Penultimate draft – forthcoming in *Philosophy Compass*.

Reliability in Machine Learning

Authors: Thomas Grote (Tübingen)

Konstantin Genin (Tübingen)

Emily Sullivan (Utrecht)

Abstract: Issues of reliability are claiming center-stage in the epistemology of machine learning. This paper unifies different branches in the literature and points to promising research directions, whilst also providing an accessible introduction to key concepts in statistics and machine learning – as far as they are concerned with reliability.

1. Introduction

Machine learning models often achieve impressive accuracy under training conditions, but fail in spectacular or unexpected ways when they are deployed in real-world settings. Is there some way to guarantee that predictive accuracy in training carries over to the settings in which models are actually deployed? In other words: can we be justified in relying on machine learning models on the basis of their performance in training?

The underlying challenge is that there are various threats to reliability, arising at the time of (i) model development, (ii) model deployment and (iii) adapting the socio-technical environment to accommodate the model. Concerning (i), unlike traditional statistical models, there is no widely accepted mathematical theory that explains why and when state-of-the-art models such as deep neural networks generalize well. As for (ii), machine learning models are commonly used in unstable environments, or even induce changes to the environment itself, and they can be fooled by humanly imperceptible manipulations to the data. Regarding (iii), one basic issue is to align the

model output with the existing epistemic norms in a given domain, which often involves aggregating the model output with other kinds of evidence. In addition, there is another overarching problem: machine learning models are notoriously opaque, which is why it is difficult to understand the inner logic of how and why they arrive at a given prediction.

What is the relationship between the individual threats to reliability? When evaluating machine learning models, what types of assurances should we prioritize? What does the alleged (un)reliability tell us about the epistemic status of machine learning models in science and society? Can failure cases in machine learning even animate conceptual expansions or revisions of existing epistemological theories? These questions, among others, claim center-stage in the emerging debate on reliability in machine learning, spanning across statistical learning theory, mainstream epistemology, and the modeling literature in philosophy of science. This paper unifies different branches in the epistemology of machine learning and points to promising research directions, whilst also providing an accessible introduction to key concepts in statistics and machine learning – as far as they are concerned with reliability.

2. Statistical Learning Theory: The Received View

Statistical learning theory is the theory of reliability that is native to machine learning. The basics of the theory were developed by Soviet mathematicians in the 1960s; by the 1990s it was a fundamental part of the education of machine learning researchers. The spectacular successes chalked up by deep learning since the 2010s is somewhat incongruous with the theoretical predictions made by statistical learning theory. Indeed, reconciling the theoretical predictions of the canonical theory with the empirical successes of deep learning[i] remains one of the deepest

theoretical problems in machine learning (Belkin et al., 2019). Lack of progress in this area has caused the prestige of statistical learning theory to suffer somewhat in recent years. Nevertheless, mastering the basics of the received view is essential for understanding how machine learners approach issues of reliability.[ii] Statistical learning theory is a theory of an "ideal case" in which threats to reliability are well-understood and can be managed with precision. Indeed, many of the approaches to reliability that we cover in subsequent sections can be seen as responses to situations that depart from the assumptions of this classical theory.

Reduced to their barest essentials, supervised learning problems in machine learning go something like this: we intend to collect *N* data points $D = \{ (X_1,Y_1), (X_2, Y_2) \ldots (X_N,Y_N) \}$; we assume that these data will be randomly sampled from a probability distribution P. The X_i take values in an input space X and the Y_i take values in a label space Y. For example, the X_i may contain email text and metadata and the Y_i label them "spam" or "not spam". We would like to find a function f: $X \rightarrow Y$ that labels unseen emails "as accurately as possible", where the *accuracy* of f is defined as the probability of sampling a point (X,Y) from P, such that f(X)=Y. Typically, we do not consider every possible function from X to Y but only those in some particular function space F.

