Noname manuscript No.
(will be inserted by the editor)

Reliability and Interpretability in Science and Deep Learning

Luigi Scorzato

Abstract In recent years, the question of the reliability of Machine Learning (ML) methods has acquired significant importance, and the analysis of the associated uncertainties has motivated a growing amount of research. However, most of these studies have applied standard error analysis to ML models—and in particular Deep Neural Network (DNN) models—which represent a rather significant departure from standard scientific modelling. It is therefore necessary to integrate the standard error analysis with a deeper epistemological analysis of the possible differences between DNN models and standard scientific modelling and the possible implications of these differences in the assessment of reliability. This article offers several contributions. First, it emphasises the ubiquitous role of model assumptions (both in ML and traditional Science) against the illusion of theory-free science. Secondly, model assumptions are analysed from the point of view of their (epistemic) complexity, which is shown to be language-independent. It is argued that the high epistemic complexity of DNN models hinders the estimate of their reliability and also their prospect of long term progress. Some potential ways forward are suggested. Thirdly, this article identifies the close relation between a model's epistemic complexity and its interpretability, as introduced in the context of responsible AI. This clarifies in which sense—and to what extent—the lack of understanding of a model (black-box problem) impacts its interpretability in a way that is independent of individual skills. It also clarifies how interpretability is a precondition for assessing the reliability of any model, which cannot be based on statistical analysis alone. This article focuses on the comparison between traditional scientific models and DNN models. However, Random Forest (RF) and Logistic Regression (LR) models are also briefly considered.

1 Introduction

Machine Learning (ML) methods in general (Hastie et al., 2009) and Deep Neural Networks (DNNs) in particular (Goodfellow et al., 2016; LeCun et al., 2015) have achieved tremendous successes in the past decade. For example, a classifier based on the *resnet* architecture (He et al., 2016) reached human level accuracy in the ILSVRC2015 challenge (Russakovsky et al., 2015). Furthermore, Neural Networks based on the idea of *transformers* (Vaswani et al., 2017) have recently spurred breakthroughs in the field of Natural Language Processing (NLP), enabling high-quality machine translation. Answers generated by Large Language Models (LLM) such as GPT-3 (Brown et al., 2020) have achieved an impressive level of similarity to those generated by humans. There is by now convincing evidence that the best ML/DNN algorithms can learn effectively highly sophisticated tasks. However, some important questions remain open concerning the reliability of DNN algorithms.

First, we cannot completely exclude the possibility that successful DNN algorithms are over-fitting the collections of datasets that are used to train and test them (Recht et al., 2018). In fact, because of the difficulty of collecting good quality labelled data, very few very popular datasets (in particular CIFAR-10, Krizhevsky et al. (2014), ImageNet, Russakovsky et al. (2015) and a few others) represent the unique benchmark for the majority of the research work on DNN. This violates one of the key statistical assumptions of all ML methods, namely that the parameters are set independently of the test data. Strictly speaking, if some data have been used in an article we read before designing our new DNN architecture, the same data should not be used to test our new architecture. In practice, it can be hard to adhere to this requirement.

Accenture AG, Genève, Switzerland. E-mail: luigi.scorzato@accenture.com

Secondly, successful applications of ML methods are much more likely to be published than unsuccessful ones. While *publication biases* (Dickersin et al., 1987) affect all types of scientific research, one can expect a much bigger impact on ML research, whose persuasive power relies crucially on its empirical success. This is related to the lack of full understanding of the mechanism that allows DNN to learn some features and ignore others.

Furthermore, assessment of confidence levels of ML predictions is notoriously difficult, especially in the case of DNN. The best evidence of these difficulties is provided by so-called *adversarial examples* (Szegedy et al., 2014): images that are misclassified with very high confidence by the DNN classifier, although they are humanly indistinguishable from other images that are correctly classified. Adversarial examples are much more pervasive in DNN than initially expected (Carlini and Wagner, 2017). The problem is not the occurrence of misclassification itself, but the high sensitivity to invisible features and the very high confidence (easily $> 99\%$) quoted by the classifier. This is clearly not a reliable error estimate[1].

Finally, the presence of social biases in some datasets widely used to train ML algorithms (Buolamwini and Gebru, 2018) is a matter of concern. Reliable error estimates could help significantly to detect earlier those predictions that rely on too limited statistics. This would be an important step toward the goal of deploying responsible AI (European Commission, 2023; High-Level Expert Group on AI, 2019) at scale.

ML methods in general, and DNN in particular, are being used successfully in many contexts where assessing their expected errors is not necessary. This applies, for example, to all those cases where the ML method improves the search efficiency of candidate solutions that can be subsequently verified by other means (Duede, 2023). This is the case in a vast range of applications spanning material discovery, drug discovery, predictive maintenance, fraud detection (Baesens et al., 2015), code suggestion (Brown et al., 2020), protein folding (Jumper et al., 2021) and many others. In those cases, assessing the reliability is the responsibility of the independent check, while the ML method merely improves the efficiency of the overall process. However, there are also applications where independent checks are not practical (e.g. for safety-critical real-time systems (Buttazzo, 2022)) or cases where it is crucial to estimate the amount of missed solutions (e.g. compliance applications). In those cases, assessing the reliability of the ML method is very important.

This paper addresses the reliability of DNN methods from a fundamental epistemological point of view. It is very important to combine this perspective with a purely statistical one because even traditional science does not offer an absolute (assumption-free) guarantee that its predictions have specific probabilities. So, we must understand to what extent DNN models rest on similar grounds as those employed by more traditional scientific disciplines and to what extent they might suffer from fundamentally different problems. It is also worth noting that the DNN approach displays many aspects of a new paradigm (Kuhn, 1962) with respect to traditional scientific disciplines. While this paper argues that the comparison with traditional science is possible and necessary, it certainly cannot be performed naively.

This paper focuses on models based on DNNs, because these have been responsible for the most impressive successes in recent years and because they pose the most interesting challenges from an epistemological point of view. However, we will occasionally consider also other ML methods, such as Logistic Regression (LR) and Random Forest (RF) models as they share some features of DNN models and some features of traditional scientific models. For the sake of concreteness, we focus on supervised ML models for binary classification. However, everything discussed in this article could be easily extended to other models of supervised ML. Unsupervised ML methods have quite different purposes and are beyond the scope of this paper.

The topic of this article falls at the intersection of the theory of complexity (Li and Vitányi, 2019), DNNs (Goodfellow et al., 2016; LeCun et al., 2015), responsible AI (MacIntyre et al., 2021; Eitel-Porter, 2021) and the epistemology of ML, which has attracted considerable attention recently (see e.g. Desai et al. (2022); Leonelli (2020)). As opposed to many works on this topic (but similar to Kasirzadeh et al. (2023); Zenil and Bringsjord (2020) among others), this article stresses the important peculiarities of DNNs within ML methods. The present approach has some similarities with the one adopted in (Watson and Floridi, 2021). Some of the most important differences are: (a) the present focus is on *reliability*, that naturally leads to consider *global*, rather than *local* interpretability; (b) for the same reason, the *relevance* defined in (Watson and Floridi, 2021) is less applicable and it is not considered here; (c) finally, Watson and Floridi (2021) consider a subjective notion of simplicity, while we focus on the hardcore complexity that no human can reduce, regardless of language and individual skills. This article investigates the objective foundations of reliability assessment (i.e. based on well-defined assumptions).

[1] It also does not help to say that adversarial examples have negligible probability among natural images, because such statement cannot be made quantitative, as we lack a suitable characterisation of the space of *possible natural images*. Moreover, adversarials have been also reproduced from natural images (Hendrycks et al., 2019).

Hence, it is largely unrelated to the literature that investigates the sociological basis for the trust in ML models, or analogies between ML and human behaviours (see e.g., Duede (2022); Clark and Khosrowi (2022); Tamir and Shech (2023)). Further comparisons with the literature are provided in the main text.

2 Assessing the reliability of model predictions in science and ML

To assess the reliability of any model we must be able to estimate the uncertainty of its predictions, in some precise and useful probabilistic sense. To start, it is convenient to distinguish between *statistical* and *systematic* uncertainties. Traditionally, statistical uncertainties are defined as those error sources which have a known statistical distribution (Bohm and Zech, 2017)[2]. These uncertainties can be safely analysed via statistical methods. Systematic uncertainties are all the others: they may stem from systematic distortions in the measurement devices, from imbalances in the data selection, from inaccurate models, from approximation errors, or from inaccurate parameter fitting. As emphasised in Bohm and Zech (2017), even random noise with unknown statistical properties must be classified as a systematic effect[3]. Statistical uncertainties can be very difficult to estimate in practice, but the process is conceptually clear. They are *knwon-unknown*, as they are expected by the model. Systematic uncertainties are *unknown-unknown* and they can be estimated only by enriching the model assumptions. This raises the question of **which assumptions are acceptable to assess our model assumptions?**

This question is crucial and problematic not only for ML models but also for traditional scientific (TS) models[4], and they have been the focus of extended analysis in the philosophy of science. It is therefore important to compare ML models to TS models in this respect, to understand to what extent ML introduces novel issues to the existing ones. It turns out that studying the reliability of ML models forces us to reconsider classic philosophical questions that are too often regarded as void of practical consequences.

Besides DNN and TS models, in the present comparison, we consider also Logistic Regression (LR) models and Random Forest (RF) models, because they represent very popular ML models whose properties often interpolate between DNN models and TS models in an interesting and instructive way.

