, the latter refers to an individual occurrence for that category. With this distinction in mind, the following type- and token-RIs are at the heart of CR:

- *Type$_1$-RI—Technical performance of algorithms* focuses on the specification, coding, execution, maintenance, and other technical features that contribute to the performance of the algorithm (e.g., high accuracy and low rate of errors, but also tolerance to domain change, repurposability, reusability, modularity, etc.). In this sense, typical cases of token$_1$-RI include practices and protocols for collecting, curating, storing, distributing, and analyzing data; the use of out-of-distribution data and data augmentation, parametrizations; benchmarking; choice of architecture; treatment of algorithmic kludges (Clark, 1987; Lenhard &

Winsberg, 2010); recasting (Durán, 2020); error treatment, and other techniques pertaining to achieving the desired performance of algorithms. Within this $type_1$-RI, one might also include an account of the employment of these practices, metrics, and methodologies, along with the specific circumstances under which the algorithm is specified and coded.

- *$Type_2$-RI—Computer-based scientific practice* focuses on securing algorithmic-based scientific research. It results from the operationalization and implementation of scientific concepts, causal structures, models and theories, laws and law-like principles, taxonomies, but also scientific metaphors and intuitions, values (epistemic and otherwise), idealizations, abstractions, and representations. This type-RI intends to capture the degree to which scientific units of analysis are implemented and operationalized into the algorithm. The selection and justification of domain knowledge are equally crucial in enhancing an algorithm's reliability. Let us note that the viability and success in doing so largely depend on algorithmic-related decisions, such as programming language choice, the use of formal techniques like verification methods (Fetzer, 1998), and the utilization of sub-modeling and multi-modeling (Durán, 2020).
- *$Type_3$-RI—Social construction of reliability* focuses on broader goals related to the acceptance—or rejection—of algorithms and their outputs by diverse communities (e.g., scientific, academic, general public), the realization of intended values and goals, and the overall assessment of an algorithm's scientific merits. Much of this occurs through $token_3$-RI, where debates about the replication of results, checks on the coherence of $\bar{o}$ with the established body of scientific beliefs and scientists' theoretical and other commitments, and other forms of intellectual exchange take place, embedded, so to speak, in diverse research cultures.[11]

Under this heading, CR is understood as a family of reliability-eliciting algorithmic-related indicators capable of crediting an algorithm as a reliable belief-forming method. It is important to note that by accepting a reliabilist epistemology, one also accepts the propensity likelihood that governs the reliability of a process. This translates into acknowledging that algorithms can occasionally be inefficient, contain errors, be unsuitable for specific purposes, misrepresent, and compute incorrect results. If failures perpetuate over time, the relative propensity governing CR shifts, rendering the algorithm ultimately unreliable.[12]

Furthermore, proponents of CR take note of human cognitive limitations in accessing some token-RIs, which conditions the claims about the reliability of

[11] To rephrase what Douglas (2004) persuasively argued: scientific and computational practices, as represented in $type_1$- and $type_2$-RI, together with the social processes associated with them, as represented in $type_3$-RI, are neither reducible to one another nor entirely independent.

[12] Humphreys notices that it is not only the high frequency, but also the quality of outputs that affects the reliability of an algorithm. He calls this problem *statistically insignificant but serious errors* (SIS-Errors) (Humphreys, 2020). My attempt to circumvent SIS-Errors can be found in Durán (2025).

an algorithm. They also need to accommodate the fact that token-RIs are neither absolute nor universally applicable. Not all token-RIs are credited, relevant, and applicable under the same criteria, nor does the same token-RI equally apply to all algorithms. CR is thus understood as perspectival, provisional, and subject to corrections, with no particular token-RI considered to have an all-or-nothing reliability-conferring property.

Thus understood, token-RIs come in degrees. The degree to which one token-RI is more relevant than another, or contributes to the overall reliability of the algorithm will depend on the context in which the algorithm is specified, coded, used, and maintained. It will depend on the epistemic and non-epistemic values and goals at stake. It will also depend on the culture of specifying, coding, maintaining, and using the algorithm of a given community (MacKenzie, 2005; Sundberg, 2010). In this sense, no individual (set of) token-RI can guarantee the reliability of all algorithms. Furthermore, even under the assumption that some token-RI is suitable for a given algorithm, this does not ensure that our reliability claims are eternally warranted. Old token-RIs can lose their appeal as new ones come to light. For these reasons, this chapter holds no pretensions to claim the completeness of the various type- and token-RIs presented here. Further arguments could be given on the need for additional type- or token-RIs not discussed here, or that some indicators are somewhat misplaced, or that some others need replacement. None of this is to say, however, that there are no stable type- and token-RIs that apply across many reliable algorithms. In fact, most of the token-RIs discussed next maintain, to my mind, a permanence in time despite changes and fine-tuning that occur with new technological and scientific developments.