One of the fundamental problems of supervised learning arises because we need to use the data D in two ways: (1) to pick a promising function f from the function space F and (2) to estimate the accuracy of our chosen candidate. Suppose for a moment that step (1) is finished: someone else has proposed a function f and we intend to sample D to probe its accuracy. That reduces our problem to a simple statistical estimation problem: the success rate of f on samples in D is a statistically unbiased estimate of its accuracy on future points sampled from P. We can appeal to standard statistical results[iii] to derive the following kinds of guarantees:

(1) the probability that the accuracy of f differs appreciably from its success rate on D is less than ε,

where ε depends on the sample size N and precisely how large a difference has to be for us to consider it appreciable. Note that this guarantee holds only *before* we sample the points in D from the distribution P. Anything might happen after the sampling — if D is unrepresentative, it might mislead us about the accuracy of f. Nevertheless, if N is large, this should happen only with small probability.

Unfortunately, this is not the typical situation. We have to use D both to *pick* a promising f *and* to estimate its accuracy. If we "double dip" by using the same data to pick an f and to estimate its accuracy, then it is not at all straightforward to derive guarantees like (1). Unless a great deal of technical sophistication is applied, such a procedure will tend to yield flattering, but biased, estimates of the accuracy of the chosen function.[iv] The way that machine learners often deal with this situation is by splitting the data D into a training set D_{train} which is used to pick a promising f, and a test set D_{test}, which is used to estimate the accuracy of the chosen candidate. It is crucial that test data is *never* used to pick the candidate. If the machine learner is very disciplined, and never peeks at D_{test} in order to pick a candidate, then statistical guarantees like (1) can be derived for whatever f is ultimately chosen. However, this means that you can use the test set *only once*. Therefore, you had better be satisfied with your choice of candidate f before you use your precious test sample to estimate its accuracy. Moreover, it creates a fundamental trade-off: if you use more of the points in D for training, you may be better able identify the best candidate in F, but only at the expense of a less accurate estimate of its accuracy; if you use more of the points in D for testing, you will get a better estimate of the accuracy of your chosen candidate, but at the expense of data you might have used for picking a better one.

Hopefully, the preceding discussion makes clear that picking a promising f on the basis of the training set is, from a statistical perspective, a rather consequential decision. But shouldn't we simply pick the f in F that performs best on the training set? That might be computationally difficult, but isn't it rather obviously the right thing to do? The problem of *overfitting* bedevils this otherwise sensible proposal. To explain this, we need to introduce a notion of the *capacity* (or sometimes *richness* or *complexity*) of the function class F.[v] Intuitively, capacity notions capture how much a small perturbation of the data perturbs the best-fitting candidate in F. In the setting of regression, a typical example of a low-capacity class is the set of linear functions: perturbing the data does not significantly change the line of best fit; if, on the other hand, we consider the set of all polynomials of degree five, then small perturbations of the data — even moving just one point — can lead to large changes in the best-fitting polynomial. The latter situation should make you worried about *overfitting*: a good fit in the training sample might be highly misleading about accuracy on future samples. If you have overfit, then what you see (in the training data) is not what you get (out of sample). The fundamental results of statistical learning theory addresses this issue. These results provide the following kinds of what-you-see-is-what-you-get guarantees:

(2) the probability that the accuracy of the best-fitting f in F differs appreciably from its success rate on D is less than ε,

where, in addition to being a function of the sample size N, and how large a difference has to be in order to be appreciable, ε is now *also* a function of the capacity of F. As before, this guarantee holds only *before* we sample the points in D from the distribution P.

Holding fixed our sample size N and minimum appreciable difference δ, guarantees like (2) tell us *how rich a function space* we can afford to search on the "budget" given by our sample size, while maintaining the quality of the estimate of the accuracy of the best-fitting function.[vi] Holding fixed the function space F and the minimum appreciable difference δ, these guarantees tell us *how large a training sample* we need for the success rate on the training data to be a good estimate of the accuracy of the best-fitting function. Holding fixed the function space F and the sample size N, they tell us *how precise an estimate* of the accuracy we can expect. If you manage these trade-offs in the way suggested by theory, then what you see (in the training data) is, with high probability, what you will get (out-of-sample). Thus, when its preconditions apply, the guarantees of the theory allow you to solve the fundamental problem sketched above: you can use the training data both to pick a promising function from the space F and to reliably estimate its accuracy. Then, if you have test data left over, you can independently corroborate your estimate of the accuracy of the chosen function.

