# Machine Learning as Epistemic Activity

Jun Kyu Lee

**Abstract**

The existing philosophical debates on the use of machine learning (ML) in science have organised around the issue of model opacity. What these debates often leave unanalysed is the source of a model's epistemic validity: the overarching ML pipeline that includes dataset collection, model training, and model evaluation. In order to carry out this analysis, I utilise the activity-based framework developed by Hasok Chang, in which the pipeline can be viewed as an integrated activity where each step has its own success criterion. In addition to the steps' individual success, the overall success of the pipeline depends on their coherent coordination. In the course of this analysis, I show that the epistemic validity of the pipeline crucially depends on the operational provenance of the data, which allows the agents to assess the data's adequacy for the pipeline's aim through two requirements (Discrimination and Traceability). The upshot is that opacity is decomposed into a number of separately diagnosable issues, of which the model complexity is only one. Furthermore, link uncertainty is revealed to be a property of dataset collection, and not of the model.

## 1 Introduction

As various machine learning (ML) models have become embedded within a range of scientific applications, philosophers have discussed the exact epistemic role these models can play in the scientific process. Much of the existing discussion has organised around the contention that ML models are *opaque*[1]. In response to this opacity, philosophers have asked whether the outputs of ML models can be trusted (Durán 2026); whether such models can confer understanding on their users (Boge 2022; Räz and Beisbart 2024); and whether they can be used to justify hypotheses (Duede 2023). Meanwhile, Sullivan (2022) has prominently argued that it is *link uncertainty*, and not opacity, that limits the amount of understanding derivable from an ML model.

To these varied contributions, I wish to bring attention to the epistemic importance of the wider context in which an ML model exists. This context is the *ML pipeline*, within which data are collected, a model is searched for, the performance of the model is evaluated, and so on. While some philosophers have already mentioned this pipeline in their writings (e.g. Hansen and Quinon 2023; Freiesleben and Grote 2023; Ratti et al. 2026)[2], the philosophical literature has yet to deeply analyse what this pipeline, in particular the decomposability of

[1] As I will describe below, there are multiple notions of opacity as applied to ML models (see also Burrell 2016; Creel 2020; Facchini and Termine 2022; Beisbart 2026).

[2] At the time of writing, Ratti et al. (2026) was a book draft in development. The version referred to here is the 29 Jul 2026 version, available at https://philsci-archive.pitt.edu/id/eprint/30610.

its steps, entails for the epistemic validity of ML use in science. Conflating the constituent steps of the pipeline muddles the epistemic situation, because each step is a distinctive task with its own success criterion that can be satisfied independently of the others.

In this paper, I develop this idea through Chang's (2022) analysis of epistemic activities and systems of practice, which supplies the distinctions required. In particular, his framework allows us to describe how different activities are *coordinated* through their *external functions*, where these functions are contingent upon the satisfaction of their *inherent aims* but nevertheless can become decoupled from them under incoherence. The main claims that I wish to establish are:

1. The epistemic status of scientific ML depends on whether the steps of the pipeline are effectively coordinated towards the relevant scientific purposes.
2. Consequently, an ML pipeline can contain individually successful steps, but remain epistemically defective as a whole due to the *incoherence* in how the steps are coordinated.
3. An important factor in this coherence is the *operational provenance* of the data that the model trains on, through which the model is connected to the target domain in the world.
4. By applying this lens to the issues of opacity and link uncertainty, we can delineate the extent to which these issues impact ML's epistemic status in a particular application.

Admittedly, this paper engages with multiple related debates in the literature, yet its goal is to show that a single relocation of focus towards activities can clarify the major philosophical issues in ML. The existing account of ML in the philosophical literature that comes closest to mine is Kieval (2026), who writes that "constructing a working DL [deep learning, a subset of machine learning] model requires skillful manipulation of multiple interconnected elements" (Kieval 2026, 265), and that "[s]uccessful DL modelling in science ... requires knowing how to use a model to pursue a particular epistemic aim" (Kieval 2026, 266). While this is a step in the right direction, he does not provide further analysis as to exactly when ML modelling is successful[3]. This is the gap that I wish to fill in with Chang's machinery.

The structure of this paper is as follows. Section 2 is a preliminary description of ML training, with a particular focus on clearing up a conflation that one often encounters in the literature. In Section 3, I show how Chang's concepts can be applied to the ML training process to clarify the epistemic situation, and contrast my account with the technical artefactualism of Ratti et al. (2026). This is followed by an in-depth discussion of the dataset collection process in Section 4, which has often been neglected in the literature. Borrowing ideas from operationalism, I argue that the operational provenance of the data is essential for the epistemic validity of the ML pipeline. In Section 5, I move on to applying the developed framework to the notion of opacity. Section 6 then discusses Sullivan's link uncertainty, and I propose a redefinition of the concept under my framework that preserves her intuitions. Section 7 compares my framework with computational reliabilism, and argues that the former answers the question left open by the latter.

[3]Kieval only goes on to suggest that de Regt's (2017) notion of pragmatic understanding is needed for success.

## 2 ML training process

In the existing philosophical literature on ML, the phrase "ML model" is sometimes ambiguously used to refer to several distinct things[4]. As this ambiguity matters for my discussion, I begin by fixing terminology. Throughout, I restrict my attention to *supervised learning*, where the goal is to find the model that relates inputs to outputs in a satisfactory manner[5].

Hence, I wish to distinguish between the following three concepts involved in training ML models:

1. The *learning algorithm*, which we denote by $A$.
2. The *family of models* from which this algorithm finds the appropriate model, denoted as

$$\mathcal{F} = \{f_\theta : \theta \in \Theta\}.$$

   Here, $\Theta$ is the *parameter space*, where each $\theta \in \Theta$ specifies the model $f_\theta$.
3. The model $f_{\hat{\theta}}$ thus chosen, which is the *trained model*.

The training process can be described in the following simplified manner, which is sufficient for my purposes. We start with the *dataset* $D$, which for supervised learning is a set of input-output pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. $D$ is randomly partitioned into the *training set* $D_{\text{train}}$ and the *held-out test set* $D_{\text{test}}$, where the random partitioning ensures that both $D_{\text{train}}$ and $D_{\text{test}}$ can be treated as being produced by the same data-generating process. Then the algorithm $A$ searches $\mathcal{F}$ for the $f_{\hat{\theta}}$ that "best" relates the $\mathbf{x}_i$s to the $y_i$s, where "best" is operationalised by the *loss* $\mathcal{L}$ comparing the model's outputs $f_\theta(\mathbf{x}_i)$ to the reference label $y_i$. A model with a lower loss is taken to be better[6].

In many ML algorithms, the optimisation problem has multiple local minima, and it is virtually impossible to discover the model that corresponds to the global minimum of the loss function (nor is it possible to conclusively establish that $A$ has found this model). Instead, ML practitioners settle for the loss that they deem to be sufficiently low for the purpose. Crucially, $\mathcal{L}$ is minimised on $D_{\text{train}}$ but the model is evaluated on $D_{\text{test}}$. The model may have a very low training loss, and yet perform badly on $D_{\text{test}}$ that it had not seen during training. As such, withholding $D_{\text{test}}$ allows the practitioners to perform an internal check of the training process. This is central to ML practices, and I will discuss it in more detail.

The model training process can be viewed as a central step in a larger *ML pipeline*, which includes other components such as dataset collection, dataset cleaning, feature engineering, and model deployment (Amershi et al. 2019; Ratti et al. 2026). In this paper, we mainly focus on dataset collection and model training, which are the steps most relevant to the

[4]Exceptions are Facchini and Termine (2022) and Termine et al. (2026), who distinguish between what they call the *training engine* (learning algorithm) and the learned model.