Finally, I recognize that CR may not be readily accepted by everyone. As a reliabilist epistemology, one might feel that it still needs to address a few concerns. For starters, there are issues pertaining to the relevance and availability of type- and token-RIs, potential conflicts emerging among token-RIs, and their precedence, order, and weight. Unfortunately, these issues will not find a complete answer here. To my mind, 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 highlight that our best epistemic efforts can be geared towards the reliability of algorithms. Little more can be expected given the fallibility and limitations of human cognition. Taking note of these caveats, I now discuss a few types- and token-RIs in more details.

4.3.1.1 Type$_1$-RI: Technical Performance of Algorithms

In earlier versions of CR (Durán, 2013, 2018; Durán & Formanek, 2018), RIs mainly focus on the specification, implementation, tractability, and overall performance of algorithms. For instance, the first three token$_1$-RI discussed in Durán and Formanek (2018) put forward defining criteria for assessing the utility value of algorithms and their outputs *qua* computational methods. Consider *validation* procedures as an

example.[13] In automated diagnosis, algorithms are used for patient prognosis. One way to increase our confidence in the output is to compare the disease progression as indicated by the algorithm with clinical data from prior patients that share the same endotype or phenotype (Myszczynska et al., 2020). This practice validates the synthetic data computed by the algorithm with empirical data collected via diverse scientific methods (e.g., observation, experimentation, intervention, measurement, and others). The utility value of the algorithm is then considered appropriate if validation standards are satisfactory.

In this respect, subjecting algorithms' outputs to validation methods increases—or reduces—our confidence in the reliability of an algorithm, as it is a good indication of the algorithm's accuracy and margin of error. Validation methods also give a fair sense of the capacity to generalize the algorithm from the training data to new, undiscovered data. From a scientific perspective, validating algorithms also contributes to the rigor and reproducibility of research, ensuring that findings are based on sound methods.

Now, it should be expected that validation methods encompass a variety of techniques and methods.[14] As such, they are not all appropriate for the same goals. This means that a given validation techniques cannot be simply applied to different algorithms without prior critical discussion. There must be agreement on how suitable a given validation technique is for the algorithm and data in question, as well as the purposed goals and tasks (Lorscheid et al., 2012; Fagiolo et al., 2007). This is an often overlooked aspect of the social dimension of engineering the performance of algorithms. In Durán and Formanek (2018), we argued for a *history of (un)successful implementations* that affords this interpretation. The idea is simple and intuitive: good practices with visible success—such as high accuracy, low margin of error, ease of implementation, and formal verification—tend to endure over time, while less successful practices tend to be eradicated.

To illustrate this token$_1$-RI a bit further, consider *design prototyping*, a subfield of software engineering that assists developers in assessing alternative design strategies and deciding which is best for a particular goal. Since there are no standard methods for choosing the best strategy, researchers need to compare the requirements of the algorithm with various design approaches to evaluate which one possesses the best characteristics for fulfilling the intended objectives. The example I used in previous publications is a computer simulation involving networking. For this, there are different topologies: ring, star, tree, and mesh. In order to pick the most suitable one, diverse performance characteristics need to be evaluated to see which topology is better at meeting performance goals and constraints (Pfleeger & Atlee, 2009, Chapter 5).

The same point can be made with an example closer to machine learning. Take the case of BenevolentAI presented earlier, which utilizes *Best First Search* (BFS), a

[13] It is important to recognize that various forms of verification and validation exist, each conferring various degrees of reliability (Oberkampf & Roy, 2010).

[14] See Chap. 14 by Manganini and Primiero in this volume.

search algorithm highly successful for navigating graphs and trees. The primary goal of BFS is to find the most promising path to a target node based on a given heuristic. In this respect, BFS has proven to be extremely effective for searching suitable drugs within BenevolentAI's knowledge graph (Segler et al., 2018, p. 604). Classified under type$_1$-RI *history of (un)successful implementations*, BFS contributes to the reliability of BenevolentAI and the justification of its outputs.