2.1 Sources of errors

It is useful to distinguish four different sources of errors and see how they affect ML and TS models in different ways. Errors may stem (a) from data measurements, including the labels of training data; (b) from the model, which may not faithfully describe the actual phenomena in scope; (c) from any approximation that may be applied to the model to derive further predictions; and (d) from fitting the parameters that are left free by the model[5].

Source (a) is the same for any models under consideration (ML or TS). Note that all measurements are theory-laden (Hanson, 1965; Duhem, 1954). But, whenever two models are employed to study the same phenomena, we can also assume that the theory behind these measurements is the same. Source (c) applies only when the original model is replaced by an approximation to enable further analytic or numerical derivation. ML models are already in a form that can be used for numerical applications and therefore source (c) applies only to some TS models.

Sources (b) and (d) are the most interesting ones when comparing TS and ML models. Model errors (b) can be reduced by extending the basic (ideal) model with further elements designed to model the deviations from the basic model. However, this enrichment comes at the expense of more assumptions and more parameters to be fitted (generating more uncertainties of type (d)). We may say that unknown model *errors* can be traded for more model *assumptions* and more parameters.

[2] The ML literature adopts sometimes similar but slightly different classification between statistical (aka random, aleatoric or 'data') uncertainties on one side and systematic (aka epistemic or model) uncertainties on the other (Hüllermeier and Waegeman, 2021; Gawlikowski et al., 2021). Here we stick to the traditional one, because it clarifies the role of the model assumptions, which is crucial in our discussion.

[3] Note that probability distributions without finite mean and variance, for which the central limit theorem does not apply, cannot be characterised empirically in most cases. Given the ubiquity of, e.g., self-similar phenomena (Barenblatt, 1996) it is clear that one cannot certainly assume that every noise generates *statistical* uncertainties.

[4] Let us be more specific on what we mean by TS models. These include any state-of-the-art textbook theories such as, e.g., the standard model of particle physics. Because we focus here on DNN models for classification, it may be useful to keep in mind also examples of TS models of classifications, such as the model of stellar classification (Gray et al., 2009) or the classification of chemical elements (Emsley, 2011) or the medical classification of diseases (WHO, 2021). In general, TS models are based on some well defined features, whose relevance is based of a wealth of background science.

[5] Type (a) and (d) may have both *statistical* and *systematic* contributions, while type (b) and (c) are purely *systematic*.

Source (d) refers to errors in the determination of the optimal parameters, even when the model might provide a correct representation of the phenomena. This may happen because of limited data, noisy data (source (a)), an inaccurate minimisation algorithm or an inexact parameter fitting process. Error source (d) has a statistical component, which is propagated from noisy data (a) and can be analysed statistically (along the lines of e.g. Marshall and Blencoe (2005) and references therein). However, source (d) also has systematic components, due to the complexity of determining the optimal parameters and also due to the unknown statistical properties of the data[6]. Moreover, increasing the number of parameters requires, in general, an exponential increase in the amount of data required for their determination[7]. In conclusion, also estimating (d) requires making suitable assumptions.

TS and ML models do not show fundamental *qualitative* differences from the point of view of the error sources above (except for the less important source (c)). But they do display *quantitative* differences: ML models tend to be very flexible and characterised by significantly more parameters than TS models. Traditionally, fitting too many parameters was considered hopeless, because of the curse of dimensionality (Giraud, 2021). But, the impressive predictive success of DNN models, despite having vastly more parameters than training data points, shows that we cannot dismiss them based on naive parameter counting. The question of this paper is whether we can also ensure the reliability of models with so many parameters. We will see that this is a very different question that will be addressed in Sec. 3 and Sec. 4. But, before coming to this, we comment, in Sec. 2.2, on the radical idea that the success of DNNs might even represent a new kind of "theory-free" science.

2.2 The illusion of predictions without assumptions

A common fallacy is to hope that we can produce error estimates which do not depend on any assumption, or, equivalently, produce estimates that take into account all possible models. In the context of ML, this fallacy becomes even more tempting because ML models can be very flexible, and, for some DNN models, a universal approximation theorem applies (Hornik et al., 1989; Leshno et al., 1993). The idea of "theory-free" science does enjoy some support (Anderson, 2008) and it is not ruled out by others (Desai et al., 2022).

The simple reason why this idea is fallacious is the well-known (but often overlooked) *underdetermination of the theory* by the data (Duhem, 1954; Quine, 1975; Stanford, 2021). No matter how much data we have collected, there are always infinite different models that fit those data within any given approximation. Moreover, for every conceivable prediction, there exist infinite models that still fit all the previous data and also deliver the desired prediction[8]. Data *alone* cannot enforce *any* conclusion.

Even ML models that enjoy the universal approximation property offer no exception to the underdetermination thesis[9]. On the contrary, they provide more evidence for it, because we can include any desired prediction among the data that we want to describe and the universal approximation property will ensure the existence of parameters that reproduce also the desired prediction within the desired approximation. The predictions of the ML model hence depend not only on the data but also on the architecture of the ML model and, in the absence of evidence to the contrary, also on the algorithmic and initialisation details.

Another frequent fallacy is the idea that we should only accept assumptions that have been tested. But how much and under which conditions should they be tested? Newton's laws were tested extensively under all conceivable circumstances for over two centuries before discovering that they do not always hold. There is no way to fully test our assumptions (or to define their scope of applicability) in a future-proof way. They may fail in contexts that we cannot formulate because we cannot foresee them.

In conclusion, any error estimate is necessarily based on assumptions that cannot be fully justified on an empirical basis. They might be narrower or broader than the assumptions of the model itself, but they must be acknowledged because the error estimates crucially depend on them.

[6] for some classes of noise sources, collecting sufficient data can be completely hopeless in practice (see, e.g., de Forcrand (2009)) or even in principle (see e.g., Gabaix (2009)).

[7] An exponential increase is expected, in the generic case, according to the curse of dimensionality (Giraud, 2021).

[8] In the context of the classical analysis of knowledge, the same problem is usually referred as the Gettier problem (see e.g. Floridi (2004) and references therein).

[9] Even assuming that the training process is able to attain the best approximation, which is not ensured by the universal approximation theorem itself.

3 Current approaches to assess the reliability of DNN predictions

After the general considerations of the previous section, let us discuss how the reliability of DNN predictions is currently assessed in the literature.

Until recently, it was very common to quote the normalised exponential function (or softmax) (Bishop, 2006) as a measure of confidence in a DNN prediction. However, it has been shown very convincingly that the softmax output lacks the basic features of a measure of confidence (see e.g. Guo et al. (2017)). In particular, it tends to be increasingly overconfident for increasingly out-of-distribution samples (Hein et al., 2019). Ad-hoc attempts to correct this pathology lead to arbitrary results.

The issue of proper estimation has received growing attention in recent years, and numerous review articles are available today (Gawlikowski et al., 2021; Abdar et al., 2021; Loftus et al., 2022). A popular view regards the Bayesian approach as the most appropriate framework, in principle (Wang and Yeung, 2016). According to this view, the main drawback of a full Bayesian estimate is its prohibitive cost, which leads to a very active search for approximations that offer the best trade-off between accuracy and computational efficiency (Titterington, 2004; Jospin et al., 2020; MacKay, 1992; Gal and Ghahramani, 2016; Blundell et al., 2015; Sensoy et al., 2018; Hartmann and Richter, 2023). Besides the Bayesian framework, the other main approaches rely either on ensemble methods (Lakshminarayanan et al., 2017; Wen et al., 2020; Michelucci and Venturini, 2021; Tavazza et al., 2021), or data augmentation methods (Shorten and Khoshgoftaar, 2019; Wen et al., 2021). An alternative is to train the DNN to specifically identify outliers or uncertain predictions (see e.g. Malinin and Gales (2018); Raghu et al. (2019)). Let us consider them separately in the following.

3.1 Bayesian Error Estimates

Consider the Bayesian approach first. The computational cost is not the only difficulty. It is well known that Bayesian error estimates depend on the assumption of *prior distributions* (or priors) for the parameters whose impact should be estimated, and the choice of priors is largely arbitrary (Gelman et al., 2013). For infinite statistics, the Doob and Berstein - von Mises theorems (Van der Vaart, 2000) state that the posterior distributions converge to prior-independent limits. In simple words: “the data overwhelms the priors”.

However, such an argument is based on the unrealistic idea that the model is fixed, while we can rather easily collect more data. This is not what usually happens in real applications. In most cases, instead, collecting valuable data can be very hard, while producing new models is much easier. This is the case in traditional science, where planning a new experiment might take decades, while new models are published every day. It is the case also in most ML applications, where collecting *labelled* data can be extremely labour intensive, while new ML models are produced at the rate of millions per second by the ML algorithm. Under these conditions, there is no guarantee that the dependency on the priors will be sufficiently suppressed by the time the model is used.

Increasing the number of parameters might look like a way to eliminate the need of changing the model. But, the Doob and Berstein - von Mises theorems do not apply when the number of parameters is of the same order as the size of the dataset or more (Johnstone, 2010; Sims, 2010). Extensions exist for high-dimensional problems (Banerjee et al., 2021) and even infinite-dimensional (aka nonparametric) ones (Ghosal and Van der Vaart, 2017). However, those extensions rely on the assumption that most parameters are very close to zero, which is achieved via LASSO priors or equivalent suppression mechanisms (Banerjee et al., 2021). Note that this is a much stronger assumption than requiring that all the relevant parameters live in *some* low-dimensional subspace of a high-dimensional model. A parameter suppression mechanism imposes a sparse realisation of the *given* parametrisation. As a result, even in the asymptotic limit of infinite statistics, the posteriors depend on the choice of the parametrisation, similar to the case of low-dimensional models. In other words, nonparametric models are not a way to eliminate model dependency. The Bayesian approach is an excellent framework to discuss the probability of a model within a given class of models. But the universe of all possible classes of models is really big and the Bayesian approach is not a way to estimate the probability of those overall model assumptions. This holds for both ML and TS models.