In practice, there are many situations where the guarantees provided by statistical learning theory do not apply. It is increasingly the habit of machine learners to search function classes of unbounded capacity, which means that the theory can provide no interesting guarantees. This is particularly true of work in deep learning. Despite the successes of deep learning, nothing magical is going on: if you train a deep neural network on white noise, it will give you a beautiful fit on the training sample and then fail spectacularly on future samples (Belkin, 2021). Thus, the predictions of statistical learning theory are borne out: if you do not bound capacity, then past results (in the training data) are no guarantee of future performance (out of sample). When a great deal of data is available, reliable estimates of accuracy can still be obtained from the test data. But, this requires the discipline never to peek at the test data during training.

In this section we have identified the reliability of a machine learning model with its predictive accuracy on samples drawn from the *same* distribution as the ones on which it was trained. Even if we search function classes of bounded capacity, significant problems are posed by the fact that in many deployment scenarios, the machine learning model is used on samples drawn from a distribution *different* from the one on which it was trained. We discuss this issue in the next section.

3. Robustness

Statistical learning theory provides no guarantees if the machine learning model is deployed in a distribution different from the one on which it was trained. Nevertheless, we might hope that the model is relatively *robust* under a range of situations. Rather than providing theoretical guarantees, research in robustness proceeds experimentally, employing a variety of evaluation techniques and mitigation strategies to detect and preempt possible performance failures. Freiesleben and Grote (2023) develop a conceptual framework for robustness in machine learning, synthesizing different strands in a fragmented research landscape. They define robustness as a multi-place concept, consisting of a *robustness target* (the machine learning model) and a *robustness modifier* (e.g., the deployment distribution). The target is robust if its performance is (relatively) stable despite (reasonable) changes to the modifier. The robustness notion is parametric and contextual: changes to the modifier are limited to a certain domain, fixed by what can be reasonably expected given the context; moreover, the acceptable loss in predictive accuracy between training and deployment conditions is fixed by the costliness of errors in the domain of interest.

On this basis, Freiesleben and Grote develop a taxonomy of different phenomena impeding robustness. They distinguish between (i) natural distribution shifts; (ii) performativity; (iii) shortcut learning; and (iv) adversarial attacks. Roughly, natural distribution shifts are changes in the distribution due to changing background conditions. Think of how the COVID-19 virus mutated

so that many of its initial symptoms lost their predictive value (Finlayson et al., 2021). Performativity, by contrast, describes a kind of distribution shift induced by the deployment of the machine learning model (Perdomo et al., 2020). To illustrate, consider how epidemiological models are used to inform policy decisions (i.e., when to release lockdowns) or how financial models can shape markets (see also Khosrowi and van Basshuysen, forthcoming).

Shortcut learning and adversarial robustness, in turn, refer to idiosyncrasies in how machine learning models *learn*. Shortcut learning describes situations in which a machine learning model achieves high predictive accuracy by associating features with the prediction target that do not hold across different settings e.g., when the model learns to use annotations on the margins of medical images for disease classification (Geirhos et al., 2020; Bellamy et al., 2022). Finally, adversarial attacks refer to the deliberate exploitation of learning idiosyncrasies by human actors: an adversarial actor searches for humanly imperceptible manipulations to the data that result in grave errors by the machine learning model. The predominant interpretation is that vulnerability to adversarial attacks arises because deep learning models detect predictively useful features in the data that humans cannot perceive (Ilyas et al., 2019; Buckner, 2020; Freiesleben, 2022).