[5]In this paper, I exclude large language models (LLMs) from consideration, although their training involves prominent supervised components. For scientific ML that I consider, the model predicts quantities in the target domain directly; LLMs output linguistic sentences about the domain instead.

[6]For neural network models, Buchholz and Raidl (2025) views this adjustment process as akin to falsification of theories.

epistemic validity of the entire pipeline. An important step of the pipeline not covered here in detail is the *architecture selection* procedure where $\mathcal{F}$ is chosen, which can be heavily guided by domain knowledge (Hansen and Quinon 2023)[7].

## 3 ML pipeline as epistemic activity

Hasok Chang defines an *activity* as "a programme of action designed for the achievement of a recognizable aim", and an *epistemic activity* as "knowledge-related activity, aimed at acquiring, assessing or using knowledge" (Chang 2022, 15–16). Each activity has a unique *inherent aim*, which partly defines the activity. The success of an activity consists in satisfying its inherent aim. But we are often interested in the *consequences* of this successful execution, which are called *external functions* of the activity. As an illustration, Chang considers the activity of striking a match. Here, the inherent aim is to light the match, which "is the whole point of the activity itself, regardless of why one is engaged in that activity" (Chang 2022, 36). At the same time, the external functions include lighting a Bunsen burner or burning down the house. Note that, unlike the inherent aim, an activity may have multiple external functions.

On a closer analysis of an activity, we usually find that it integrates other activities. These constituent activities are coordinated towards the goal of satisfying the integrated activity's inherent aim, although these constituent activities still retain their own inherent aims. While this conception of activities-inside-activities may seem unworkable due to infinite regress, Chang defends it by its pragmatic fruitfulness for analysing practices: "we do gain useful insights from analysing an activity into its *apparent* constituents, and the analysis should be carried out wherever it is productive, and only and as far as it is productive" (Chang 2022, 38, emphasis in original). A consequence of this characterisation is that we may have epistemic activities that are quite large-scale, as long as they have a single inherent aim. On the other hand, a *system of practice* is "formed by the coordination of activities without a single overarching aim", and it "has multiple system-level aims" (Chang 2022, 39).

An important quality of activities and systems of practice is *operational coherence*, which consists in "aim-oriented coordination", or "in doing what makes sense to do in specific situations of purposive action" (Chang 2022, 40). More concretely, it "consists in an effective coordination of the external functions of various activities" (Chang 2022, 39) in achieving the aims of integrated activities or systems of practices[8]. It is the operational coherence with respect to the inherent aim that identifies and individuates an activity. Some relevant properties of operational coherence are as follows: it is not underwritten by a unified theory, and can be achieved in an *ad hoc* manner; nevertheless, it is achieved through deep understanding of the surroundings, not by mere conventions; it is a matter of degrees. Finally, coherence "is design for success, and that design is based on empirical learning" (Chang 2022, 44) but does not *cause* success, and coherent activities may fail—and incoherent ones succeed—by extrinsic accidents.

[7] Note that the three steps mentioned here roughly correspond to Kieval's (2026) three "decision procedures" needed for constructing a working ML model.

[8] In this quoted passage, Chang only mentions the satisfaction of system-level aims, but I take it that this characterisation also pertains to activity-level aims.

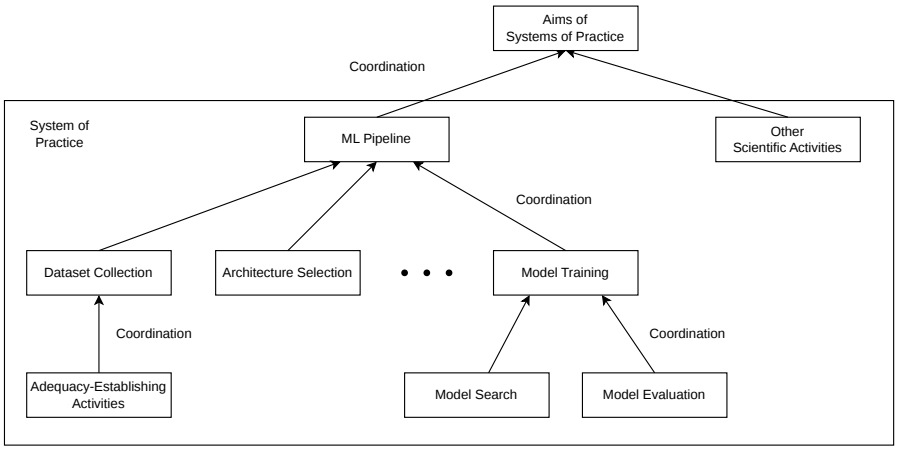



Figure 1: Machine learning as an epistemic activity within a larger scientific system of practice.

Let us see how this framework can be applied to ML. The ML pipeline can be construed as an integrated epistemic activity that incorporates activities such as dataset collection and model training (see Figure 1). Its inherent aim is to create a model that predicts the output values from the input values with an indication of its predictive performance, where we take these values to refer to quantities, properties, or categories (which I will denote simply as "quantities") of the *outside world*[9]. This aim is epistemic, for its satisfaction grants us knowledge about how, and how well, the quantities relate to each other[10]. But crucially, it is not part of the inherent aim to create a model that predicts the output *well*, despite the common characterisation of ML modelling as being solely aimed at prediction accuracy. For example, suppose that we are using ML modelling to test the efficient market hypothesis (Fama 1970). Then the failure of the trained model to be better than random directly supports the hypothesis with respect to the variables considered. So what makes ML modelling suitable is whether the system of practice can utilise the resulting model for its purposes.

Now, consider the *model training* step within the pipeline. Its inherent aim is to create a model that relates the inputs to the outputs *as* numbers (not as real-world quantities, which is the aim of the pipeline as a whole); where this model must generalise beyond the immediate dataset $D$, to further data that we treat as produced by the same data-generating process as $D$. This activity incorporates the constituent activities of *model search* and *model evaluation*. During model search, we utilise the learning algorithm $A$ to search for $f_{\hat{\theta}} \in \mathcal{F}$, which involves optimising $\mathcal{L}$ on $D_{\text{train}}$ alone[11]. The coherence of this activity independently of the other constituents of the pipeline underwrites Ratti et al.'s argument when they write

[9]One exception is when we are using synthetic data to evaluate the performance of the learning algorithms themselves.

[10]However, I bracket the question of whether the constituent activities of the pipeline are *epistemic* activities, although I think many of them are defensibly so. An integrated epistemic activity may have non-epistemic activities as its constituents: establishing a certain chemical claim is an epistemic activity, although its constituent match-lighting activity is not.

[11]This often requires further sampling of $D_{\text{train}}$ to create the *validation dataset*.

that "MLMs [ML models] are *optimization machines*" (Ratti et al. 2026, 72, emphasis in original). (Note that their usage of the term "ML model" is somewhat inconsistent across the text. Here, it refers to the learning algorithm that optimises the loss function.)

If $\mathcal{F}$ contains highly complex functions, as is the case with modern ML pipelines, then the model evaluation step becomes important. This is because $A$ could have found a model that merely memorises $D_{\text{train}}$ in a way that does not generalise, which is the phenomenon known as *overfitting*[12]. The inherent aim of model evaluation is to gauge the generalisation performance by testing the model on $D_{\text{test}}$. Meanwhile, the relevant external function of model evaluation is to check how the model performs on data that it had not seen during the training process. The inherent aim and the external function do not have the same satisfaction criteria, and a good performance on $D_{\text{test}}$ does not automatically imply that the model is satisfactory for the aim of the ML pipeline.