3.2 Frequentist Error Estimates

The frequentist approach (Lakshminarayanan et al., 2017; Wen et al., 2020; Michelucci and Venturini, 2021) is not better equipped to answer the questions above (Sims, 2010). In fact, the frequentist error assessment relies on the same assumptions that characterise the model that has produced the predictions (and, possibly, additional assumptions). By construction, the frequentist approach is not designed to discuss the probability of the selected model. Frequentist error estimates do have a value, but only under the assumption that the model is valid for some parameters in the neighbourhood of the value selected by maximal likelihood.

This limitation is especially problematic for a model that displays high sensitivity to details, which are not understood. Because any little change to the parameters, as it happens, e.g., for any new step of training, may bring the system to a configuration for which the previous error analysis is not relevant anymore.

3.3 DNN-based Error Estimates

The spectacular and inexplicable success of DNNs represents a strong temptation to try to use DNN models to answer just about any difficult question, provided that we have at least some labelled examples to train the model. The fact that we do not understand why DNNs work, or when they do, makes almost any new application an interesting experiment in itself.

In this sense, using another DNN to detect outliers (Malinin and Gales, 2018) or uncertain predictions (Raghu et al., 2019) represents an interesting further application of DNN, but it does not provide an error estimate that relies less on DNN model assumptions. On the contrary, we must rely on *two* DNN models, in this case.

On the other hand, the same criticism may be directed equally well against traditional science: we rely on the laws of physics to build experimental devices that we use to test the laws of physics themselves. And we rely on other laws of physics to determine the accuracy of our measurements. Why is this problematic for DNN, if it is not for traditional science? Or should it be? Again, the analysis of DNN forces a reappraisal of classic epistemological questions concerning traditional science. We do it on Sec. 4, but, before that, we consider one last potential approach.

3.4 Predictions are NOT all you need

Can't we just use the success rate of the test dataset as an estimate of the expected prediction error? Intuitively, one expects that *novel predictions* on unseen data do provide a meaningful test. However, formulating this idea precisely is fraught with riddles (Crupi, 2021). In fact, this intuitive expectation is based on some hidden assumptions. In particular, we are implicitly assuming that the distribution of the data *will not change significantly*. But this is not a useful assumption, because we cannot deduce anything from the fact that "the distribution of the data does not change" without assuming anything about the distribution itself.

This is obviously a problem for DNN models, but also for TS models. In fact, as already remarked, we cannot define the scope of applicability of any theory in a way that is future-proof: the model may fail in circumstances that we currently cannot even imagine. But for DNNs the problem is somewhat harder because the domain of applicability is difficult to define also with reference to features that we do know.

In the philosophy of science, the idea that "predictions are all you need" goes under the name of *strong predictivism* (Barnes, 2022). However, the strong version of predictivism has essentially no support and it is used only as a target for criticisms. In fact, predictions *alone*—no matter how impressive—are never sufficient to warrant trust in a model (Barnes, 2022; Votsis et al., 2014). Trustworthy predictions require trustworthy premises. Imagine that a crazy newly proposed theory predicts, correctly, something as unlikely as the exact outcome of a national lottery (and otherwise agrees with standard theories). Would we feel compelled to embrace the new theory? Certainly not. Likelihood by itself is never sufficient to determine theory selection. Other non-empirical aspects are also important. But which ones?

This problem applies also to TS models, but it becomes really disturbing for DNN models whose persuasive power relies almost entirely on their empirical success. Although DNN successes are impressive, it is impossible to derive any quantitative level of confirmation from them without additional assumptions. For example, published results certainly suffer a positive selection. Kaggle competitions (Goldbloom, 2015) are very interesting also from a philosophical point of view because they provide a controlled framework for assessment. However, they also

cannot be used directly to confirm either general DNN architectures or specific DNN models, because submissions are strongly biased in terms of the approach used and the competencies of the participants. The fact that some successful DNN architectures keep being successful over time is certainly convincing evidence that they can learn something valuable. However, also in those cases, we cannot exclude that the DNNs have actually learned unwanted features that happened to have a strong spurious correlation with the labels (see, e.g., Xiao et al. (2021) for an example of this phenomenon). This confirms that also DNN predictions are only valuable when they stem from a valuable model. What characterises a valuable model, both in TS and in ML? This question is addressed in more depth in Sec. 4.

4 Assumptions, simplicity and interpretability

The previous discussions started from different perspectives which all led to the same conclusion: the reliability of a model necessarily depends on the reliability of all its model assumptions, which are never self-justifying. Empirical evidence alone is never enough, neither in ML nor in TS. In Sec. 4.1 we review which classes of assumptions characterise TS models, DNN models and other ML models. While we cannot assess the reliability of any of these assumptions, in Sec. 4.2 we propose a measure of their simplicity as the only surrogate that we can access. In Sec. 4.3 we identify a close relation between the simplicity of the assumptions and the concept of interpretability (now intensively discussed within the field of responsible AI).

Any model, whether TS or ML is essentially characterised by all its assumptions and its accuracy with respect to the existing empirical data[10]. In the following, the expression "all assumptions" refers to all those that are necessary to reproduce the outcome of the model, including the comparison with the empirical evidence and the expression of the input data[11]. The training data themselves are not part of the model assumptions[12]: they contribute to the empirical accuracy of the model just like the experiments that contribute to the development of TS models.

4.1 Assumptions

TS models rely on multiple categories of assumptions: besides the specific assumptions of the *model* itself, they rely also on multiple scientific disciplines that are not the main focus of the specific TS model but are essential to constrain the TS model. These disciplines are referred to as *background science* and they include a variety of domains from logic, mathematics, basic physics, chemistry and the modelling of any relevant experimental device. TS models typically include also assumptions to restrict their own *domain* of applicability[13]. Finally, TS models assume that their free parameters should be selected based on either a Maximum Likelihood principle (in a frequentist approach) or via a Bayesian argument that includes the assumption of Bayesian priors. Notably, TS models do not normally question the parameters inherited from background science, unless it becomes really compelling. In fact, those parameters must be consistent with a much broader spectrum of empirical evidence from very different domains. In this sense, traditional science adopts a kind of *divide et impera* strategy to parameter fitting.

On the other hand, DNN models require very little assumptions from background science (besides modelling the data collection). They are very generic and they are essentially defined only by the architecture parameters and hyper-parameters. However, their predictions are not determined by the training data, the DNN architecture and the Maximum Likelihood principle alone, not even approximately. They may depend also on subtle details of the training process. In fact, weight initialisation and pre-training techniques are key design choices when training and deploying a DNN model (Narkhede et al., 2022; Glorot and Bengio, 2010). Adversarials (Szegedy et al., 2014) are also evidence of high sensitivity to details. If we had evidence that the training of a DNN model depends only on some simple rules and is independent of the specific initial values of the weights (within some clearly defined bound and approximations), we would need to include only these rules as part of the assumptions of the DNN

[10] Defining a scientific or ML model by its assumptions might suggest that our discussion is restricted to a *syntactic* approach to scientific models. However, following (Lutz, 2017, 2014), it also applies to a *semantic* approach. Other views exist in the philosophical literature (Winther, 2021). But, there is currently no compelling evidence of real models that do not admit a formal description.

[11] If the model predicts an outcome with some statistical distribution, as it is mostly the case, the requirement is to reproduce the statistical distribution and not the individual outcomes.

[12] More precisely, they do not matter in any comparison. See footnote 31.

[13] Note that scope restrictions are logically equivalent to additional assumptions and are treated here as such.

model. Instead, the lack of understanding of the training dynamics enforces the inclusion of the full specification of the DNN (initial) parameters as part of the assumptions.

It is interesting to compare the DNN training process to other scientific applications of Markov Chain Monte Carlo (MCMC)[14]. For many TS models, there is no *proof* of convergence of the MCMC algorithm to a definite outcome. However, best-practice rules and diagnostic tools have been developed that enable a rather accurate formulation of the conditions that must be fulfilled for the algorithm to converge to a stable outcome (see e.g. Roy (2020)). Formal proofs of independence from details, if available, are very desirable, because no further assumption is needed in that case. But, also assuming a few semi-empirical rules is satisfactory, if they are de-facto accurate. Despite considerable effort, researchers have not been able to identify any simple set of rules that makes the outcome of DNN models independent of any further detail. Unfortunately, research is very difficult on this topic, because it requires testing different initial conditions, which is computationally very expensive. Moreover, this difficulty might actually be an intrinsic price to pay for the great flexibility of DNN models (see also Hartmann and Richter (2023). High sensitivity to details is the key aspect of the so-called *black-box* problem (Desai et al., 2022) of some ML methods: **lack of understanding is a serious limitation to the extent that we cannot tell which details actually matter for a conclusion**.

Furthermore, TS models strive to be consistent, when applied to different data and different domains. In particular, it is typically ensured that the values of the parameters remain consistent, within expected uncertainties, across different applications. Too large deviations are interpreted as failures of the model. DNN parameters, on the contrary, are usually not regarded as something that should be consistent across applications. Although DNN training often starts from DNN models that were pre-trained on other datasets, further training is always performed for new applications and no constraint is usually imposed to keep the parameters close to the original values. That means that different assumptions are made by the DNN models for different domains of application.