The different stumbling blocks for robustness highlight why it is so challenging to establish guarantees for the reliability of machine learning models: distributions may change in many different ways and for many different reasons. Machine learning researchers have developed an advanced armory of game theoretic and causal inference techniques to predict distribution shifts or to identify stable points, but these methods are proven to work only under stylized conditions (Perdomo et al., 2020; Garg et al., 2022). Since we often have incomplete causal knowledge of a given domain, it can be hard to discern spurious from meaningful statistical relationships. Therefore, apart from paradigm cases of shortcut learning, it can be hard to tell whether relying on a statistical relationship threatens or supports robustness. The upshot is that ensuring the

robustness of machine learning models typically involves a trial-and-error process, requiring a combination of domain knowledge, external evaluation with out-of-distribution data, data augmentation, explainable AI (xAI) techniques, and continuous retraining.[vii]

Philosophers are increasingly connecting robustness in machine learning with modal conditions for knowledge familiar from traditional epistemology. For example, Vandenburgh (2023) claims that, just as justified true belief is insufficient for knowledge, a guarantee of predictive accuracy is also insufficient for machine learning models to produce knowledge. Drawing parallels with safety conditions in epistemology, Vandenburgh argues that, if they are to yield knowledge, machine learning models must be robust to errors by registering the right features and predicting accurately in counterfactual scenarios. Buijsman (2023) uses examples from machine learning to refine theories of process reliabilism (see also Goldman, 1979). He argues that in order for a belief to count as a result of a reliable belief-forming process, it is sufficient that the relevant output is reliable only in a local range of circumstances. The local range, which is determined by a similarity metric, must be large enough to ensure a non-accidental connection between the actual input and the output. The intersection of machine learning robustness and the modal conditions of epistemology is a promising new research avenue. Looming in the background, however, are questions about the extent to which the modal conditions for knowledge are (in)compatible with best practices in statistical science (Mayo-Wilson, 2018).

4. Socio-Technical Accounts of Reliability

The preceding discussion has focused on predictive accuracy, whether in the training distribution or the distribution arising in deployment. But a focus on predictive accuracy is not sufficient to capture everything that is at stake in machine learning reliability. While the model makes predictions, there are typically humans *in the loop* that base their decisions on the output of the

model. The distinction between predictions and decisions is an important one: the human decision-maker is typically not just a passive receiver of information, but an expert in her own right. For example, she might be a scientist, interested in finding out whether the model's predictions are vindicated by the evidence, a clinician trying to select the optimal treatment for a patient, or a judge making a sentencing decision. In all likelihood, the model's predictions will not be the only piece of evidence available to the human decision-maker, who forms a preliminary judgment independent of the model. Therefore, issues of model robustness are joined by questions of how to assess the evidential strength of machine learning predictions or how optimally to aggregate human and machine judgment. It is no coincidence that many of these issues are structurally similar to traditional problems of testimony and peer disagreement from social epistemology, with the important difference being that the testimony is generated by machines.

'Computational reliabilism' has emerged as a frontrunner in socio-technical accounts of reliability. Initially born of the debate about the trustworthiness of computer simulations (Durán and Formanek, 2018; see also Boge, 2021), it has recently been applied to machine learning in healthcare (Durán and Jongsma, 2021). Computational reliabilism seeks to answer the question of when a user is justified in trusting model outputs. The notion of justification is here loosely inspired by theories of reliabilist epistemology (Goldman and Beddor, 2021): we are justified in trusting a model if it has a history of producing correct outputs. Another distinctive feature of computational reliabilism is that it tries to provide a justification for trusting the output of models, while bypassing the so-called "black box problem".