For instance, if we continue to reuse $D_{\text{test}}$ in evaluating the model, then $D_{\text{test}}$ indirectly becomes "seen" by the model training process. Subsequently, the satisfaction of the inherent aim becomes increasingly mismatched to the external function that the model training activity (and hence, the ML pipeline) requires from model evaluation. In other words, model training becomes *incoherent*, and this is a well-known problem in the ML community: success on the public benchmarks for model performance becomes increasingly meaningless, as researchers effectively end up optimising their models for those benchmarks. From this example, we see how activity-level analysis allows us to capture some serious epistemic failures in the ML pipeline. If we had instead characterised ML models as products of optimisation without further context, then it would be difficult to specify the fault with test set reuse.

As a further illustration of the advantages of my account, we may contrast it with a different goal-oriented account of ML models, namely technical artefactualism (TA) developed by Ratti et al. (2026). Their starting point is the *ICE theory* (Vermaas and Houkes 2006), in which the user of an ML model must be justified in believing that the model has the capacity to, and actually contributes towards, the successful execution of the goal under the use plan[13]. The *use plan* refers to "a series of ... actions in which manipulations of $x$ are included as contributions to realising the given goal" (Vermaas and Houkes 2006, 7). To the extent that the introduction of use plans acknowledges that ML models are built within the context of a wider aim, the ICE theory is congenial to my framework.

However, this approach is quickly met with difficulties, as we see from Ratti et al.'s response to Hurshman's (2024) objection. Hurshman argues that the ICE theory is unsuitable for ML, because the opacity of ML models prevents the users from obtaining the justification required. Ratti et al. effectively concede this point, and constrain the goal's scope to the optimisation of the loss function (in effect re-interpreting "ML model" to refer to the learning algorithm, as noted previously). While they are undeniably correct that the learning algorithm $A$ acts as an "optimization machine", this move only defers the crucial question of why the ML model $f_{\hat{\theta}}$ is valid for the purpose of its deployment. For this

[12]Note that memorisation of the training set is not fatal for generalisation in itself. The phenomenon of *double descent* (Belkin et al. 2019) shows that the generalisation capability of a neural network model can further increase after the training set has been fully memorised.

[13]This is a very compressed description that is sufficient for my purposes. The full description is available in Section 3.4 of Ratti et al. (2026).

question, they gesture towards reliabilism (see Section 7 for my discussion of reliabilism) and domain-dependent considerations, but do not provide further details[14]. An activity-based approach allows us to continue Ratti et al.'s project beyond optimisation, as it views optimisation as one activity (i.e. model search) among others; we may then analyse the other activities systematically, and discover how they contribute to the epistemic validity of the entire pipeline.

# 4 Dataset collection

## 4.1 How ML models speak about the world

While ML model training is concerned with relating data as numbers stored in the computer, the ML pipeline is concerned with quantities in the world[15]. In this paper, I wish to draw an explicit structural distinction between the data-as-numbers and the real-world quantities, as it is central to my argument. This stance is similar to Andrews (2025), who writes that "data is a form of mathematical representation, it does not intrinsically hold semantic meaning or refer to empirical phenomenon" (Andrews 2025, 2694). Yet once this distinction separates the model from the world, we are left with a question about how the ML pipeline can say anything about the target domain.

As I stated, architecture selection may be informed by domain knowledge. But while the successful choice of $\mathcal{F}$ can inform us of the structural dependencies within the data, it fails to tell us what the data are actually about. For instance, the successful application of convolutional neural networks (CNNs) in a given task suggests that the relevant information in the input exists in spatially local patterns. But it says nothing about whether the input data come from animal images or audio spectrograms, although this information decides the target domain of the pipeline. Instead, I argue that what the pipeline claims about the target domain originates from the activity of *dataset collection*, whose inherent aim is to create digital representations of the quantities measured from the world. This activity is often executed without ML modelling in mind, as when researchers merely upload the measurement data to a database. When this dataset is utilised in an ML pipeline, the external function of this activity is to provide an *adequate* dataset for the aim of the ML pipeline.

Before defining this adequacy, let me first discuss the constituent activities of dataset collection. These activities are far more heterogeneous compared to model training. Each variable in the dataset should be considered a *measurement* of the quantity of interest, and as such, the activities are individuated by the measurement procedure up to empirically accepted equivalence. For instance, if air pressure appears in the dataset, then *air pressure measurement* becomes a constituent activity. The inherent aim of this activity is to produce values that can serve the operationally coherent activities in which "air pressure at region X" appears. In practice, this can be satisfied by following the relevant protocols for measuring air pressure, whose codified form appears in engineering standards.

[14] At this point, Ratti et al. introduce a tripartite typology of artefact functions (adapted from Vermaas (2009) and Van Eck (2015)), distinguishing between *behaviour function*, *effect function*, and *purpose function*. But it is unclear what justificatory work this distinction does in this context.

[15] Note that, as in the climate model example presented below, the "world" may be the output of a separate simulation model.

For this example, "empirically accepted equivalence" refers to the different ways of measuring the air pressure that scientists and engineers have accepted as equivalent within a given situation. To illustrate, two different types of pressure meters are taken to measure the "same" quantity, as long as the target air pressure lies within their ranges of validity and they are correctly calibrated. The emphasis on empirical acceptance is in accordance with Bridgman's dictum that "the practical justification for retaining the same name [for a measured quantity] is that within our present experimental limits a numerical difference between the results of the two sorts of operations has not been detected" (Bridgman 1927, 16). In other words, the equivalence of two measurement operations is something that must be empirically validated, which is a claim that can be accepted without the problematic semantic doctrine that is often attached to operationalism.

This operational approach allows us to diagnose certain problems that arise from using "big data" with heterogeneous provenance, which are often fed to ML models for training. Leek et al. have pointed to "batch effects" in high-throughput biology, which are "subgroups of measurements that have qualitatively different behaviour across conditions and are unrelated to the biological or scientific variables in a study" (Leek et al. 2010, 733), commonly correlated with the differences in lab group or date of experiment. If we bracket the operational information and simply take each variable to be measuring the "same target", then we become blind to the potential sources of variation—such as ambient temperature and reagent quality—that mislead the outcomes of subsequent analyses. To counter the batch effect, they specify that "[i]nformation about changes in personnel, reagents, storage and laboratories should be recorded and passed onto data analysts" (Leek et al. 2010, 737), in addition to statistical analysis and careful experimental design. This means that all operational information relevant to the purpose should be attended to and recorded, in order to make different measurements as closely empirically equivalent as possible while correctly accounting for the residual discrepancies.

In tandem with tracking operational provenance, *certification* is an activity that is crucial in this context. Even when a variable measures a real-world quantity, this quantity need not necessarily correspond to a theoretical quantity like the air pressure[16]. That requires relevant certification to be in place, which justifies our belief that the measurements are indeed tracking pressure. In some scientific contexts, this certification process goes by the name of establishing the *construct validity* of certain measurement operations (Cronbach and Meehl 1955; Coston et al. 2023). While it is beyond the scope of this paper to describe this process in detail (see e.g. Tal 2017 for a model-based account), we can note that certification *qua* activity is not simpler than measurement itself: as Chang notes, even for weighing objects with a balance pan, certification involves "ordering the weights from a reliable supplier, or comparing them to a more trusted set of weights, or checking them against certain natural phenomena" (Chang 2022, 38).