Likewise, past failures must be, and typically are, avoided by competent programmers. The history of computing is littered with cases of failed software that changed specification and coding practices. Therac-25 is one tragic case (Leveson & Turner, 1993). As reported, the algorithm used by Therac-25 was not thoroughly validated, and the testing process was insufficient to catch critical bugs that led to radiation overdoses. Furthermore, there was poor error handling and reporting in the software. Error messages were often cryptic, and operators were not adequately trained to understand and respond to them. Finally, there were no redundancy safety mechanisms that could ensure that software failures do not result in such catastrophic outcomes. From the perspective of CR, these all amount to diverse indicators of the unreliability of the algorithm used in Therac-25.

4.3.1.2 Type$_2$-RI: Computer-Based Scientific Practice

Assessing the technical performance of algorithms facilitates justification in terms of increasing accuracy, predictive power, low error rates, tolerance to domain change, and the ability to multi-purpose algorithms and data, among other factors. However, this assessment is silent on the adequacy of algorithms for scientific purposes. A reliabilist epistemology must offer standards by which the algorithm used in a scientific context can be warranted to a greater or lesser degree.

Let me illustrate these ideas with a familiar example. Wu and Zhang's automatic facial recognition system exemplifies how accuracy alone does not exhaust the reliability of an algorithm. As mentioned in Sect. 4.2.1, the AUC-ROC measured 0.9540 predictive accuracy for their CNN. Such tremendous results cemented these researchers' confidence in the scientific merits of the algorithm. However, as discussed, there is no basis for such optimism. Criminality is a socially constructed concept that 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 concepts and frameworks, scientists are licensed to hold that the prediction—however accurate—lacks the merits required for legitimate scientific claims.

Under CR, the reliability of algorithms is not assessed based on high predictive accuracy. Science involves more than just measuring and classifying algorithmic outputs. In this respect, I believe that algorithms cannot and should not operate in isolation from the broader context of scientific undertakings. We need to delve not only into standard non-algorithmic scientific practice, but also into a form of scientific practice that evolves with and heavily depends on algorithms. The example

that immediately comes to mind is AlphaFold 2 (Jumper et al., 2021), but the use of BenevolentAI in discovering drugs to combat COVID-19 also illustrates the scientific significance of algorithms. Type$_2$-RI is an attempt to capture the family of token-RI connected to a larger body of scientific theories, beliefs, and practices within which algorithms are specified, coded, utilized, and maintained. In what follows, I lay out two obvious candidates.

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 specification, coding, use, and maintenance of algorithms in scientific contexts. Understood as a reliability indicator, *expert knowledge* reports on the many ways in which scientific expertise, technical expertise, and general competencies can be implemented into an algorithm.

To best understand this indicator, we must look at its various functions. For starters, it puts forward the algorithm's competencies, scope, and theoretical assumptions as conceptualized by the researchers involved in the specification, coding, maintenance, 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 algorithms, aptitudes to anticipate their merits intelligibly, and abilities of agents to manipulate them. For instance, setting up the variety of initial conditions, datasets, parameters and hyper-parameters (epochs, batch size, number of neurons, number of layers, dropout rate, etc.), all of which are complex yet critical for the performance and scientific merits of the algorithm. Consider determining which parameter to prioritize as a reliability indicator. Their selection and optimization are not trivial and yet fundamental for the general performance of algorithms (convergence of results, accuracy, overall performance) (Hutter et al., 2014). van Rijn and Hutter have conducted an informative experiment to show that the final performance metrics for deep learning models vary according to how different researchers select and optimize algorithmic parameters and instantiations (van Rijn & Hutter, 2018). Thus understood, experts contribute to the overall reliability of an algorithm by specifying relevant internal data-types, structures, relations, operations, and the like. They also credit reliability (or might identify instances of unreliability) by their pick and choose of datasets, parameters, and other variables.

As a reliability indicator, expert knowledge also attempts to accommodate the complexities of algorithms through the division of cognitive labor. Rather than being developed in isolation, algorithms involve 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). A core team

specifies and codes 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 make use of mathematical and computational methods others have devised and tested. There is no development of algorithms in solitude. Teams with diverse cognitive strengths and talents collectively collaborate in a variety of ways. Hence, the success or failure of algorithms 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.

Interestingly, the role and value of experts are being increasingly recognized in philosophical studies on algorithms. Ratti and Graves (2022) argued that documenting developers' motives and the code and specification of an ML are indicators of the reliability of the system. Newman (2016) has argued along similar lines with respect to computer simulations. 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 somewhat 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 algorithmic 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 specified and implemented. Their expertise is, however, crucial for the assessment of the reliability of the algorithm, and thus must be considered. Typically, their role is to inform, supervise, and sometimes even test the specification and coding of algorithms. But of course, these roles and interactions vary among cultures of research.