In broad strokes, the strategy of computational reliability is to identify a number of epistemic safeguards, ensuring that the machine learning output is generated by a reliable process. These safeguards include (i) verification and validation, (ii) robustness analysis, (iii) a history of (un)successful implementations, and (iv) the role of expert knowledge. The first two conditions

are model-centric safeguards, while the latter are concerned with the model's socio-technical embedding. Roughly, (i) ensures reliability via formal methods and performance on benchmarks (see 'section 2'); (ii) involves testing the model under heterogeneous settings (see 'section 3'); (iii) requires that the model is used in accordance with the epistemic standards of a (scientific) domain (see also Winsberg, 2003); and (iv) is about whether the model merits the trust of relevant experts (see also Beisbart, 2017). Importantly, these conditions are not derived from an overarching epistemological theory but can be understood as an assemblage of epistemic best practices. Furthermore, how exactly these conditions are operationalized is domain-specific. As a case in point, the epistemic norms for how machine learning models are to be used vary significantly from particle physics to clinical settings.

In combining model-centric and socio-technical aspects, computational reliabilism is an ambitious research program. Like any ambitious research program, it faces some challenges: as discussed earlier, deep learning models lack proper theoretical foundations, making it difficult to obtain formal guarantees, as required by (i). More critically, (iii) and (iv) presuppose that the machine learning model is embedded in an epistemically well-ordered environment. The crux of the matter, however, is that in the process of being implemented in domains like criminal justice, clinical medicine, or science, machine learning models disrupt domain-relevant epistemic norms. For example, it is contested how much judges should rely on algorithmic risk estimates (Holm, 2023; Schmidt et al., 2023) or whether (and what kind of) scientific understanding can be provided by machine learning models (Sullivan, 2022a; Boge, 2022). These difficulties raise the possibility that computational reliabilism presupposes what it means to provide: a better grip on the epistemic standing of machine learning models.

An adjacent problem is discussed by Genin and Grote (2021): when machine learning models provide clinical decision-support it is pivotal to ensure that this enables clinicians to make more

reliable decisions. In order to determine this, we cannot just rely on standard evaluation techniques in machine learning, but ideally we would conduct randomized clinical trials. However, when compared to drugs, randomized controlled trials for machine learning models raise various methodological issues, possibly undermining the validity of the generated data.

5. Opacity and Reliability

The remaining central approach for assessing the reliability of machine learning models is by evaluating the internal decision process of the model. As discussed above, it might be that machine learning models achieve high predictive accuracy but do so by learning what we consider the *wrong* features, like in the case of shortcut learning. In one well discussed case, a machine learning model trained for classifying wolves versus huskies based its classifications on spurious features in the background of the image. More precisely, the model used the fact whether there was snow in the background to be predictive for wolves (Ribeiro et al., 2016). If we can inspect *why* a machine learning model makes the predictions that it does, then this is one central, if not necessary, way for assessing model reliability. In the case of the wolf v. husky classifier, the model is unreliable, because it relies on the wrong features, which is why it is prone to overfit. While this seems like a simple solution to the reliability problem, since machine learning models are opaque, there are several challenges for this method of reliability assessment.[viii]

In a recent paper, Duede (2022a) argues that opacity in machine learning prevents us from exploiting more traditional avenues of assessing model reliability. For example, he argues that machine learning models share a large resemblance to scientific instruments, such as simulations, that follow a "theoretically informed procedure to arrive at an output" (2022, p. 7). But in order to rely on such theoretically mediated instruments, we must assess the procedural processes that underlie them. However, due to opacity, he argues, we cannot gain access to the high-level logical rules of a machine learning model, and thus cannot assess the procedural processes internal to the

ML model. As a result, we cannot assess its reliability.[ix] Similarly, he argues, machine learning models cannot be assessed as reliable in the same way that we assess experts, because again, machine learning models lack transparency.[x] This leaves Duede to conclude that we need to expand our concept of reliability that incorporates principles distinct from those in brute induction, instrumentation, and expertise. Nevertheless, how this conceptual expansion of reliability might look, besides simply developing the right interpretability tools, is not addressed.