As a concrete illustration, consider the case study provided by Mussgnug (2022), who discusses an ML model that purported to predict poverty. Since "poverty" is a theoretical

[16]Note that this certification process is not *necessary* for the variable to be usable in an ML pipeline. Instead, the variable can simply be described in a purely operational manner: "I have placed this pressure meter in such-and-such a location and recorded the output voltage with a voltmeter". The resulting ML model trained on such data could be sufficient if it is only used instrumentally to predict within a fixed setup. That is, such data *can* be adequate-for-purpose, depending on the aims of the system of practice.

construct, the researchers operationalised its magnitude by a proxy quantity extracted from phone survey results. However, no explanations were provided as to why this quantity would track the concept of "poverty" as theoretically understood. As Mussgnug argues, while we can say that this ML model was *calibrated* using this phone survey data, we are not justified in saying that it correctly *measures* poverty (due to the lack of certification). Yet this criticism of the model does not make sense unless we look *beyond* the model training process, and consider the aim of the ML pipeline (do we have a model that relates measurable variables to poverty?) and furthermore the aims of the system of practice (how do we describe, understand and intervene on poverty?). After all, the model is good at predicting an operationally well-defined variable, and it would indeed be useful if we are interested in this proxy for some other reason. Under the framework developed above, we may precisely state what is wrong about the model. We see that the inherent aims of dataset collection and model training were satisfied, for (presumably) the quantity was well-measured and the model was trained in an acceptable manner. The issue arises in their coordination towards the aim of the integrated activity (ML pipeline); it is here where the researchers' actions become incoherent, because the aim asks for a certified variable that tracks poverty but the dataset collection activity does not provide one[17].

This activity-based analysis of dataset collection clarifies in which sense ML models function as *epistemic artefacts* (Kieval 2026), which are defined as "external tools for thinking, the construction and manipulation of which are crucial to their epistemic functioning" (Knuuttila 2011, 263). An example of computational epistemic artefact is a natural language parser, which is described in Knuuttila and Voutilainen (2003). Can an ML model be construed as an epistemic artefact? It can be, but only when we view the model as the material output of the entire ML pipeline. Knuuttila notes that much linguistic (i.e. domain) knowledge can be gained by attempting to construct a parser, starting with how ambiguous natural language often is. Similarly, when a researcher builds an ML pipeline based on the aim they wish to achieve, they can learn a lot about the target domain. As mentioned previously, creating an adequate dataset requires them to grapple with whether their question is sufficiently aligned with the collected data, which in turn clarifies what the question actually meant in the first place. So ML models satisfy the *construction*-side requirement to be an epistemic artefact[18].

### 4.2 Determining adequacy of data

With this operational background in mind, I can now define the notion of data adequacy. I start from the account provided by Bokulich and Parker, who state that "a dataset or data

[17]There is another common source of incoherence, which has been covered by other philosophers (Freiesleben and Grote 2023). It might be the case that the statistical distribution of the variables shifts from that of $D_{\text{test}}$ when the model is deployed. In this case, the inherent aim of model training decouples from its external function of providing a model that works on the deployment distribution. Other activities are required to validate that such *distribution shifts* do not occur under deployment.

[18]After epistemic artefacts are constructed, Knuuttila writes that they become "open-ended things that have their own history and dwell in our research practices in manifold ways as both tools and objects of inquiry" (Knuuttila and Voutilainen 2003, 1494). A key to this open-endedness of an artefact is its manipulability as a materially embodied object. Through the affordances provided by the material medium, the manipulation of an artefact is supposed to give us epistemic access to the target domain. Whether this characterisation applies to ML models is a nuanced matter, and is further discussed in a footnote of Section 5.

model $D$ is adequate$_I$-for-$P$ just in case the use of $D$ in instance $I$ would (or would be very likely to) result in the achievement of $P$" (Bokulich and Parker 2021, 31). Henceforth, I drop the subscript $I$, taking it to be the relevant system of practice. The purpose $P$ is the inherent aim of the ML pipeline mentioned above: the creation of a model that relates real-world quantities. For this $P$ to be achieved, we need to ascertain that the data and the corresponding operational provenance satisfy the following two requirements:

1. *Discrimination*: The data, given their operational provenance, must be such that the states considered distinct by the aim can be discriminated;
2. *Traceability*: As for the differences in the data that do not correspond to states taken to be distinct by the aim, they must be traceable and accounted for based on our knowledge of the operational provenance.

In Mussgnug's example, the operations involved in collecting the target variable were such that the degree of poverty could not be *discriminated* on the basis of the data. In the batch effect, a spurious distinction that does not appear in the biological claim was *traced* to the lab-to-lab and other sources of variability, and needed to be accounted for with operational knowledge. Without suitable operational provenance, these data would have been inadequate.

The enumeration of the relevant states is usually very theory-laden[19]: the different degrees of poverty, or the different qualities of the chemical reagent. But once the states have been enumerated, many of the procedures required for establishing adequacy involve what Bogen and Woodward call assessing the *reliability* of data for inferring phenomena, for which they argue "the construction of systematic explanations of facts about such data" (Bogen and Woodward 1988, 327) is not needed. Consider Woodward's (1989) example, where scientists introduced thick iron shielding in their neutral current detection experiment (which we may note is a physical operation on the equipment), because their theoretical knowledge was insufficient to compute the spurious effect of hadron punchthrough. While the justification of such shielding requires *some* theory, the theory need not fully explicate the effect of shielding. This contrasts with the states "neutral current exists" and "neutral current does not exist", whose formulation requires substantial theory. Likewise, scientists only need to have good reasons to believe that the data are adequate-for-purpose, without deriving full guarantees based on theoretical grounds. Of course, the establishment of adequacy is fallible, like all other scientific judgements.

In order to adjudicate adequacy, the operational provenance may include the following pieces of information:

1. Production: Which instruments, surveys, or procedures generated the values, and what are their certifications (if necessary)?
2. Sampling: From which population, and under which conditions, were the samples obtained?
3. Processing: How were the samples cleaned and transformed?

What I propose is in line with Boyd's (2018) proposal that the empirical adequacy of theories should be decided upon "enriched lines of evidence" that includes "a sequence of

[19] An exception is when we are working with purely operational variables.

empirical results including the records of data collection and all subsequent products of data processing generated on the way to some final empirical constraint" along with "the metadata regarding the provenance of the data records and the processing workflow that transforms them" (Boyd 2018, 406–7). In other words, the inferences warranted by a piece of data are largely derived from the operations that produced it. As Boyd also notes, this position is also broadly in agreement with Leonelli's (2009, 2016) view of "metadata".

Finally, let us note a potential danger with epistemic artefactualism. By characterising ML models as artefacts, it is implied that they can function as independent epistemic tools. For instance, Kieval approvingly cites Morgan and Morrison (1999) in writing that "models can take on 'a life of their own' as material mediators between theory and observation" (Kieval 2026, 263). In one sense, it is trivially true that a mathematical function $f_{\hat{\theta}}$ is an independent entity. But we cannot say that it is epistemically independent, as its warrant lies in the pipeline. A similar danger is noted by Boyd (2018) when she gently objects against Leonelli's (2016) use of the term "decontextualisation" in describing how data can travel. A model alone is no more inherently meaningful about the world than a sine function is.