In connection with this, local practices and vernacular terminology often exceed what is captured by software engineering standards. Consider for instance how algorithms might naturalize or 'fossilize' 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 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 same concept 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 & Budson, 2000). Implementing a cogent definition of health is no trivial matter.[15]

Lastly, anyone who has written a piece of code knows all too well that not every line is documented. And even if algorithms were exhaustively documented, this would not guarantee an understanding of the code and its various functions. Thorough documentation—when it occurs—and well-intentioned software engineering may still fall short of capturing the methodological and epistemological competencies embedded in algorithmic systems. We still need to highlight the subtle interpretations, gentle disagreements, and non-verbal practices that permeate computational and scientific work, and that ultimately find their way into the algorithm.

Let me finish by noticing that this reliability indicator brings about another important aspect of algorithmic systems, namely, that they might only be *locally* reliable. The idiosyncrasies attached to documenting, specifying, coding, executing, and maintaining algorithms might make them only 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 analyzing 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 & Swetlitz, 2017, 2018). When implemented in South Korea and Denmark, only a fraction of the outputs computed by the algorithm matched—or closely matched—the local clinician's best diagnosis (Hamilton et al., 2019).

Knowledge-Based Integration

Scientific results do not come in discrete bits, nor are the objects of scientific inquiry independently sanctioned. Instead, scientific theory and practice constitute 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 algorithms 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" (Durán et al., 2024; Wu & Zhang, 2017). Given "race, gender and age, the faces of [the] general law-abiding public have a greater degree of resemblance

[15] Another, perhaps more thoroughly studied example is the implementation of the notion of fairness in algorithmic systems (Narayanan, 2018; Hao & Stray, 2019).

compared with the faces of criminals" (Durán et al., 2024; Wu & Zhang, 2017). It is, however, doubtful whether Wu and Zhang's CNN has any scientific merits. One reason (to add to those previously mentioned) is that Wu and Zhang's CNN is largely disconnected from accepted bodies of scientific knowledge. More precisely, the categories their CNN purports to use (i.e., {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. Thus, the CNN's outputs are based solely on picking out facial traits from selected photos, rather than being premised on a larger body of knowledge implemented in the algorithm.

I will call approaches that conceive of algorithms as disconnected from the larger body of scientific knowledge *just a bunch of data analysis* (JBDA). By doing this, I intend to emphasize that an algorithm performing mere data analysis, but disconnected from concepts, theories, law-like principles, hypotheses, and other scientific units of analysis, is unlikely to merit scientific credentials, regardless of its predictive accuracy. To my mind, JBDA ignores the 'bigger picture' of knowledge integration, interpretation, and operationalization into algorithms. In fact, I consider JBDA as misleadingly portraying algorithms as an objective, unambiguous, and scientifically grounded examination of data that generates scientifically meaningful outputs. Nothing could be further from the truth. 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. To these authors' minds, "like most technologies, machine learning is neutral" (Durán et al., 2024; Wu & Zhang, 2017). JBDA approaches advocate a form of scientific practice that is non-perspectival, socially disinterested, and impartial (i.e., epistemically and normatively neutral[16]), and disembodied from a 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 insight 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 algorithms 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 algorithm.[17] To briefly illustrate this idea, as this point encroaches on issues discussed under type$_3$-RI (see Sect. 4.3.1.3), consider BenevolentAI*, an algorithm whose working principles are JBDA. Suppose that BenevolentAI* puts forward baricitinib* as a

[16] For a critical view on this perspective in the context of algorithms, (see Pozzi & Durán, 2025).

[17] Meskhidze makes a similar claim in the context of ML applications in astrophysics. The chapter discusses "physics-informed machine learning," where physical laws and domain-specific knowledge are embedded within the algorithm and are crucial to its performance. From the standpoint of an epistemology of algorithms, my approach diverges in that Meskhidze prioritizes fostering transparency and interpretability, in contrast to externalist accounts such as reliabilism (see Chap. 18 by Meskhidze in this volume).

drug with high chances of combating COVID-19 symptoms. Would researchers be justified in believing baricitinib*? Surely not at face value, but only once it is embedded in a larger body of knowledge about COVID-19 and after some clinical trials show its scientific worth. This interpretation, however, shifts the reliability of the algorithmic system to a human agent capable of re-interpreting the outputs in light of background knowledge. Examined in isolation, algorithmic systems whose working principles are solely based on JBDA can sustain claims about single justification, but they are less likely to be *reliable* in any sense of the term.