Perhaps one promising approach for having internal reliability is through a mechanistic interpretability approach (Langer et al., 2021). This approach borrows tools from neuroscience to do systematic experiments on deep neural networks to uncover mechanisms that contribute to the model's decision. The idea is that if we have mechanistic knowledge of how models make decisions then we can use that to judge whether the informed procedures constitute a reliable versus unreliable process. However, any research program in interpretability is still young, so if we need to know the internal properties of a model, then we may not be able to assess reliability in the short term.

Looking slightly beyond reliability, others have argued that interpretability is required in order to assess the epistemic success of a machine learning model (see also Räz and Beisbart, 2022). In one influential paper, Krishnan (2020), discusses model reliability in terms of justifying the use of a model without the need for internal transparency, but there are two cases where she argues model transparency is important: scientific discovery and public trust. First, she argues that the public often believes that knowing the inner logic of a machine learning model is paramount to assessing whether it is trustworthy and reliable. Turning to scientific discovery, Krishnan argues that interpretability methods are necessary to help uncover hypotheses, and knowing the internal logic of the model is necessary. However, assessing the reliability of a machine learning model used for scientific discovery seems like a paradigmatic case of external reliability. It does not matter if the

hypothesis we generate is actually linked at all to the model, all that matters is if the hypothesis turns out to be scientifically fruitful, which must be done by additional scientific work outside of the machine learning model (Duede, 2022b, Sullivan, 2022a, Zednik and Boelsen, 2022).

The problem of opacity in machine learning also introduces a second-order epistemic problem. xAI techniques *themselves* can be more or less reliable. Recent work in computer science has exposed that the leading xAI techniques are subject to adversarial attacks (Slack et al., 2020, Slack et al., 2021), can be just plain wrong (Rudin, 2019), and disagree (Krishna et al., 2022). If it is necessary to judge the reliability of a machine learning model through yet another model that *also* needs to be assessed for reliability, then we start to enter into epistemically pernicious territory.

While there is growing work around assessing the reliability of xAI, exploring the dependence and tension between xAI reliability and machine learning reliability is underdeveloped. Fleisher (2022) provides a nice argument that xAI models are idealizations of machine learning models, but he stops short of taking a critical stance on xAI models, or how we might evaluate if the idealizations are successful. Watson (2022), in turn, explores how Mayo's (2018) severe testing approach can be used to guide testing procedures that establish the fidelity of xAI models. Lastly, Sullivan (2022b) argues that the stakes of the domain, or even the stakes of the data subject, can influence how much opacity, or xAI reliability is a problem for model use or gaining knowledge or understanding of the model. However, this simply shows that there are context dependent ways of assessing how much this second-order problem matters.

But perhaps internal reliability is just a red herring. Other work calls into question the need for solving the internal reliability problem. For example, the work on computational reliabilism discussed above, downplays internal reliability: if we validate the model with data that is representative of the deployment distribution, and install the right external epistemic guardrails,

achieving transparency may not be necessary for justified trust in the model output. At best, XAI techniques might act as an additional plausibility check. One area for future work is connecting the collection of issues in reliability, model evaluation, and scientific discovery. Defining reliability in terms of modeling purposes along the lines of an adequacy-for-purpose model evaluation framework may help (Parker, 2020). In that case, there could be times where the internal reliability matters and times where it may not.

6. Taking Stock and Looking Ahead

This paper unified different strands of the epistemology of machine learning, concerned with reliability. Issues discussed ranged from the (lacking) theoretical foundations of deep learning models, robustness issues under deployment conditions, and problems of testimony and disagreement. These reliability issues are further complicated by the opacity of machine learning models' internal decision-processes. Regardless of which level we are referring to, a recurring theme is that, so far, we neither have the right technical tools nor a sufficiently pronounced conceptual understanding of the problems involved, to provide *clean* guarantees for the reliability of machine learning models.

Moreover, the rapid developments in machine learning are constantly giving rise to new developments, new conceptual problems will emerge. A pertinent example here are 'foundation models', referring to exceedingly large machine learning models (when compared to traditional deep learning models), whereby a 'foundation core' gets initially trained by vast amounts of un-curated data (often virtually the whole internet), which is then subsequently refined via high quality data from a domain of interest. What is distinct of foundation models is that they are not confined to a single task but can be applied to a broad range of downstream tasks (see also Bommasani et al., 2021). To illustrate, think of how new large language models are able to write texts on different topics, at varying length, style, and sophistication.