## 5 Dealing with opacity

Despite the widespread interest, there remain philosophical disagreements on what it exactly means for an ML model to be (epistemically) *opaque*, beyond the fact that it relates to the bounded cognitive capacity of humans (Humphreys 2009). Recently, Beisbart (2026) suggested a typology of opacity in ML as the difficulties in attaining the following epistemic states:

1. Knowing the training data.
2. Knowing the relations among variables and understanding the workings of the model.
3. Predicting the behaviour/the outputs of the models.
4. Having a sense of/understand why an outcome was produced.
5. Pinning down mistakes if results are wrong.
6. Understanding the system that is investigated.

While the epistemic states corresponding to (2)-(4) can be differentiated, I will group them under a single heading of (2), because they all arise from the agent's lack of cognitive access to the input-output relation. In what follows, I explain how these issues are separate, and to which extent they are problematic.

We see that (2)-opacity and (5)-opacity pertain to models. Let us begin with (2)-opacity, which is what one usually means with the statement "ML models are opaque" (see e.g. Boge 2022). Søgaard (2023) pins down the reason for this unknowability of the input's influence on the output to the *model size*, i.e. the number of components in the parameter vector $\theta$. He argues the model size leads to (2)-opacity because it prevents humans from holding these parameters in memory and thereby reconstruct the inferential process of the model. (While this kind of opacity is often referred to as the "black box"-ness of a complex ML model, I believe this analogy is unhelpful. As a mathematical function, an ML model is fully inspectable by the programmer. As an analogy, an unopenable watch is epistemically different from a watch with very intricate but inspectable machinery.)

In face of (2)-opacity, we should recall that the inherent aim of model training is to find a generalisable model that relates the inputs-as-numbers to the outputs-as-numbers. (2)-opacity neither hinders this aim, nor does it create problems for the constituent activities of search and evaluation. It also leaves the coherence of the ML pipeline intact. If this opacity is to pose a problem, it must be interfering with the external function that the *system of practice* is requiring of the ML pipeline. That is, a large model size is epistemically problematic only if the aims of the system require us to find a *humanly graspable* relationship between the input and the output. But this is not always the case, and it is often justified to prioritise predictive accuracy over this graspability, even in scientific contexts.

For example, consider the ML models proposed for use in climate science, as described by Lusk (2026). One of the tasks in climate science is *downscaling*, where the coarse-grained information from a general circulation model (GCM) is used as the boundary conditions for the higher-resolution regional climate model (RCM). There are two main approaches to this procedure: *dynamical downscaling* that uses physical simulation, and *statistical downscaling* that uses empirical statistical relationships. Against these approaches, scientists have proposed training neural network models on the data obtained from dynamical downscaling. The appeal is that it may provide a way to combine the computational efficiency of the statistical approach with the accuracy of the dynamical approach, especially with respect to the non-stationarity of future climate projections.

This gain in computational efficiency is not merely a practical matter, for it allows the scientists to run more simulations within the available timeframe. When the expensive dynamical downscaling is the only option, we are constrained to using a single "best guess" as the justification for decision; multiple simulations allow better grasp of the uncertainties involved. We see that the opacity of the ML model does not hinder its scientific use in this case, provided that the suitable validation procedures are in place. In particular, it shows that opaque models need not necessarily be restricted to a heuristic role. This view contrasts with that of Duede, who argues that opaque models should only "be used to guide attention and scientific intuition toward more promising hypotheses" (Duede 2023, 1094), and cannot be used for *justifying* them. However, the use of neural network downscaling is an acceptable method for obtaining knowledge about the uncertainties of climate projections.

So a complex model does not necessarily signal a defect of the modelling procedure that disqualifies its role in the scientific process. After all, the relationship between the input and the output *could* be irreducibly complex[20], and simpler models may distort the relationship beyond an acceptable level. Instead of focusing solely on the model, determining the appropriateness of an ML model requires consideration of the system of practice's aims. At times, graspability is indeed a paramount consideration[21]; but what we often need in

[20]However, some (2)-opacity may represent *needless* complexity: many situations allow for simplifications after further efforts by the researchers (Rudin 2019).

[21]Speaking of graspability, it would be remiss to ignore the recent literature on *mechanistic interpretability*, which aims to "provid[e] human-intelligible descriptions of the functional modules that drive the emergence of model behaviors" (Millière and Buckner 2024, 10). Some promising results (Schut et al. 2023; Simon and Zou 2024) hint at the possibility that we may retrieve intelligible knowledge about the target domain by these methods, although it would be inappropriate to adjudicate on this matter in this paper. If successful, we would be extracting graspable knowledge directly from the opaque model.

Mechanistic interpretability has another philosophical import: as it involves manipulating ML models to gain knowledge about the target domain, its success would turn ML models into full-blooded epistemic artefacts.

those cases is empirical investigation aimed at discovering data that relate to the phenomena in a simpler manner, which lies outside the ML pipeline.

Moving on to (5)-opacity, there are two possible senses in which this issue can be understood. The first sense asks *how* we should fix a broken model, and this relates to (2)-opacity: because the input-output relation is so complex, we cannot locally revise the model to cover the cases that the model gets wrong. If the mistakes accumulate beyond what can be tolerated, the main option is to retrain the model, which may induce wholesale changes to the function $f_{\hat{\theta}}$[22]. Again, whether this is problematic depends upon the wider context. In the second sense, we are asking *why* the model is not performing as desired. This question is what the activity-based analysis helps us answer, as it allows us to localise the failure to a decoupling between our goals and our activities.

Next, we see that (1)-opacity is not concerned with the model at all, but rather the dataset. However, it is unclear what "knowing the data" consists of. After all, it cannot mean memorising the data. My interpretation is that it refers to understanding the operational provenance of the dataset as described previously. If we do not sufficiently understand how the dataset was collected, then we are prevented from assessing *Discrimination* and *Traceability*, which is required to determine whether $P$ (the aim of the ML pipeline) has been achieved. So (1)-opacity is a matter of absent information, which is an absolute state of affairs. Meanwhile, (2)-opacity arises from the complexity relative to the bounded human cognition.

Finally, (6)-opacity is a failure of achieving an aim (understanding) of the system of practice through the ML pipeline activity, which points to incoherent coordination of this activity.

In summary, we see that the issue of opacity can be decomposed as follows:

1. Objective access to the data and their operational provenance (1);
2. Cognitive access to the trained model (2);
3. Diagnosis and repair of the pipeline (5);
4. Achieving understanding through scientific activity (6).

Opacity is not a blanket verdict on the epistemic validity of scientific ML. Rather, it identifies various limitations in the pipeline, whose importance depends on the extent to which they interfere with the inherent aims of the activities (and of the system of practice) and the coordination of external functions. In particular, the model opacity *qua* complexity is only a part of what we mean by "opacity", and it is not always problematic. This de-emphasis on model opacity is shared by computational reliabilism described in Durán and Formanek (2018) and Durán (2026). I return to computational reliabilism in Section 7.

Regardless of the final verdict on mechanistic interpretability, we see that epistemic artefactualism provides a useful angle through which we can organise our thoughts about ML, and as such I take it to be complementary to my account.

[22]This reveals a crucial difference between an ML model and a toy model (cf. Sullivan 2024). While a toy model has identifiable components and can be either de-idealised or be provided with some phenomenological corrections, such progressive corrections are mostly not possible for ML models.

## 6 Link uncertainty

Emily Sullivan defines *link uncertainty* as "the extent to which the model fails to be empirically supported and adequately linked to the target phenomena" (Sullivan 2022, 110). Her thesis is that link uncertainty is what prevents us from achieving understanding through ML models, and not opacity. But taken at face value, it is not clear if link uncertainty is what causes epistemic trouble for ML models. As Jensen (2026) also points out, ML models seem to have transparent empirical references because they are trained on empirical variables. So it is difficult to see in what manner the model is not empirically supported.