What the examples of BenevolentAI and BenevolentAI* show, I believe, is that we might still be justified under JBDA-like algorithms 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, and/or (b) its outputs are later assessed by the relevant community and within a corpus of scientific knowledge (more on this in Sect. 4.3.1.3.)

Let me further illustrate this reliability indicator. Take again BenevolentAI, which utilizes information gathered from scientific research papers, structured and unstructured biomedical data, and drug and pharmaceutical industry data. BenevolentAI also implements knowledge graphs that structure data into causal relationships between known diseases, genes, environmental factors, and approved drugs (Smith et al., 2021). Furthermore, BenevolentAI aligns with auxiliary assumptions, theories, and structures of drug molecular profiles, as well as mechanisms integral to the process of damaging healthy cells and tissues. It also incorporates theories about genetics, medical studies of disease, and biological models relating genes to drug effects. Through this knowledge-based integration, outputs generated by BenevolentAI are better justified than those by BenevolentAI*. Indeed, the JBDA-like version does not implement any accepted model or concept into its algorithm, does not operationalize knowledge graphs, and does not represent mechanisms integral to knowledge about damaging healthy cells and tissues. It is the theoretical rigor, along with a history of (un)successful implementations (Durán & Formanek, 2018), domain and expert competence, and possibly some skilled insight that leverages the reliability of BenevolentAI over and above BenevolentAI*.[18] To put the same idea rather bluntly, the lack of such token$_2$ reliability indicators places BenevolentAI* at a justificatory disadvantage compared to BenevolentAI.

4.3.1.3 Type$_3$-RI: Social Construction of Reliability

The performance of algorithms (type$_1$-RI) is undoubtedly required for justification. But while a necessary condition, it is certainly not sufficient. The paradigmatic example is Wu and Zhang's CNN. This algorithm leverages a subset of RI$_1$ (most prominently, validation) but makes claims about an alleged law of facial recognition that is difficult to justify. Expert and knowledge integration (type$_2$-RI) are also

[18] Thanks go to Emanuele Ratti for pressing on clarifying this point.

fundamental to the reliability of algorithms. But again, necessary but not sufficient. BenevolentAI furnishes a good example. The algorithm is robust and built on a solid scientific basis. However, baricitinib was later flagged as counter-prescribed for immunocompromised patients. So, what is missing? To my mind, the reliability of algorithms must also be assessed within social processes that aim to achieve standards of scientific value and thresholds for acceptance of ō. Let me put the same idea in different form. The performance of algorithms along with expert knowledge and knowledge-based integration observe that the relevant scientific structures and processes, commitments and categories, entities and concepts are—to an acceptable degree–correctly implemented into, and computed by the algorithm. But the justification of ō requires something else. We need to further seek for the consistency of ō with a larger corpus of knowledge. In this way, we observe that the output aligns with accepted scientific commitments and satisfies standards of quality, evidential support, and relevance, along with other key scientific qualifications.

Let me quickly illustrate how this reliability indicator would work using the story behind BenevolentAI. After announcing barinitricib, diverse groups within the scientific community began to debate the benefits—and dangers—of this drug. A major concern that emerged was that the mechanisms of action of baricitinib would block JAK-STAT signaling pathways (mainly via JAK1 and JAK2), thus impairing interferon-mediated antiviral responses (Favalli, 2020, p. 1013). Blocking interferon would allow attack by other viruses (e.g., herpes zoster and herpes simplex) which in some cases may be more harmful than COVID-19. Favalli (2020) later reported on the potential harms of administering baricitinib to some patients, most importantly immunodeficient patients. As a response, Richardson and colleagues accepted the conditions under which BenevolentAI was a reliable indicator—and reasonably use this debate and incorporate new functions into the algorithm.

As a reliability indicator, this scientific debate regulated on how much evidence was required to accept BenevolentAI's outputs. It draw thresholds of which errors and artifacts can be tolerated, and to what extent. It also determined which assumptions are fit for purpose. Commitments to reliable algorithms are commitments to a network of scientific methodologies, standards, and traditions expressed through scientific debate. As Elgin points out, this network enables scientists to build on each other's work. They can be confident that justified outputs have the epistemic value their discipline prescribes (Elgin, 1996, p. 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 administering 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—one that is not always successful. It involves situated background assumptions and perspectives, scientific inquiry permeated by contextual values and interests, and a myriad of concepts implemented in algorithms that are socially constructed, historically contingent, and often idiosyncratic to particular disciplines. Take again variations in the definition of 'health' and 'disease' (Boorse, 2011; Sisti & Caplan, 2017). Each