A further key difference between TS and DNN models is the specification of the domain of application. TS models typically define their domain of applicability in terms of the features that play a role in the model and they are typically measurable. If the domain is defined in terms of measurable features, it becomes possible to suspend a prediction for out-of-domain data. For example, domain restrictions usually enter the detailed specifications of the experimental set-ups that are required to collect valid data. Normally, the domain of applicability can be formulated in terms of a limited set of additional assumptions. This does not ensure that some overlooked features may not play an unexpected role and compromise the accuracy of the model, but this scenario happens rarely for state-of-the-art TS models.

In the case of DNN models, a measurable specification of the domain of application is much more difficult, and it is practically never provided. One key idea of DNNs is that the relevant features do not need to be specified explicitly. But this complicates the formulation of the domain of applicability. At least the experimental set-up for valid data collection must be specified with sufficient precision to ensure the relevance of the training dataset. For DNN, however, it is more difficult to define the domain by describing the experimental set-up, because, again, we have a very limited understanding of which features are learned during the DNN training process. Hence, the risk of omitting relevant prescriptions is much higher than for typical TS models. If, on the other hand, we omit the specification of the domain, we should test the performance of the DNN on any possible phenomena, even those totally different from the scope in which the DNN was actually trained.

It is worth including LR models in this comparison because they represent an interesting alternative to both TS and DNN models that live somewhere in the middle ground. LR models are defined unambiguously by their features[15]. Because the optimal regression coefficients are uniquely defined by the Maximum Likelihood principle, no coefficient must be included to define the assumptions of the model. Domain restrictions are also specified via the relevant features.

Another interesting comparison is with RF models: they also display significant sensitivity to some details (initial conditions and algorithmic hyperparameters), but they are based on limited and well-defined features. This enables, in the first place, a clear definition of the domain of applicability. It also makes it easier to identify conditions of robustness, for some applications and contexts. A deep analysis of the stability of RF models is beyond the scope of this article. However, it should be noted that their sensitivity to details seems closer to traditional applications of MCMC rather than DNN. Tab. 1 briefly summarises the content of this section.

[14] For other aspects of the comparison between DNN and computer simulations, see also (Boge, 2022).

[15] Normally, features are assumed to be directly measurable for LR models. If not, the assumptions of the LR model must include any specification needed to define them in terms of measurable ones. See also App. A.

Assumptions	DNN	RF	LR	TS
(a) background	Little/None	Features	Features	Many domains
(b) domain	Data collection set-up	Features	Features	Features
(c) model	Network Architecture	Features, hyper-param	Features	Many
(d) parameters	Init. weights & biases	Init. params	MaxLH	MaxLH / Priors

Table 1 A brief overview of the main categories of assumptions needed for DNN, RF, LR and TS models.

4.2 Simplicity of the assumptions

The previous section shows that the key difference between TS models and DNN models is that the former use **few assumptions for a wide range of phenomena**. A classic scientific-philosophical tradition singles out precisely this aspect as the main non-empirical epistemic value that scientific models should try to achieve, in addition to empirical accuracy (see e.g., Galilei (1962), p. 397, Newton (1964), p. 398, Lavoisier (1862), p. 623, Poincaré (1902), Mach (1882), Weyl (1932), Barnett (1950), Kemeny (1953), Quine (1963), Lewis (1973), Walsh (1979), Derkse (1992), Swinburne (1997), Nolan (1997), Scorzato (2013) also reviewed e.g. in Baker (2004), Fitzpatrick (2014), Zellner et al. (2001)).

An obvious problem with this view is that the number (and the length) of the assumptions is catastrophically language-dependent: it can always be made trivial—hence meaningless—by a suitable choice of language[16]. However, it was observed in Scorzato (2013) that requiring the inclusion of a basis[17] of measurable concepts among the assumptions enforces a minimal achievable complexity (see details in App. A and App. B). The original argument of Scorzato (2013) is applicable to a wide class of TS models, but not immediately to DNN models. It is shown in the App. A that the mere possibility of the existence of adversarials enables the extension of the argument to DNN as well. This enables the reference to the epistemic complexity of a model in a precise and language independent way (see App. B).

How does the epistemic complexity of a model impact its reliability? This question does not have a simple answer. Certainly, we cannot dismiss DNN models on the grounds of their high complexity, because we mostly do not have simpler TS models that cover the same topics and complexity-based model selection applies only to empirically equivalent models (see App. B for more details). However, a high complexity affects reliability at least in two indirect ways. The first one is *interpretability*. If it is challenging for any human to even understand which are all the assumptions of the model or memorise them, it becomes difficult to even formulate precisely the question: *what is it* whose reliability we are investigating? In this sense, a manageable epistemic complexity should be a precondition to assessing reliability. The next Sec. 4.3 looks in greater detail into the relation between the epistemic complexity introduced here and the concept of interpretability that plays a prominent role in the current discussion about responsible AI.

The second indirect relation between simplicity and reliability is via (scientific) *progress*. It was observed already in (Popper, 1959) that simpler models offer a more direct path toward scientific progress. The present framework offers further support for this idea: simpler models provide fewer ways in which they can be changed. Hence, they point more clearly to what must be changed, or they point to a fundamental deficiency of the model. Complex models might be more flexible and survive more empirical challenges, but that is not an advantage if we are looking for a better model. As a matter of fact, TS models are the result of a long history of small and big unambiguous improvements, where different models are unified under a single one (unification) or where the parameters of some empirical model are derived from a more fundamental one (reduction) or radically new models supersede the the existing ones via revolutionary changes. This history of progresses is certainly part of their perceived reliability. On the other hand, DNN models are too complex to allow the identification of this kind of unambiguous progress: every fine tuning of paramters or architecture adjustment might bring some improvements, but it is difficult to tell what might be lost. This topic is discussed further in Sec. 5.

[16] It is sometime noted that Kolgomorov complexity depends on the language only via a constant (Li and Vitányi, 2019). However, this is not sufficient for the purpose of model selection (Votsis, 2016).

[17] A *basis* is a set of quantities assumed to be measurable, that is sufficient to define all measurable quantities that are needed for the comparison with existing empirical evidence.

4.3 Interpretability

Recently, the notions of interpretability, explainability, explicability, transparency and related concepts have attracted considerable attention. Together, they build the core taxonomy of the topic of responsible AI (Arrieta et al., 2020). There is still considerable debate and obscurity around the precise definition of each of these terms, which have also started to acquire a legal connotation. The need for clarification has been analysed, from the philosophical point of view, by Beisbart and Räz (2022). In fact, the concept of *interpretation* has a long history in the philosophical literature (see, e.g., (Lutz, 2023) as a starting point). It is not the goal of this paper to try to clarify the relations between these concepts.

Instead, we focus here only on what is relevant for reliability. To this end, it is worth noting that one of the most quoted definitions of *interpretability* reads "the degree to which a human can understand the cause of a decision" (Miller, 2019; Lipton, 2018). Here a "decision" translates to what we previously called a "model prediction". What are the causes of model predictions? They are exactly the full set of model assumptions discussed in the previous section. In fact, they must include any information necessary to reproduce the model's predictions[18]. Anything less, would be an incomplete cause; anything more would be superfluous[19].

It is essential to appreciate that what matters is to understand the *causes* (i.e. the *model assumptions*) and not the entire *internal mechanism* that leads, step by step, from a given input to the corresponding model output. Most of our best scientific theories do not allow such "understanding", which is often encoded in millions of operations executed by computers. Not even the best world experts would be able to reproduce the model output without the support of those computers. But understanding these details is pointless[20], because all these operations are entirely determined by a few equations, which are the true *causes*, i.e. the *model assumptions*[21]. Of course, this is true also because we know that little details, such as the specific value of the random generator used in the simulation code, do not matter. Requiring a full understanding of the detailed mechanism is an act of desperation in cases where we do not know which details actually matter, but it is not a sensible requirement when we do know that.

Another interesting definition reads: "providing either an understanding of the model mechanisms and predictions, a visualisation of the model's discrimination rules, or hints on what could perturb the model" (Arrieta et al., 2020). As mentioned above, "understanding the model mechanisms" (or visualising them) is too strong a requirement, that would unfairly penalize most modern science. However, this definition also emphasises an important aspect: control of "what can perturb the model". As discussed in the previous section, if a set of simple rules ensures the stability of the output (not necessarily a unique exact output, but at least a unique statistical distribution which does not depend on other details), then those rules are sufficient causes/assumptions for the model's predictions. If not, more details must be included among the causes/assumptions.

Let us examine one more definition, which reads: "the degree to which a human can consistently predict the model's result" (Kim et al., 2016). This definition tries to remove the ambiguity of the word "understanding", which is used elsewhere, but without success: is the human allowed to utilise a tool to reproduce the model's results? If she is, then she can just use another, identical, DNN model and every DNN model becomes perfectly interpretable. If she is not, then most modern science is not interpretable, according to this definition, because no human being can go very far without any instrument. However, this attempt contains a very interesting aspect: the concept of understanding must be clarified; we must understand enough causes (i.e. model assumptions) to reproduce the outcome, and the outcome must be "consistent", i.e. robust to changing details. This is achieved if we know all the model assumptions, if these do not require the inclusion of too many details and if the output is computable for the relevant inputs (i.e. the input is letigimete for the model). So, also this last definition supports an identification of the concept of *interpretability* with a measure of the simplicity of all the assumptions.

To conclude this section, the *reliability* of a model crucially depends on its *interpretability* and the latter can be given a clear meaning that doesn't depend on the vague notion of "understanding", doesn't depend on any language and doesn't depend on individual skills.

[18] Another popular definition, that reads "the ability to explain or to present in understandable terms to a human" (Doshi-Velez and Kim, 2017; Hall, 2018), is more vague, but essentially consistent with the one of (Miller, 2019; Lipton, 2018).