Instead, let us try to reconstruct what link uncertainty could mean by considering her examples. She states that the melanoma classifier trained on the photos of moles (Esteva et al. 2017) has small link uncertainty, while the classifier for self-identified sexual orientation trained on facial photos (Wang and Kosinski 2018) has great link uncertainty. In both cases, the trained models were able to discriminate the classes well. Despite this parity, Sullivan's contention is that, while there is extensive "background knowledge linking the appearance of moles to instances of melanoma" (Sullivan 2022, 126), scientific evidence fails to suggest the existence of "a link between facial features and origins of sexual orientation" (Sullivan 2022, 128). Consequently, she argues that the latter model fails to provide how-actually explanations of the phenomenon (which her paper equates with gaining understanding).

However, it is not certain in what sense the melanoma model can provide a how-actually explanation, even with the background knowledge. If the fault with Wang and Kosinski's model is that it "is not able to provide understanding of how innate facial features reflect sexual orientation" (Sullivan 2022, 128), then a similar charge can be raised against the melanoma model[23], for it does not help us understand what the underlying causes of melanoma are.

The framework developed in this paper can help clarify this situation. If the system of practice aims at understanding the causes of melanoma, then the activity of constructing a model that classifies moles based on external images is incoherent with respect to its aim. It is widely understood that the appearance of the mole is a symptom of the disease, not its cause. It is analogous to asking a handwritten digit classifier *why* a certain number was written, which is clearly outside what the model can answer. But Esteva et al. do not claim that their model allows understanding of melanoma[24]. They state that their model "may allow medical practitioners and patients to proactively track skin lesions and detect cancer earlier" (Esteva et al. 2017, 115). For such an aim, creating a model that relates external appearances of moles to types of melanomas is far more coherent.

Taking our focus away from understanding allows us to see more clearly the sense in which Wang and Kosinski's model is epistemically inferior to Esteva et al.'s one. As with Esteva et al., their model cannot answer questions about the causes of sexual orientation. But Wang and Kosinski's explicitly stated aim is to alert people to the potential invasion of privacy under the progress of ML. For that purpose, it appears that the model is quite effective, because such alerts only require displaying the possibility of danger.

[23] As Jensen (2026) notes, Sullivan equivocates between two targets of explanation: the medical professionals' ability to diagnose diseases, and the disease development. I take it that Sullivan's intended target is the latter, as Jensen does.

[24] As I see it, ML models can have scientific value even when they do not provide *understanding*.

However, *if* the aim is to show that sexual orientation is determinable from the morphological factors of the face, then there are diagnosable issues from the operational point of view. On the model output side, we note that the "sexual orientation" of each sample is obtained by matching the gender of the person with that of their sought partner in their dating profile. Miller (2018) argues such operationalisation based on self-identification and binary framing is misleading in light of current understanding of sexual orientation. This is the failure of *Discrimination*, analogous to that displayed in Mussgnug (2022). On the input side, several issues have been raised: the dating site population may not be representative of the general population (for instance, it only includes sexual minorities who have come out), there might be systematic differences in cultural grooming patterns, or in head orientation (Leuner 2019). In consequence, there is great doubt as to whether the collected dataset is adequate for the purpose of discovering biological origins of sexual orientation, due to failures in both *Discrimination* and *Traceability*.

It should be noted that there are input-side problems in melanoma classification as well. As Winkler et al. (2019) have found, these models respond to the external markings that clinicians make on the skin adjacent to the mole. This is problematic because it interferes with the stated aim of Esteva et al., for which patients should be able to track lesions on their own. Nonetheless, the output side involves fairly uncontested classifications based on histopathological investigations. (Such investigations should be viewed as certification in the sense described above.) Furthermore, rather minimal understanding of dermatology allows us to know that such marking should be irrelevant to classification, and we know how to remove this type of contamination from the dataset: by cropping the markings from the image. But in Wang and Kosinski's model, such separation is not possible to the extent that we cannot distinguish the effects of biology and culture from a facial photograph. So in the former case, the failure in *Traceability* is remediable, but the latter is not.

Under my framework, we can re-interpret link uncertainty as follows. As per the preceding discussion, I take link uncertainty to pertain to the activity of dataset collection. This contrasts with Sullivan, who writes of link uncertainty as a property of the model. Then we may say that link uncertainty corresponds to both

1. the degree to which we doubt the adequacy of the dataset for the purpose, and
2. the degree to which we are uncertain how to reduce this doubt by intervening in the dataset.

Since knowing how to reduce doubts presupposes that we can identify the doubt, a reduction in (2) will lead to a potential reduction in (1). For instance, further understanding of the cultural effects on face morphology will help us control for these factors in the dataset. In contrast, a reduction in (1) could simply be due to a careful application of our existing knowledge. For example, we see that Winkler et al.'s result increases link uncertainty in the sense of (1) for Esteva et al.'s model, but leaves (2) untouched.

Another two-tiered clarification of link uncertainty is provided by Tamir and Shech (2022), who argue that Sullivan conflates *empirical link uncertainty*, which simply follows Sullivan's concept above, and *misrepresentation uncertainty*, which refers to the "unintended misrepresentation of target features by the data" (Tamir and Shech 2022, 334). While this disambiguation can be helpful in traditional scientific modelling, there are a number of problems when it is applied to ML modelling. Firstly, as per Jensen's objection noted

above, empirical link uncertainty is seemingly inapplicable to ML models. Tamir and Shech also appear to be aware of this problem, as they only apply this notion to the Schelling model of segregation. Secondly, they do not clarify the concept of intention involved in misrepresentation uncertainty. My account can explicate this intention as that of achieving the aims of the system of practice. Finally, Tamir and Shech make *representation* fundamental to link uncertainty, but representation remains a contentious issue in philosophy of ML and beyond. By focusing on the *use* of data, my account can bracket this issue.

## 7 Computational reliabilism

Under my framework, the epistemic warrant of a scientific process is obtained from the *activities* external to the ML model itself, and so the opacity of the model is no longer integral to the warrant. This move is similar to computational reliabilism (CR), which does not "depend on revealing the internal mechanisms and properties of the algorithm", instead relying on the "reliability indicators that are external to it" (Durán 2026, 59). Therefore, readers may question if there is any difference between the two. To discuss this issue, let me start by describing the three types of reliability indicators (RIs), which according to the latest formulation are defined as (Durán 2026, 63–64):

- $\text{Type}_1$-RI, technical performance of algorithms: "focuses on the specification, coding, execution, maintenance, and other technical features that contribute to the performance of the algorithm".
- $\text{Type}_2$-RI, computer-based scientific practice: "focuses on securing algorithmic-based scientific research ... intends to capture the degree to which scientific units of analysis are implemented and operationalized into the algorithm".
- $\text{Type}_3$-RI, social construction of reliability: "focuses on broader goals related to the acceptance—or rejection—of algorithms and their outputs by diverse communities".

At the outset, I should state that this paper provides a practice-based account of ML use, and should not be seen as replacing a more traditional epistemological account that Durán provides. As I see it, my account proposes an answer to the invitation made by Durán, when he states that "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 ... these issues will not find a complete answer here ... 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" (Durán 2026, 65). By individuating inherent aims and the issues of coordination, my account can both localise the defective token-RI and delimit its epistemic damage.