concept operates under a myriad of cultural, political, economic, and moral values. Caruana et al. (2015) discuss a neural network specified 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). Now, while these findings are statistically accurate—as evidenced by the presence of type$_1$-RI and type$_2$-RI indicators—the system is nonetheless perceived as unreliable, owing to physicians' divergent background assumptions about what a predictive algorithm ought to provide. According to Theunissen and Browning (2022), physicians often presume that the algorithm predicts outcomes based on a shared baseline of care, rather than adapting to differential care relative to patients' backgrounds. As a result, the algorithm's outputs may be called into question due to oversimplifications in one or more of its embedded assumptions. This further underscores that neither type$_1$-RI nor type$_2$-RI is individually—or jointly—sufficient for establishing the reliability of algorithmic systems.

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 also find their way into the use and perception of the outputs of algorithms. Just like many other scientific methodologies, entrenching the reliability of algorithms requires a delicate balancing act. We bring together our best technical knowledge, theories, methodologies, and social skills. We do so in an attempt to justify believing that the output of a given algorithmic system can be scientifically valuable, represent, or even true.

4.4 Final Thoughts

I have presented CR as a reliabilist epistemology for the justification of algorithms' outputs. I also presented and discussed diverse types and token reliability indicators that form the basis of reliable algorithmic systems. In sum, I set out to defend the following claim: we have—or increasingly have—justification for believing algorithms' outputs when they are generated by a reliable belief-forming method. To be reliable here means that the algorithm is specified, coded, used, and maintained utilizing specific, tailor-made reliability indicators designed for the purpose at hand.

Admittedly, CR construes justification as inherently provisional. As discussed, reliability indicators might change over time and even be replaced. They might also be hard to measure or resolve conflicts. But I believe this is part of the self-critical and self-correcting endeavors that we find in scientific research. It might also be the best epistemic effort we can offer given a context and our limited resources. These activities constitute our best knowledge, metrics, methodologies, and practices, even if subjected to further scrutiny and revision.

I also listed a few issues with CR that I was unable to address but which might be considered conditional to its acceptability. I can accept that. But CR is a step in the right direction, if not as a more adequate epistemology of algorithms, at least as an alternative to internalist epistemologies. In this regard, the chapter has achieved its goal of setting up an externalist epistemology of algorithms, a goal that should be appreciated in its own right.

Acknowledgments This paper has been in the making for quite some time, and many people deserve acknowledgment for their thoughtful comments and suggestions. Let me begin with my co-editor: thank you, Giorgia, for your many close readings of this chapter. I am also grateful to Jack Casey for engaging discussions on these topics and for helping me avoid several missteps.

Many others have contributed in various ways: Federica Russo, Emanuele Ratti, Edoardo Datteri, Giuseppe Primiero, Viola Schiaffonati, Rawad El Skaf, Manuel Barrantes, Karin Jongsma, Andrea Ferrario, Nico Formanek, Charles Rathkopf, Emma-Jane Spencer, and Atocha Aliseda. Some of them believe in CR, some do not, but all have been supportive and encouraging of my ideas. Naturally, all wrongs—and all rights—are mine.

Finally, thanks go to Kass and Diego. To Kass, for endlessly listening and sharing thoughts as I structured and shaped CR. I am sure I bored you on a cosmic scale—thank you for still listening, under all conditions. And to Diego, for teaching me one absolute truth: that everything is silly.

References

Alston, W. P. (1995). How to think about reliability. *Philosophical Topics*, *23*(1), 1–29.

Alvarado, R., & Humphreys, P. (2017). Big data, thick mediation, and representational opacity. *New Literary History*, *48*(4), 729–749. https://doi.org/10.1353/nlh.2017.0037

Beisbart, C. (2021). Opacity thought through: On the intransparency of computer simulations. *Synthese*, *199*, 11643–11666.

Beisbart, C. (2025). In which ways is machine learning opaque? In Durán, J. M., & Pozzi, G. (Eds.), *Philosophy of science for machine learning: Core issues and new perspectives*. Synthese Library. Springer.

Blanz, V., & 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* (pp. 187–194). ACM Press/Addison-Wesley Publishing Co.

Boge, F. J. (2022). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*, *32*(1), 43–75.

Boorse, C. (2011). Concepts of health and disease. In Gifford, F. (Ed.), *Handbook of the philosophy of science* (pp. 13–64). Elsevier.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1). https://doi.org/10.1177/2053951715622512

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (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* (pp. 1721–1730). Association for Computing Machinery. https://doi.org/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. https://doi.org/10.1086/709729

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

Duede, E. (2022). Instruments, agents, and artificial intelligence: novel epistemic categories of reliability. *Synthese*, *200*(6), 491.