[19] The *interventionist* framework introduced by Woodward (2005) is useful to identify the causes of a *specific* output, which is relevant for *local* interpretations as in (Watson and Floridi, 2021). However, when we consider the causes of *all* potential outputs produced by a model, as we must do to assess its reliability, then all non superflous model assumptions must be considered part of the causes.

[20] See also the discussion in Beisbart (2021) leading to similar conclusions concerning computer simulations.

[21] Knowledge of the model assumptions is a precondition for counterfactual explanations (Baron, 2023; Buijsman, 2022), but it is not sufficient to ensure it. Full visibility of how the output would change by any change of input may require unlimited computational power. This holds for DNN models as much as for most modern TS models.

5 Conclusions and way forward

DNNs are being used effectively in many contexts where it is not necessary to assess their reliability precisely. This is the case when DNN models generate candidate solutions that are eventually checked via independent tests (Jumper et al., 2021), or when they generate data that do not need to be sampled with a strictly defined distribution, because the conclusion has no strict statistical value or because the distribution is corrected afterwards (Albergo et al., 2019)[22]

However, when assessing the expected errors is necessary, it is important to understand how traditional science achieves that. The reliability of traditional science does not depend only on a statistical analysis of the uncertainties. It depends also on the fact that scientific models rely on *few assumptions* that remain the same for a very large amount of phenomena. This is accomplished by building minimal models for different domains of phenomena. Of course, these different domains are interdependent and they must be combined to describe more complex phenomena. However, this *divide et impera* strategy turns out to be quite efficient. Moreover, TS models strive to employ as much analytical understanding as possible (enabled by simpler models), next to empirical testing to ensure that empirical successes aren't ephemeral. In other words, **this strategy enables progress towards models that are gradually more and more accurate and/or more and more interpretable (i.e. simpler)**.

DNNs often display impressive predictive power, which can be possible only because, out of their vast amount of parameters, only a few combinations can be actually relevant. However, while the *existence* of a submanifold of relevant parameters (and the ability of the algorithm to *find* it, somehow[23]) is sufficient for their powerful predictivity, it is not sufficient for their reliability. The latter relies also on actually *identifying* that manifold of relevant parameters. This is necessary for interpretability, which is not just a desirable property for non-expert users: it is a precondition to assess the reliability itself. Identifying the assumptions is also necessary for scientific progress (Scorzato, 2016), because we need to know which assumptions we may want to replace.

Impressive predictive power in new domains, combined with weak foundations and difficult synthesis with background-science can be a signal of an ongoing (and incomplete) scientific revolution that may lead to a paradigm change (Kuhn, 1962), as was the case for the birth of Quantum Mechanics. What could be the way forward? Predicting the retirement of the scientific method itself (towards something like a theory-free science) appears misguided, according to the present analysis.

One natural hope is to better understand the conditions that ensure the robustness of DNN predictions, hence enabling an interpretable formulation of the assumptions. However, this might be inherently impossible, because the high flexibility of the DNN might be intrinsically coupled with their high sensitivity to details (Hartmann and Richter, 2023).

One interesting alternative approach is to use DNN as a tool to suggest features for other models, along the lines of (Huang and LeCun, 2006; Notley and Magdon-Ismail, 2018). However, it is important to ensure that the features obtained through this process are directly measurable and not simply the output of the last DNN layer, otherwise the DNN component cannot be removed from the formal counting of the assumptions needed to produce the results. The features extracted in this way would ease the connection with background science (imposing valuable constraints) and enable the building of simpler models, possibly unrelated to DNN models.

Another interesting approach is to focus on those models where the entire possible input space is potentially completely available for training. This is the case, for instance, when reading printed characters (from a limited choice of fonts), or scanning objects that belong to a finite list of possibilities (possibly rotated and translated in space). To cover the full input space, it is crucial to identify and implement all possible symmetry transformations, which include not only spatial transformations but also background and noise transformations. Even when all these steps are put in place, it is still difficult to check that the entire possible input space has been covered. Adversarials may still lurk, as long as the DNN is not exactly invariant under the transformation of redundant parameters. Ideally, to ensure robustness, one should also reduce the number of DNN parameters to the actual degrees of freedom of the input space. This is not easy, but it might be possible along the lines of (Liu et al., 2015).

Finally, another promising avenue of research is to study those limits that can be computed exactly, as the limit of infinitely large width of fully connected DNNs (Jacot et al., 2018). Such analytical results are essential to test the behaviour of a DNN where we know exactly how it should behave. So, their first epistemological advantage is

[22] If the independent check is performed by humans, it becomes important to compare the reliability of DNN with the one of humans. This a very important and interesting question, but beyond the scope of this paper, that focuses on comparing DNN models to TS models.

[23] The observation that DNN seem to be able to identify the right features, sometimes, is equivalent to state that it seems to solve the *reference class problem*, sometimes (Buchholz, 2023). However, our current problem is to characterise *sometimes*.

to increase the empirical accuracy of the DNNs (or rule them out). Furthermore, these limit cases could potentially also suggest how to realise simpler ML methods that are as powerful as DNN under specific circumstances. Another possibility is to use exactly solved limit cases as a starting point to set up DNNs that are small deviations from those limit cases, which might potentially enable a simpler formulation (with fewer assumptions) of the DNN itself.

A On the irreducible complexity of DNN representations

One may question the meaningfulness of assessing the complexity of the assumptions by counting the words used to express them. In fact, the length of the expression of a set of axioms is language-dependent. Even worse, for any set of axioms, one can always find a language where the same axioms assume a trivially concise form (Kelly, 2009; Votsis, 2016). This is the origin of an old paradox: why do scientists employ relatively complex formulations, even for their most fundamental theories, when a much simpler one is available? Is it due to hidden cultural bias? An answer was given in Scorzato (2013), where it was shown that the language that makes the axioms trivially short typically employs concepts that are not *measurable*. If we require the axioms to include sufficient relations to measurable concepts, finding a concise formulation becomes non-trivial and the shortest formulation among all available languages becomes meaningful. In fact, the standard scientific formulations are typically the most concise available, under these constraints[24].

The argument in Scorzato (2013) applies only to scientific theories that are sufficiently complex to allow the emergence of chaotic phenomena. In particular, any scientific theory that includes even the simplest form of Newton's equations (or similar differential equations) is already complex enough to fall within the scope of the argument in Scorzato (2013). Because most TS models typically rely on background science and the latter typically includes at least some minimal notions of basic physics, it follows that, for most real-life TS models, the length of the assumptions in their native scientific language is a meaningful measure of complexity.

However, DNN models rely on very little or no background science. Hence, it is an interesting question whether the formulation of a DNN model can be made arbitrarily concise or not. Interestingly, the mere possibility of the existence of adversarial examples enables an argument similar to the one in Scorzato (2013). This is not surprising, because adversarial examples have strong analogies with chaotic phenomena. As we will show below, a coordinate system does exist that makes the formulation of the DNN trivially concise (with just one parameter) and it is easy to find. But that formulation is not measurable (in the sense specified below). Moreover, finding a coordinate system that makes the DNN significantly more concise than the standard formulation, while provably preserving measurability, is a major challenge with no known solution. Therefore, the best estimate we have of the complexity of a DNN is the one that we can derive from its standard formulation.

Let us see how a simple formulation may run into conflict with measurability. It is not difficult to find a coordinate system $\xi \in [0,1]^n$ describing the objects to be classified (e.g. images of dogs and cats) such that the DNN classifies an image as cat (dog) when $\xi_0 = 0$ ($\xi_0 = 1$) and classifies it with decreasing (increasing) probability as cat (dog) when ξ_0 increases from 0 to 1. We can not only prove that such coordinates exist: they are essentially equivalent (up to trivial transformations) to the last layer representation of a DNN. In these coordinates, the DNN would have an extremely simple representation. In fact, it would have just one parameter $w = 1/2$ and the final output would be `cat` if $\xi_0 < w$ and `dog` otherwise.

Note that the ξ_0 coordinate does not tell whether the image represents a cat or a dog, it tells whether the DNN *predicts* it to be a cat or a dog (we will call these images *p-cats* and *p-dogs* for short). In fact, the more concise formulation that employs ξ does not change the predictivity of the DNN model, it only re-expresses it with a simpler notation.

Now, are the concepts of p-cat-ness and p-dog-ness *measurable*? In a model for visual perception of images, any measurement must be based on the observation of the colour of each picture element. The colour and position of each picture element can be determined only up to some limited precision, that is defined by human sensitivity. Hence, the claim that the attributes of p-cat-ness and p-dog-ness are measurable may be plausible only if small *imperceptible* changes in the picture cannot induce a change on their assessment from almost certainly p-cat ($\xi_0 \simeq 0$) to almost certainly p-dog ($\xi_0 \simeq 1$). However, this is exactly what the adversarials do: they are instances very close to p-cats which are humanly indistinguishable from dogs, and actually very close to p-dogs. In this sense, the concepts of p-cat-ness and p-dog-ness cannot be measurable. Here we have assumed that *measurable* quantities cannot assume different values, with high confidence, for imperceptibly different data points.

Note that we have *not* proved that any DNN model *cannot* be represented with coordinates that are both very concise and also measurable. We have only shown that an easy, very concise, formulation is not measurable. But this is sufficient for our goals, namely to refute the claim that there is always an easy way to formally trivialise the formulations of any system of axioms (Kelly, 2009; Votsis, 2016). Although that claim remains correct from a logical point of view, it does not hold anymore once the constraints of measurability are enforced. Discovering a simpler formulation might be possible and it would be a significant achievement[25].