For instance, consider Mussgnug's example. As the target variable (phone survey results) has no theoretical basis, CR would say that there is a problem with $\text{Type}_2$-RI, such that scientists are not justified in taking the outputs of the resulting ML model at face value. My account can subsequently supply what the problematic token-RI is, namely the lack of certification procedures for establishing construct validity. Recall that I have stated this lack of certification is only fatal to the pipeline if its inherent aim is to predict poverty *qua* theoretical construct, and could be defensible under other use-cases. I take this proviso

to be in line with Durán's observation that "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 epistemic and non-epistemic values and goals at stake" (Durán 2026, 65). In fact, while discussing the case of a drug risk-scoring model, he lists the potential RIs as "design provenance (i.e., information about how, why, and for whom the algorithm was constructed), epistemic stratification (e.g., monitoring whether the algorithm preserves relevant epistemic distinctions, such as different individuals taking different drugs)" (Durán 2025, footnote 10), which are the factors that my account systematically distinguishes.

Additionally, my account may also help with clarifying the type-RIs. Note that $\text{Type}_2$-RI refers to the factors that reduce link uncertainty as understood by Sullivan: "$\text{Type}_2$-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" (Durán 2026, 68). He states that these scientific units are "implemented and operationalized into the algorithm" (Durán 2026, 64), but I have pointed out that data adequacy is the focal point for domain expertise, and not the model nor the training algorithm. However, data-related indicators are explicitly categorised as $\text{Type}_1$[25], which suggests that the typology may require refining. Meanwhile, it is unclear what separates $\text{Type}_3$-RIs from $\text{Type}_2$-RIs, insofar as both measure adherence to scientific standards; and any stronger reading of the phrase "social construction of reliability" beyond agreed scientific standards raises the familiar objections of constructivism. My account separates the inherent aim of the ML pipeline from the aims of the surrounding system of practice, and the reliability of the pipeline becomes a different matter from whether the broader scientific aims are satisfied through the external function of the pipeline; subsequently, we can limit $\text{Type}_3$-RI to the latter issue only.

In summary, CR provides the larger epistemological framework in which we may safely assert that the justifications for relying on an ML model's output lie outside the model; my account then describes what these justifications consist of.

## 8 Conclusion

In this paper, I have argued that an activity-based analysis can provide insights into the epistemic validity of ML practices. An ML pipeline should be considered as an integrated epistemic activity consisting of various activities with their own aims. The successful completion of the activities is necessary for the pipeline's aim to be satisfied, but it is not sufficient; in addition, a coherent coordination of the external functions of these activities is necessary. Whether the pipeline helps us achieve the broader scientific aims is distinct from the achievement of the pipeline's inherent aim, and should be evaluated separately. Additionally, I have displayed how this framework can clarify the philosophical debates surrounding ML:

1. Opacity in ML arises from different aspects of the ML pipeline, and model complexity is only a part of the problem (Beisbart 2026).
2. Link uncertainty is not a property of an ML model, but rather that of dataset collection.

[25] "[T]ypical cases of $\text{token}_1$-RI include practices and protocols for collecting, curating, storing, distributing, and analyzing data" (Durán 2026, 63).

3. Opaque models are not confined to a heuristic role, and climate downscaling shows how they may be justificatory.
4. The limitation of technical artefactualism arises from its artefact-centric nature, and overcoming it requires consideration of activities outside model search.
5. The token-level reliability indicators of computational reliabilism can be described through activity-based analysis.

Finally, the arguments made in the paper can be considered as evidence that activity-based analysis can be put to "productive use by those who want to understand and promote good practices relating to knowledge" (Chang 2022, 8). I have shown that this framework can be applied not just to historical case studies, but also to methodological debates at the forefront of scientific practice. Apart from the timeliness, ML provides an especially interesting case for activity-based analysis because the individual activities are *engineered* and their inherent aims are explicitly formalised by the practitioners, unlike in other types of scientific practice where they have to be rationally reconstructed by the philosopher. The application has also revealed subtleties in the existing account; the test set reuse shows how the external function of an activity can become degraded by a mindlessly repeated satisfaction of its inherent aim.

A limitation of this paper is the lack of discussion of LLMs, despite their growing influence on the scientific practice. The construction of an LLM involves more steps than the general scientific ML models, which provides richer opportunities for an activity-based analysis. Furthermore, there is greater uncertainty as to what the aim of an LLM pipeline should be in the first place, because we are no longer directly predicting the quantity of interest. Clarifying the epistemic role of LLMs within this framework is left to future work.

## References


Amershi, Saleema, Andrew Begel, Christian Bird, et al. 2019. "Software Engineering for Machine Learning: A Case Study." *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)* (Montreal, QC, Canada), May, 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042.

Andrews, Mel. 2025. "The Immortal Science of ML: Machine Learning and the Theory-Free Ideal." *Erkenntnis*, ahead of print, September 25. https://doi.org/10.1007/s10670-025-01010-x.

Beisbart, Claus. 2026. "In Which Ways Is Machine Learning Opaque?" In *Philosophy of Science for Machine Learning*, edited by Juan M. Durán and Giorgia Pozzi, vol. 527. Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-03083-2_1.

Belkin, Mikhail, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. "Reconciling Modern Machine Learning Practice and the Bias-Variance Trade-Off." *Proceedings of the National Academy of Sciences* 116 (32): 15849–54. https://doi.org/10.1073/pnas.1903070116.

Boge, Florian J. 2022. "Two Dimensions of Opacity and the Deep Learning Predicament."

*Minds and Machines* 32 (1): 43–75. https://doi.org/10.1007/s11023-021-09569-4.

Bogen, James, and James F. Woodward. 1988. "Saving the Phenomena." *The Philosophical Review* 97 (3): 303. https://doi.org/10.2307/2185445.

Bokulich, Alisa, and Wendy Parker. 2021. "Data Models, Representation and Adequacy-for-Purpose." *European Journal for Philosophy of Science* 11 (1): 31. https://doi.org/10.1007/s13194-020-00345-2.

Boyd, Nora Mills. 2018. "Evidence Enriched." *Philosophy of Science* 85 (3): 403–21. https://doi.org/10.1086/697747.

Bridgman, P. W. 1927. *The Logic of Modern Physics*. Macmillan Paperbacks. Macmillan. https://books.google.co.kr/books?id=zGUGAQAAIAAJ.

Buchholz, Oliver, and Eric Raidl. 2025. "A Falsificationist Account of Artificial Neural Networks." *The British Journal for the Philosophy of Science* 76 (4): 1011–35. https://doi.org/10.1086/721797.

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

Chang, Hasok. 2022. *Realism for Realistic People: A New Pragmatist Philosophy of Science*. Cambridge University Press. https://doi.org/10.1017/9781108635738.

Coston, Amanda, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. 2023. "A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms." *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (Raleigh, NC, USA), February, 690–704. https://doi.org/10.1109/SaTML54575.2023.00050.

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

Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52 (4): 281–302. https://doi.org/10.1037/h0040957.

Duede, Eamon. 2023. "Deep Learning Opacity in Scientific Discovery." *Philosophy of Science* 90 (5): 1089–99. https://doi.org/10.1017/psa.2023.8.

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

Durán, Juan M. 2026. "Beyond Transparency: Computational Reliabilism as an Externalist

Epistemology of Algorithms." In *Philosophy of Science for Machine Learning*, edited by Juan M. Durán and Giorgia Pozzi, vol. 527. Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-03083-2_4.