Durán, J. M. (2013). *Explaining simulated phenomena: A defense of the epistemic power of computer simulations* (PhD thesis). Institut für Philosophie, Universität Stuttgart. Retrieved from https://elib.uni-stuttgart.de/bitstream/11682/5409/1/Thesis_Duran.pdf

Durán, J. M. (2018). *Computer simulations in science and engineering. Concepts - practices - perspectives*. Springer.

Durán, J. M. (2020). What is a simulation model? *Minds and Machines*, *30*, 301–323.

Durán, J. M. (2021). Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare. *Artificial Intelligence*, *297*, 103498. https://doi.org/10.1016/j.artint.2021.103498

Durán, J. M. (2025). In defense of reliabilist epistemology of algorithms. *European Journal for Philosophy of Science*, *15*, 37.

Durán, J. M., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines*, *28*(4), 645–666.

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, *47*(5), 329–335. https://doi.org/10.1136/medethics-2020-106820

Durán, J. M., van der Vloed, D., Ruifrok, A., & Ypma, R. J. F. (2024). From understanding to justifying: Computational reliabilism for AI-based forensic evidence evaluation. *Forensic Science International: Synergy, 9*, 100554. ISSN:2589-871X. Available at: https://www.825sciencedirect.com/science/article/pii/S2589871X24001013

Eden, A. H. (2007). Three paradigms of computer science. *Minds and Machines*, *17*, 135–167.

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

Elgin, C. (2017). *True enough*. MIT Press.

Emani, S., Rui, A., Rocha, H. A. L., Rizvi, R. F., Juacaba, S. F., Jackson, G. P., & Bates, D. W. (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. https://doi.org/10.2196/31461

Fagiolo, G., Moneta, A., & Windrum, P. (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., Maioli, G., & Caporali, R. (2020). Baricitinib for COVID-19: a suitable treatment? *The Lancet*, *20*, 1012–1013.

Fetzer, J. H. (1998). Program verification: The very idea. *Communications of the ACM*, *37*(9), 1048–1063.

Goldman, A. (1979). What is justified belief? *The Justification of Belief*, *105*(9), 1–23.

Goldman, A. I. (2012). *Reliabilism and contemporary epistemology*. Oxford University Press.

Hamilton, J. G., Genoff Garzon, M., Westerman, J. S., Shuk, E., Hay, J. L., Walters, C., Elkin, E., Bertelsen, C., Cho, J., Daly, B., Gucalp, A. (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.

Hao, K., & Stray, J. (2019). *Can you make AI fairer than a judge? Play our courtroom algorithm game*. MIT Technology Review. Retrieved June 9, 2025, from https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/

Humphreys, P. W. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. Oxford University Press.

Humphreys, P. W. (2009). The philosophical novelty of computer simulation methods. *Synthese*, 169(3), 615–626.

Humphreys, P. (2020). *Neural Nets: Why reliabilism is an inappropriate epistemology for them*, YouTube video, 13 October 2020. Retrieved August 15, 2024, from https://www.youtube.com/watch?v=2VFPXbrCqzM&t=830s

Hutter, F., Hoos, H., & Leyton-Brown, K. (2014). An efficient approach for assessing hyperparameter importance. In E. P. Xing & T. Jebara (Eds.), *Proceedings of the 31st International Conference on Machine Learning* (Vol. 32(1), pp. 754–762). PMLR.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S., Ballard, A., Cowie, A.,

Romera-Paredes, B., Nikolov, S., Ain, R., Adler, J., ..., Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*, 583–589.

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. Oxford University Press.

Lenhard, J., & Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in the History and Philosophy of Modern Physics*, *41*, 252–262.

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

Lorscheid, I., Heine, B.-O., & Meyer, M. (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.

MacKenzie, D. (2005). Computing and the cultures of proving. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, *363*(1835), 2335–2350.

Massimi, M., & 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*. Retrieved from https://www.wired.co.uk/article/future-health-trends

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

Myszczynska, M., Ojamies, P., Lacoste, A., Neil, D., Saffari, A., Mead, R., Hautbergue, G., Holbrook, J., & Ferraiuolo, L. (2020). Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature Reviews Neurology*, *16*, 440–456. https://doi.org/10.1038/s41582-020-0377-8

Narayanan, A. (2018). Translation tutorial: 21 Fairness definitions and their politics. In *Proc. Conf. Fairness, Accountability, and Transparency (FAT*)*, New York, NY, USA. Tutorial presentation.