B A simple model of progress

This appendix specifies the concepts of *(epistemic) complexity*, *theory selection* and *progress* introduced in the main text and it is essentially a summary of (Scorzato, 2016). The first step is to define a rule for *model selection*. We have seen that *empirical accuracy* is not enough to determine model selection: there are always infinite empirically equivalent theories. The key is then to identify at least one[26] non-empirical epistemic value that is sufficient to rule out empirically equivalent alternatives that the scientists would not consider valuable[27].

[24] Indeed, the scientific language evolves with science in a global optimisation process (Kvasz, 2008).

[25] Note also that, in general, it is not possible to tell whether a given formulation is the shortest possible: this question is not computable, as shown in Chaitin (1969). However, what matters to judge the complexity of a real scientific model is the shortest formulation *available*, which certainly exists and it is what Def. 1 refers to.

[26] One single additional value, besides empirical accuracy, offers an acceptable simple model of progress. See (Scorzato, 2013) for a discussion of other potential epistemic values that appear to be non-independent from simplicity and empirical accuracy.

[27] This criterion is *descriptive*. But, like any scientific law, it might be considered *normative* to the extent that it represents a good description.

The idea that *simpler assumptions* should be preferred to more complex ones—if they are empirically equivalent—is supported by a strong scientific/philosophical school of thought (see e.g., Galilei (1962), p. 397, Newton (1964), p. 398, Lavoisier (1862), p. 623, Poincaré (1902), Mach (1882), Barnett (1950), Weyl (1932), Quine (1963), Kemeny (1953), Lewis (1973), Walsh (1979), Derkse (1992), Swinburne (1997), Nolan (1997), also reviewed e.g. in Baker (2004), Fitzpatrick (2014), Zellner et al. (2001)).

A natural way to quantify the simplicity of the assumptions of a model is via Kolmogorov complexity (KC) (Li and Vitányi, 2019; Kolmogorov, 1965; Chaitin, 1969; Solomonoff, 1964; Hutter, 2005; Votsis, 2016). In general, the KC of a data string D is defined as the length of the shortest program p (in a given programming language) that produces D. In this context, the KC simply measures the length of the assumptions of the model $\mathcal{M}$ in the shortest formulation available in the given formal language (Chaitin, 1975; Scorzato, 2013). Comments about other frameworks can be found in the references cited in the App. C.

A key question is the dependence on the language in which $\mathcal{M}$ is formulated. In App. A it is shown that, if we require that $\mathcal{M}$ includes a definition of a basis[28] of measurable quantities, there is no practical way to find a language that makes the Kolmogorov complexity of $\mathcal{M}$ trivial. This extends to DNN models a similar argument for TS models (Scorzato, 2013) and justifies the definition of the complexity of $\mathcal{M}$ as the minimum length of the axioms across all available (measurable) formulations. See App. A for more details. Accordingly, we define:

Definition 1 *The* **(epistemic) complexity** *of a model $\mathcal{M}$ is the minimal length—across all available formulations (in any language) of $\mathcal{M}$—of all the assumptions of $\mathcal{M}$ that are needed to derive all the available results of $\mathcal{M}$. The assumptions must contain all references needed to ensure that any comparison between the empirical data D and the corresponding results of $\mathcal{M}$ are measurable*[29]*. The* **(epistemic) simplicity** *of a model $\mathcal{M}$ is the inverse of its complexity.*

Being defined as the minimum across all available formulations, Def. 1 is obviously language independent. The interesting observation is that it is also in non-trivial, in general, thanks to the analysis in App. A.

The same analysis justifies an estimate of the complexity by relying on the natural language in which the assumptions of a model are originally formulated (Kvasz, 2008). In the same spirit, we can assume a standard alphabet of symbols that can be used to express rather efficiently any existing model. A rough estimate of the model complexity can then be done by estimating the number of characters that are necessary to express the full assumptions. To be concrete, the specific model assumptions of a typical TS model might include about $O(10{-}1k)$, to which we must add, say, $O(10)$ scientific models from background science[30], which leads to a total of $O(100 - 10k)$ characters.

Model selection can then be defined as a Pareto efficient selection of models, where the Pareto agents are represented by simplicity and all the dimensions of empirical accuracy. Specifically,

Definition 2 **(Model selection)** *Given a set of empirical questions (i.e. a topic t), a model A is preferred to model B if A is neither more complex nor less empirically accurate than B on the topic t, while being strictly better than B in at least one of these aspects. In this case, we say also that model A is* **better** *than model B and B is* **worse** *than A*[31].

Note that this model selection does not require any trade-off[32]: it defines a Pareto frontier (state-of-the-art)[33]. Progress occurs whenever a Pareto improvement is achieved, thanks to a new model or new measurements:

Definition 3 *Given a topic t, the* **State-of-the-art** *is the ensemble of models which are not worse than any other model for the topic t. The models that are not state-of-the-art are* **obsolete***. There is* **progress** *when a model that was in the state-of-the-art becomes enduringly obsolete.*

These definitions are compared to a variety of real cases of theory selections and scientific progress in Scorzato (2016).

C Alternative notions of complexity

Kolmogorov complexity is a very popular measure of complexity, also widely used to estimate the complexity of a theory (Chaitin, 1975). Nevertheless, it is not the only possibility and one may question how the present analysis would change if we had used a different measure. For example, any specific compression software, like, e.g., gzip, implicitly defines an alternative notion of complexity. However, in all these cases, our conclusions about the complexity of DNN models would not change. Counting the atomic sentences (Carey, 2023) would also not change our conclusions.

In a ML context, it is common to refer to the *VC dimension* (Vapnik and Chervonenkis, 1971; Harman and Kulkarni, 2012), which is also interpreted as a measure of the complexity of the hypothesis (Vapnik, 1999; Corfield et al., 2009). To the extent that the VC dimension actually measures the complexity of the hypothesis, it is equivalent to our Def. 1. However, such interpretation has major limitations, if applied to our context. In this context we need a measure of *all* the assumptions of *any* model, in the language that is optimal for the model itself. The VC

[28] A *basis* is a set of quantities assumed to be measurable, that is sufficient to define all measurable quantities that are needed for the comparison with existing empirical evidence.

[29] For a discussion of the concept of "measurable" see App. A and (Scorzato, 2013).

[30] The advantage of using background science is that the same assumptions cover a much wider scope of phenomena.

[31] Note that, when comparing two models on the same data, the expression of the input data should be included as part of the assumptions, to prevent the possibility that one model artificially pushes the model complexity in the expression of the input data. This technical remark is only needed to ensure that the complexity of the input data does not actually matter.

[32] Well known rules exist that do define trade-offs between accuracy and model size (applicable for similar models that differ only in their dimension) (Akaike, 1973; Schwarz, 1978; Rodríguez, 2005). Different rules correspond to different trade-offs defined by different choices of priors (Burnham and Anderson, 2004). The epistemic complexity defined above extends the measure of complexity to any model, but it refrains from proposing a trade-off, which necessarily contain an additional arbitrary choice.

[33] Note that for two models entailing different empirical consequences, their comparison in terms of simplicity does not lead to a selection. However, it is often possible to alter ad-hoc the assumptions on both models to make them both empirically accurate and equivalent. The new (more complex) models are then comparable in terms of simplicity and the comparison may lead to model selection if one is better than the others beyond approximation errors.

dimension is designed for ML models only. In particular, the VC dimension of any parameter-free model is zero, which is very misleading for TS models that contain many ad-hoc hypothesis. Even if restricted to ML models, the VC dimension has problems of both overestimation and underestimation of the real complexity of the model assumptions. Concerning the former, it has been noted that the VC dimension is not suitable to describe the complexity of DNN models that contain a very large number of parameters (Dziugaite and Roy, 2017). This is the case whenever the algorithm is able to reduce the possible configuration that the model is able to reduce in practice. On the other hand, the VC dimension does not cover all those hypothesis that are not expressed in terms of explit parameters in the ML model. A number of extensions have been proposed (Bartlett, 1998; Bartlett and Mendelson, 2002; Wang et al., 2022) and, as long as they cover *all* the model assumptions, they are implicitly included among the possible expressions over which the minimum of Def. 1 is computed. A similar comment is applicable to the model dimension that is used by classic rules that define a trade-off between accuracy and complexity (Akaike, 1973; Schwarz, 1978).

One may also question the choice of using *any* notion of complexity of the assumptions to select among empirically equivalent models. In fact, philosophers of science have proposed a wide range of alternative epistemic values potentially relevant for theory selection. However, Scorzato (2013) argues that also these different values—to be meaningful—implicitly rely on the existence of a non-trivial notion of the complexity of the assumptions. Moreover, once the constraints of measurability are imposed, most classic epistemic values turn out to be not independent of epistemic complexity and empirical accuracy. The question is not settled, but we take the current evidence to justify the restriction of this paper.

Disclosure and Disclaimer The author works for a company that undertakes business in the deployment of AI systems as part of its commercial activities. The views expressed in this article are those of the author alone and do not necessarily represent the views of his employer.

References

Abdar, M., F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion 76*, 243–297.

Akaike, H. (1973). Information Theory as an Extension of the Maximum Likelihood Principle. In B. Petrov and F. Csaki (Eds.), *Second International Symposium on Information Theory*, pp. 267–281. Budapest: Akademiai Kiado.

Albergo, M. S., G. Kanwar, and P. E. Shanahan (2019, Aug). Flow-based generative models for Markov chain Monte Carlo in lattice field theory. *Phys. Rev. D 100*, 034515.