Durán, Juan M., and Nico Formanek. 2018. "Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism." *Minds and Machines* 28 (4): 645–66. https://doi.org/10.1007/s11023-018-9481-6.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, et al. 2017. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." *Nature* 542 (7639): 115–18. https://doi.org/10.1038/nature21056.

Facchini, Alessandro, and Alberto Termine. 2022. "Towards a Taxonomy for the Opacity of AI Systems." In *Philosophy and Theory of Artificial Intelligence 2021*, edited by Vincent C. Müller, vol. 63. Springer International Publishing. https://doi.org/10.1007/978-3-031-09153-7_7.

Fama, Eugene F. 1970. "Efficient Capital Markets: A Review of Theory and Empirical Work." *The Journal of Finance* 25 (2): 383. https://doi.org/10.2307/2325486.

Freiesleben, Timo, and Thomas Grote. 2023. "Beyond Generalization: A Theory of Robustness in Machine Learning." *Synthese* 202 (4): 109. https://doi.org/10.1007/s11229-023-04334-9.

Hansen, Jens Ulrik, and Paula Quinon. 2023. "The Importance of Expert Knowledge in Big Data and Machine Learning." *Synthese* 201 (2): 35. https://doi.org/10.1007/s11229-023-04041-5.

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

Hurshman, Clint. 2024. "Do Opaque Algorithms Have Functions?" *Synthese* 204 (3): 91. https://doi.org/10.1007/s11229-024-04745-2.

Jensen, Sara Pernille. 2026. "Uncertainties about Link Uncertainty: ML Models as Phenomenological Models." *Synthese* 207 (3): 111. https://doi.org/10.1007/s11229-026-05509-w.

Kieval, Phillip Hintikka. 2026. "Representation Learning Without Representationalism: A Non-representational Account of Deep Learning Models in Scientific Practice." In *Philosophy of Science for Machine Learning*, edited by Juan M. Durán and Giorgia Pozzi, vol. 527. Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-03083-2_12.

Knuuttila, Tarja. 2011. "Modelling and Representing: An Artefactual Approach to Model-Based Representation." *Studies in History and Philosophy of Science Part A* 42 (2): 262–71.

https://doi.org/10.1016/j.shpsa.2010.11.034.

Knuuttila, Tarja, and Atro Voutilainen. 2003. "A Parser as an Epistemic Artifact: A Material View on Models." *Philosophy of Science* 70 (5): 1484–95. https://doi.org/10.1086/377424.

Leek, Jeffrey T., Robert B. Scharpf, Héctor Corrada Bravo, et al. 2010. "Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data." *Nature Reviews Genetics* 11 (10): 733–39. https://doi.org/10.1038/nrg2825.

Leonelli, Sabina. 2009. "On the Locality of Data and Claims about Phenomena." *Philosophy of Science* 76 (5): 737–49. https://doi.org/10.1086/605804.

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study*. The University of Chicago Press.

Leuner, John. 2019. "A Replication Study: Machine Learning Models Are Capable of Predicting Sexual Orientation From Facial Images." Version 1. Pre-published. https://doi.org/10.48550/ARXIV.1902.10739.

Lusk, Greg. 2026. "Artificial Intelligence in Climate Science: From Machine Learning to Neural Networks." In *Philosophy of Science for Machine Learning*, edited by Juan M. Durán and Giorgia Pozzi, vol. 527. Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-03083-2_21.

Miller, Arianne E. 2018. "Searching for Gaydar: Blind Spots in the Study of Sexual Orientation Perception." *Psychology & Sexuality* 9 (3): 188–203. https://doi.org/10.1080/19419899.2018.1468353.

Millière, Raphaël, and Cameron Buckner. 2024. "A Philosophical Introduction to Language Models - Part II: The Way Forward." Pre-published May 6. https://doi.org/10.48550/arXiv.2405.03207.

Morgan, Mary S., and Margaret Morrison, eds. 1999. *Models as Mediators: Perspectives on Natural and Social Sciences*. Ideas in Context. Cambridge University Press.

Mussgnug, Alexander Martin. 2022. "The Predictive Reframing of Machine Learning Applications: Good Predictions and Bad Measurements." *European Journal for Philosophy of Science* 12 (3): 55. https://doi.org/10.1007/s13194-022-00484-8.

Ratti, Emanuele, James Ladyman, and Alberto Termine. 2026. "The Modelling Structure of AI-Based Science." July 29. https://philsci-archive.pitt.edu/30610/.

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

Regt, Henk W. de. 2017. *Understanding Scientific Understanding*. Oxford Studies in Philosophy of Science. Oxford University Press. https://doi.org/10.1093/oso/9780190652913.001.0001.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1 (5): 206–15. https://doi.org/10.1038/s42256-019-0048-x.

Schut, Lisa, Nenad Tomasev, Tom McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. 2023. "Bridging the Human-AI Knowledge Gap: Concept Discovery and Transfer in AlphaZero." Pre-published October 25. https://doi.org/10.48550/arXiv.2310.16410.

Simon, Elana, and James Zou. 2024. "InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders." Pre-published November 13. https://doi.org/10.48550/arXiv.2412.12101.

Søgaard, Anders. 2023. "On the Opacity of Deep Neural Networks." *Canadian Journal of Philosophy* 53 (3): 224–39. https://doi.org/10.1017/can.2024.1.

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

Sullivan, Emily. 2024. "Do Machine Learning Models Represent Their Targets?" *Philosophy of Science* 91 (5): 1445–55. https://doi.org/10.1017/psa.2023.151.

Tal, Eran. 2017. "Calibration: Modelling the Measurement Process." *Studies in History and Philosophy of Science Part A* 65–66 (October): 33–45. https://doi.org/10.1016/j.shpsa.2017.09.001.

Tamir, Michael, and Elay Shech. 2022. "Understanding from Deep Learning Models in Context." In *Scientific Understanding and Representation*, 1st ed., edited by Kareem Khalifa, Insa Lawler, and Elay Shech. Routledge. https://doi.org/10.4324/9781003202905-28.

Termine, Alberto, Emanuele Ratti, and Alessandro Facchini. 2026. "Machine Learning and Theory-Ladenness: A Phenomenological Account." *Synthese* 207 (3): 94. https://doi.org/10.1007/s11229-026-05454-8.

Van Eck, Dingmar. 2015. "Mechanistic Explanation in Engineering Science." *European Journal for Philosophy of Science* 5 (3): 349–75. https://doi.org/10.1007/s13194-015-0111-3.

Vermaas, PE. 2009. "The Flexible Meaning of Function in Engineering." In *Proceedings of the International Conference on Engineering Design (ICED 09)*, edited by L Leifer, P Skogstad, M Norell Bergendahl, and M Grimheden. The Design Society.

Vermaas, Pieter E., and Wybo Houkes. 2006. "Technical Functions: A Drawbridge Between

the Intentional and Structural Natures of Technical Artefacts." *Studies in History and Philosophy of Science Part A* 37 (1): 5–18. https://doi.org/10.1016/j.shpsa.2005.12.002.

Wang, Yilun, and Michal Kosinski. 2018. "Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation from Facial Images." *Journal of Personality and Social Psychology* 114 (2): 246–57. https://doi.org/10.1037/pspa0000098.

Winkler, Julia K., Christine Fink, Ferdinand Toberer, et al. 2019. "Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition." *JAMA Dermatology* 155 (10): 1135. https://doi.org/10.1001/jamadermatol.2019.1735.

Woodward, James. 1989. "Data and Phenomena." *Synthese* 79 (3): 393–472. https://doi.org/10.1007/bf00869282.