Newman, J. (2016). Epistemic opacity, confirmation holism and technical debt: Computer simulation in the light of empirical software engineering. In F. Gadducci & M. Tavosanis (Eds.), *History and Philosophy of Computing. HaPoC 2015. IFIP Advances in Information and Communication Technology* (Vol. 487, pp. 256–272). Springer.

Oberkampf, W. L., & Roy, C. J. (2010). *Verification and validation in scientific computing*. Cambridge University Press.

Parker, W. S. (2020). Model evaluation: An adequacy-for-purpose view. *Philosophy of Science*, *87*, 457–467.

Humphreys, P. *Epistemic opacity and epistemic inaccessibility*, unpublished manuscript. Available at: https:\/wordpress.its.virginia.edu/Paul_Humphreys_Home_Page/files/2016/02/epistemic-opacity-and-epistemic-inaccessibility.pdf

Pfleeger, S. L., & Atlee, J. M. (2009). *Software engineering: Theory and practice* (4th ed.). Pearson.

Pozzi, G., & Durán, J. M. (2025). From ethics to epistemology and back again: Informativeness and epistemic injustice in explanatory medical machine learning. *AI & Society, 40*, 299–310. https://doi.org/10.1007/s00146-024-01875-6

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

Räz, T., & Beisbart, C. (2022). The importance of understanding deep learning. *Erkenntnis*. https://doi.org/10.1007/s10670-022-00605-y

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

Richardson, P., Griffin, I., Tucker, C., Smith, D., Oechsle, O., Phelan, A., Rawling, M., Savory, E., & Stebbing, J. (2020). Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. *The Lancet*, *395*(10223), e30–e31. https://doi.org/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., & Budson, A. E. (2000). Health of organisms and health of persons: An embedded instrumentalist approach. *Theoretical Medicine and Bioethics*, *21*(4), 339–352.

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

Ross, C., & Swetlitz, I. (2018). IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. *Statnews*. Retrieved from https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf

Russo, F., Schliesser, E., & Wagemans, J. (2024). Connecting ethics and epistemology of AI. *AI & Society*, *39*(4), 1585–1603.

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

Sisti, D., & Caplan, A. L. (2017). The concept of disease. In Solomon, M., Simon, J. R., & Kincaid, H. (Eds.), *The Routledge companion to philosophy of medicine* (pp. 5–15). Routledge.

Smith, D. P., Oechsle, O., Rawling, M. J., Savory, E., Lacoste, A. M. B., & Richardson, P. J. (2021). Expert-augmented computational drug repurposing identified baricitinib as a treatment for COVID-19. *Frontiers in Pharmacology*, *12*, 709856. https://doi.org/10.3389/fphar.2021.709856

Sullivan, E. (2022). Understanding from machine learning models. *British Journal for the Philosophy of Science*, *73*(1), 109–133.

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*, *41*, 273–281.

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

Theunissen, M., & Browning, J. (2022). Putting explainable AI in context: Institutional explanations for medical AI. *Ethics and Information Technology*, *24*(2), 23.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71–86.

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

Vulsteke, C., Ortega Arevalo, M., Mouton, C., Stam, K., Goethals, R., Ameye, F., Populaire, C., Peeters, M., & Verdonck, P. (2018). Artificial intelligence for the oncologist: Hype, hubris, or reality? *Belgian Journal of Medical Oncology*, *12*(7), 330–333.

Wachter, S., Mittelstadt, B., & Russell, C. (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., & Zhang, X. (2016). Automated inference on criminality using face images. arXiv: 1611.04135v1. https://arxiv.org/pdf/1611.04135v1

Wu, X., & Zhang, X. (2017). *Responses to critiques on machine learning of criminality perceptions* (Addendum of arXiv:1611.04135). arXiv:1611.04135v3. https://doi.org/10.48550/arXiv.1611.04135

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, *34*(2), 265–288.

Juan M. Durán is an Assistant Professor in the Faculty of Technology, Policy and Management at TU Delft. His research focuses on epistemology and the philosophy of science in relation to computer-based scientific and engineering practices, including computer simulations, artificial intelligence, and Big Data. He also actively investigates the ethical dimensions of AI technologies and their impact on society.

In 2019, Durán received the Herbert Simon Award from the International Association for Computing and Philosophy (IACAP) in recognition of his outstanding early-career contributions. The award honors scholars whose original research is transforming the dialogue at the intersection of computing and philosophy.