Anderson, C. (2008, June). The end of theory: The data deluge makes the scientific method obsolete. `https://www.wired.com/2008/06/pb-theory/`. [Online; posted 23-June-2008].

Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion 58*, 82–115.

Baesens, B., V. Van Vlasselaer, and W. Verbeke (2015). *Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection*. John Wiley & Sons.

Baker, A. (2004). Simplicity. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Winter 2004 Edition ed.). Stanford University.

Banerjee, S., I. Castillo, and S. Ghosal (2021). Bayesian inference in high-dimensional models.

Barenblatt, G. I. (1996). *Scaling, self-similarity, and intermediate asymptotics: dimensional analysis and intermediate asymptotics*. Number 14 in Cambridge Texts in Applied Mathematics. Cambridge University Press.

Barnes, E. C. (2022). Prediction versus Accommodation. In E. N. Zalta and U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2022 ed.). Metaphysics Research Lab, Stanford University.

Barnett, L. (1950). The Meaning of Einstein's New Theory – Interview of A. Einstein. *Life Magazine 28*, 22.

Baron, S. (2023). Explainable AI and Causal Understanding: Counterfactual Approaches Considered. *Minds and Machines 33*, 347–377.

Bartlett, P. (1998). The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory 44*(2), 525–536.

Bartlett, P. L. and S. Mendelson (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research 3*(Nov), 463–482.

Beisbart, C. (2021). Opacity thought through: on the intransparency of computer simulations. *Synthese 199*(3-4), 11643–11666.

Beisbart, C. and T. Räz (2022). Philosophy of science at sea: Clarifying the interpretability of machine learning. *Philosophy Compass 17*(6), e12830.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Volume 4 of *Information science and statistics*. Springer.

Blundell, C., J. Cornebise, K. Kavukcuoglu, and D. Wierstra (2015). Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR.

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

Bohm, G. and G. Zech (2017). Introduction to Statistics and Data Analysis for Physicists; 3rd revised. *Verlag Deutsches Elektronen-Synchrotron 488*, 978–3.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems 33*, 1877–1901.

Buchholz, O. (2023). The deep neural network approach to the reference class problem. *Synthese 201*(3), 111.

Buijsman, S. (2022). Defining explanation and explanatory depth in XAI. *Minds and Machines 32*(3), 563–584.

Buolamwini, J. and T. Gebru (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In S. A. Friedler and C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR.

Burnham, K. P. and D. R. Anderson (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research 33*(2), 261–304.

Buttazzo, G. (2022). Can We Trust AI-Powered Real-Time Embedded Systems? In M. Bertogna, F. Terraneo, and F. Reghenzani (Eds.), *Third Workshop on Next Generation Real-Time Embedded Systems (NG-RES 2022)*, Volume 98 of *Open Access Series in Informatics (OASIcs)*, Dagstuhl, Germany, pp. 1:1–1:14. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Carey, R. (2023). Bertrand Russell: Metaphysics. *The Internet Encyclopedia of Philosophy ISSN-2161-0002*, 1.

Carlini, N. and D. Wagner (2017). *Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods*, pp. 3–14. New York, NY, USA: Association for Computing Machinery.

Chaitin, G. (1969). On the Length of Programs for Computing Finite Binary Sequences: Statistical Considerations. *Journal of the ACM 16*, 145–159.

Chaitin, G. J. (1975). Randomness and mathematical proof. *Scientific American 232*(5), 47–53.

Clark, E. and D. Khosrowi (2022). Decentring the discoverer: how AI helps us rethink scientific discovery. *Synthese 200*(6), 463.

Corfield, D., B. Schölkopf, and V. Vapnik (2009). Falsificationism and statistical learning theory: Comparing the Popper and Vapnik-Chervonenkis dimensions. *Journal for General Philosophy of Science 40*, 51–58.

Crupi, V. (2021). Confirmation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021 ed.). Metaphysics Research Lab, Stanford University.

de Forcrand, P. (2009). Simulating QCD at finite density. *PoS LAT2009*, 010.

Derkse, W. (1992). *On Simplicity and Elegance: An Essay in Intellectual History*. Eburon.

Desai, J., D. Watson, V. Wang, M. Taddeo, and L. Floridi (2022). The Epistemological Foundations of Data Science: A Critical Review. *Synthese 200*(6), 1–27.

Dickersin, K., S. Chan, T. Chalmers, H. Sacks, and H. J. Smith (1987). Publication bias and clinical trials. *Control Clin Trials 8*(4), 343–353.

Doshi-Velez, F. and B. Kim (2017). Towards a rigorous science of interpretable machine learning. https://arxiv.org/abs/1702.08608.

Duede, E. (2022). Instruments, agents, and artificial intelligence. *Synthese 200*(6), 491.

Duede, E. (2023). Deep Learning Opacity in Scientific Discovery. *Philosophy of Science 90*(5), 1089–1099.

Duhem, P. M. M. (1954). *The Aim and Structure of Physical Theory*. Princeton: Princeton University Press.

Dziugaite, G. K. and D. M. Roy (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, pp. 1–14.

Eitel-Porter, R. (2021). Beyond the promise: implementing ethical AI. *AI and Ethics 1*(1), 73–80.

Emsley, J. (2011). *Nature's building blocks*. Oxford: Oxford University Press.

European Commission (2023). EU Artificial Intelligence Act. https://artificialintelligenceact.eu/the-act/.

Fitzpatrick, S. (2014). Simplicity in the Philosophy of Science. *The Internet Encyclopedia of Philosophy ISSN-2161-0002*, 1.

Floridi, L. (2004). On the logical unsolvability of the gettier problem. *Synthese 142*, 61–79.

Gabaix, X. (2009). Power laws in economics and finance. *Annu. Rev. Econ. 1*(1), 255–294.

Gal, Y. and Z. Ghahramani (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR.

Galilei, G. (1962). *Dialogue Concerning the Two Chief World Systems*. Berkeley: University of California Press.

Gawlikowski, J., C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, and X. X. Zhu (2021). A Survey of Uncertainty in Deep Neural Networks. https://arxiv.org/abs/2107.03342.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Ghosal, S. and A. Van der Vaart (2017). *Fundamentals of nonparametric Bayesian inference*, Volume 44. Cambridge University Press.

Giraud, C. (2021). *Introduction to high-dimensional statistics*. CRC Press.

Glorot, X. and Y. Bengio (2010, 13–15 May). Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh and M. Titterington (Eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Volume 9 of *Proceedings of Machine Learning Research*, Chia Laguna Resort, Sardinia, Italy, pp. 249–256. PMLR.

Goldbloom, A. (2015). What algorithms are most successful on Kaggle?

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Gray, R., C. Corbally, and A. Burgasser (2009). *Stellar Spectral Classification*. Princeton Series in Astrophysics. Princeton University Press.

Guo, C., G. Pleiss, Y. Sun, and K. Q. Weinberger (2017). On Calibration of Modern Neural Networks. https://arxiv.org/abs/1706.04599.

Hall, P. (2018). On the art and science of machine learning explanations. https://arxiv.org/abs/1810.02909.

Hanson, N. (1965). *Patterns of Discovery: An Inquiry Into the Conceptual Foundations of Science*. Cambridge University Press.

Harman, G. and S. Kulkarni (2012). *Reliable reasoning: Induction and statistical learning theory*. MIT press.

Hartmann, C. and L. Richter (2023). Transgressing the boundaries: towards a rigorous understanding of deep learning and its (non-) robustness. In *Artificial Intelligence–Limits and Prospects*, pp. 43–82. transcript Verlag.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference and prediction* (2 ed.). Springer.

He, K., X. Zhang, S. Ren, and J. Sun (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE.

Hein, M., M. Andriushchenko, and J. Bitterwolf (2019). Why RELU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41–50.

Hendrycks, D., K. Zhao, S. Basart, J. Steinhardt, and D. Song (2019). Natural Adversarial Examples. https://arxiv.org/abs/1907.07174.

High-Level Expert Group on AI (2019, April). Ethics guidelines for trustworthy AI. Report, European Commission, Brussels.

Hornik, K., M. Stinchcombe, and H. White (1989, July). Multilayer Feedforward Networks Are Universal Approximators. *Neural Netw. 2*(5), 359–366.

Huang, F. J. and Y. LeCun (2006). Large-scale learning with svm and convolutional for generic object categorization. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Volume 1, pp. 284–291. IEEE.

Hüllermeier, E. and W. Waegeman (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning 110*(3), 457–506.

Hutter, M. (2005). *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Berlin: Springer. 300 pages, http://www.hutter1.net/ai/uaibook.htm.

Jacot, A., F. Gabriel, and C. Hongler (2018). Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 31. Curran Associates, Inc.

Johnstone, I. M. (2010). High dimensional bernstein-von mises: simple examples. *Institute of Mathematical Statistics collections 6*, 87.

Jospin, L. V., W. L. Buntine, F. Boussaïd, H. Laga, and M. Bennamoun (2020). Hands-on Bayesian Neural Networks - a Tutorial for Deep Learning Users. *CoRR abs/2007.06823*, 1–20.

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature 596*(7873), 583–589.

Kasirzadeh, A., S. Rosenstock, and D. Danks (Eds.) (2023). *Philosophy of Science in Light of Artificial Intelligence*, Synthese Collections. Springer.

Kelly, K. T. (2009). Ockham's Razor, Truth, and Information. In J. van Behthem and P. Adriaans (Eds.), *Handbook of the Philosophy of Information*. Dordrecht: Elsevier.

Kemeny, J. G. (1953). The use of simplicity in induction. *The Philosophical Review 62*, 391.

Kim, B., R. Khanna, and O. O. Koyejo (2016). Examples are not enough, learn to criticize! Criticism for Interpretability. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 29, pp. 1–9. Curran Associates, Inc.

Kolmogorov, A. N. (1965). Three Approaches to the Quantitative Definition of Information. *Problems Inform. Transmission 1*, 1–7.

Krizhevsky, A., V. Nair, and G. Hinton (2014). The CIFAR-10 dataset. `http://www.cs.toronto.edu/~kriz/cifar.html`.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kvasz, L. (2008). *Patterns of change*. Science Networks. Historical Studies. Dordrecht: Springer.

Lakshminarayanan, B., A. Pritzel, and C. Blundell (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30, pp. 1–12. Curran Associates, Inc.

Lavoisier, A. (1862). Rèflexions sur le phlogistique. In *Oeuvres*, Volume 2, pp. 623–655. Paris: Imprimerie Impériale.

LeCun, Y., Y. Bengio, and G. Hinton (2015). Deep Learning. *Nature 521*(7553), 436–444.

Leonelli, S. (2020). Scientific Research and Big Data. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2020 ed.). Metaphysics Research Lab, Stanford University.

Leshno, M., V. Y. Lin, A. Pinkus, and S. Schocken (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks 6*(6), 861–867.

Lewis, D. K. (1973). *Counterfactuals*. Cambridge, MA, USA: Blackwell.

Li, M. and P. Vitányi (2019). *An Introduction to Kolmogorov Complexity and Its Applications* (4th ed.). New York: Springer.

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

Liu, B., M. Wang, H. Foroosh, M. Tappen, and M. Pensky (2015, June). Sparse Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Loftus, T. J., B. Shickel, M. M. Ruppert, J. A. Balch, T. Ozrazgat-Baslanti, P. J. Tighe, P. A. Efron, W. R. Hogan, P. Rashidi, G. R. Upchurch Jr, et al. (2022). Uncertainty-aware deep learning in healthcare: a scoping review. *PLOS digital health 1*(8), e0000085.

Lutz, S. (2014). What's Right with a Syntactic Approach to Theories and Models? *Erkenntnis 79*(8), 1475–1492.

Lutz, S. (2017). What Was the Syntax-Semantics Debate in the Philosophy of Science About? *Philosophy and Phenomenological Research 95*(2), 319–352.

Lutz, S. (2023). Interpretation. In *The SAGE Encyclopedia of Theory in Science, Technology, Engineering, and Mathematics*, Sage Reference, pp. 407–411. SAGE Publications, Inc.

Mach, E. (1882). Über die Ökonomische Natur der Physikalischen Forschung. *Almanach der Wiener Akademie*, 179.

MacIntyre, J., L. Medsker, and R. Moriarty (2021). Past the tipping point? *AI and Ethics 1*, 1–3.

MacKay, D. J. C. (1992, 05). A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation 4*(3), 448–472.

Malinin, A. and M. Gales (2018). Predictive Uncertainty Estimation via Prior Networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 31, pp. 7047–7058. Curran Associates, Inc.

Marshall, S. L. and J. G. Blencoe (2005). Generalized least-squares fit of multiequation models. *American Journal of Physics 73*(1), 69–82.

Michelucci, U. and F. Venturini (2021). Estimating Neural Network's Performance with Bootstrap: A Tutorial. *Machine Learning and Knowledge Extraction 3*(2), 357–373.

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

Narkhede, M. V., P. P. Bartakke, and M. S. Sutaone (2022). A review on weight initialization strategies for neural networks. *Artificial intelligence review 55*(1), 291–322.

Newton, I. (1964). *The Mathematical Principles of Natural Philosophy (Principia Mathematica)*. New York: Citadel Press.

Nolan, D. (1997). Quantitative Parsimony. *British Journal for the Philosophy of Science 48*(3), 329–343.

Notley, S. and M. Magdon-Ismail (2018). Examining the use of neural networks for feature extraction: A comparative analysis using deep learning, support vector machines, and k-nearest neighbor classifiers. `https://arxiv.org/abs/1805.02294`.

Poincaré, H. (1902). *La Science et l'Hypothèse*. Paris: Ernest Flammarion Ed.

Popper, K. (1959). *The Logic of Scientific Discovery*. New York: Basic Books.

Quine, W. (1975). On Empirically Equivalent Systems of the World. *Erkenntnis 9*, 313.

Quine, W. V. O. (1963). On simple theories of a complex world. *Synthese 15*(1), 103–106.

Raghu, M., K. Blumer, R. Sayres, Z. Obermeyer, B. Kleinberg, S. Mullainathan, and J. Kleinberg (2019). Direct uncertainty prediction for medical second opinions. In *International Conference on Machine Learning*, pp. 5281–5290. PMLR.

Recht, B., R. Roelofs, L. Schmidt, and V. Shankar (2018). Do CIFAR-10 Classifiers Generalize to CIFAR-10? *CoRR abs/1806.00451*, 1–25.

Rodríguez, C. C. (2005). The ABC of Model Selection: AIC, BIC and the New CIC. *Bayesian Inference and Maximum Entropy Methods in Science and Engineering 803*, 80–87.

Roy, V. (2020). Convergence Diagnostics for Markov Chain Monte Carlo. *Annual Review of Statistics and Its Application 7*(1), 387–412.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV) 115*(3), 211–252.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics 4*, 461–464.

Scorzato, L. (2013). On the role of simplicity in science. *Synthese 190*, 2867–2895.

Scorzato, L. (2016). A simple model of scientific progress. In L. Felline, F. Paoli, and E. Rossanese (Eds.), *New Developments in Logic and Philosophy of Science*, Volume 3 of *SILFS*. College Publications.

Sensoy, M., L. Kaplan, and M. Kandemir (2018). Evidential Deep Learning to Quantify Classification Uncertainty. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 31, pp. 3183–3193. Curran Associates, Inc.

Shorten, C. and T. M. Khoshgoftaar (2019). A survey on image data augmentation for deep learning. *Journal of big data 6*(1), 1–48.

Sims, C. (2010). Understanding non-bayesians. `http://sims.princeton.edu/yftp/UndrstndgNnBsns/GewekeBookChpter.pdf`. Department of Economics, Princeton University.

Solomonoff, R. J. (1964). A Formal Theory of Inductive Inference. Parts I and II. *Inform. Contr. 7*, 1–22, 224–254.

Stanford, K. (2021). Underdetermination of Scientific Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 ed.). Metaphysics Research Lab, Stanford University.

Swinburne, R. (1997). *Simplicity as Evidence of Truth*. Milwaukee: Marquette University Press.

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Tamir, M. and E. Shech (2023). Machine understanding and deep learning representation. *Synthese 201*(2), 51.

Tavazza, F., B. DeCost, and K. Choudhary (2021). Uncertainty Prediction for Machine Learning Models of Material Properties. *ACS Omega 6*(48), 32431–32440.

Titterington, D. (2004). Bayesian methods for neural networks and related models. *Statistical Science 6*(1), 128–139.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.

Vapnik, V. N. and A. Y. Chervonenkis (1971). On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability & Its Applications 16*(2), 264–280.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Votsis, I. (2016). Philosophy of Science and Information. In L. Floridi (Ed.), *The Routledge Handbook of Philosophy of Information*. Routledge.

Votsis, I., L. Fahrbach, and G. Schurz (2014). Introduction: Novel Predictions. *Studies in History and Philosophy of Science Part A 45*, 43–45.

Walsh, D. (1979). Occam's razor: A principle of intellectual elegance. *American Philosophical Quarterly 16*(3), 241–244.

Wang, H. and D.-Y. Yeung (2016). Towards Bayesian Deep Learning: A Framework and Some Existing Methods. *IEEE Transactions on Knowledge and Data Engineering 28*, 3395–3408.

Wang, R., L. Yang, B. Zhang, W. Zhu, D. Doermann, and G. Guo (2022). Confidence Dimension for Deep Learning based on Hoeffding Inequality and Relative Evaluation. `https://arxiv.org/abs/2203.09082`.

Watson, D. S. and L. Floridi (2021). The explanation game: a formal framework for interpretable machine learning. In *Ethics, Governance, and Policies in Artificial Intelligence*, pp. 185–219. Springer.

Wen, Q., L. Sun, F. Yang, X. Song, J. Gao, X. Wang, and H. Xu (2021). Time Series Data Augmentation for Deep Learning: A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization.

Wen, Y., D. Tran, and J. Ba (2020). BatchEnsemble: an Alternative Approach to Efficient Ensemble and Lifelong Learning. In *International Conference on Learning Representations*.

Weyl, H. (1932). *The Open World: Three Lectures on the Metaphysical Implications of Science*. Dwight Harrington Terry Lectures. Yale University Press.

WHO (2021). International classification of diseases. `https://icdcdn.who.int/icd11referenceguide`. Eleventh Revision (ICD-11).

Winther, R. G. (2021). The Structure of Scientific Theories. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021 ed.). Metaphysics Research Lab, Stanford University.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford university press.

Xiao, K., L. Engstrom, A. Ilyas, and A. Madry (2021). Noise or Signal: The Role of Image Backgrounds in Object Recognition. In *International Conference on Learning Representations*.

Zellner, A., H. A. Keuzenkamp, and M. McAleer (2001). *Simplicity, Inference and Modelling: Keeping It Sophisticatedly Simple*. Cambridge University Press.

Zenil, H. and S. Bringsjord (Eds.) (2020). *Philosophy and Epistemology of Deep Learning*, Volume 4 of *Philosophies Special Issues*, Basel, Switzerland. Multidisciplinary Digital Publishing Institute.