Nevertheless, while their performance is often unprecedented, said models raise various reliability concerns – e.g., they are prone to fabricate facts that may seem credible at first glance. More interesting, however, is that they also raise profound conceptual problems that challenge the standard paradigm of evaluation in machine learning: if the foundation core has been trained by the whole internet, is it even possible to test its performance on *unseen* data? Can we still draw a meaningful distinction between data from within and outside of the training distribution? How can we obtain precise performance estimates if there are no clear bounds to the scope of application? And the list continues. Another interesting development in the technical literature is to test the *knowledge* of foundation models in exam-style questions (for example, see Singhal et al., 2023). Arguably, this entails a shift in the reliability assessment of machine learning models, from a mere focus on predictive accuracy towards their reasoning capacities.

In the end, we hope that is clear that the issues of reliability will take center stage for years to come in the debate. And we hope that this paper proves to be a useful point of reference for future work in the epistemology of machine learning.

---

[i] See Buckner (2019) for an introduction to deep learning.

[ii] For an excellent introductory article with more technical detail, see Von Luxburg and Schölkopf (2011). For a book-length introduction, see Shalev-Shwartz and Ben-David (2014). For philosophical connections with the problem of induction, see Harman and Kulkarni (2012) and Sterkenburg and Grünwald (2021). For connections with Popper and falsificationism, see Corfield et al. (2009).

[iii] We have in mind "concentration inequalities" such as Hoeffding's inequality.

[iv] In statistics, this is known as the problem of valid post-selection inference.

[v] There are many notions of capacity, but the most prominent of these is the VC dimension (Corfield et al, 2009).

[vi] Some philosophers and machine learners interpret these results as a mathematical expression of Ockham's razor (Steel, 2009; Sterkenburg, 2023). If capacity is a measure of complexity, then the advice of the theory is not to "live beyond your sample size" by searching a more complex function class than you can afford. But this is a rather different recommendation than the usual exhortation to select the simplest hypothesis compatible with the data.

[vii] We will discuss the relationship between xAI and reliability in machine learning in section 5.

[viii] There is also a rich debate regarding the nature of ML opacity itself (Boge, 2022, Creel, 2020), but for this we direct our readers to (Beisbart and Räz, 2022; Buchholz 2023).

[ix] Though Sullivan (2023) argues that ML models can be assessed in a similar way as simulations, or 'toy models.'

[x] There is a rich debate on whether there is actually a difference between opacity in human reasoning and the reasoning of machine learning models and whether this difference matters. Call this the 'double-standard problem' (Günther and Kasirzadeh, 2022; Peters, 2023; Zerilli, 2019).

References:

Beisbart, C. (2012). How can computer simulations produce new knowledge? *European Journal for Philosophy of Science*, *2*, 395–434.

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

Bellamy, D., Hernán, M. A., & Beam, A. (2022). A structural characterization of shortcut features for prediction. *European Journal of Epidemiology*, *37*(6), 563-568.

Belkin, M., Hsu, D., Siyuan M., & Soumik, M. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, *116(32),* 15849-15854.

Belkin, M. (2021). Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, *30*, 203-248.

Boge, F. J. (2021). Why trust a simulation? Models, parameters, and robustness in simulation-infected experiments. *The British Journal for the Philosophy of Science*.

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

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Buchholz, O. (2023). A Means-End Account of Explainable Artificial Intelligence. *Synthese*, *202*(2), 33.

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy compass*, *14*(10), e12625.

Buckner, C. (2020). Understanding adversarial examples requires a theory of artefacts for deep learning. *Nature Machine Intelligence*, *2*(12), 731-736.

Buijsman, S. (2023). Over What Range Should Reliabilists Measure Reliability?. *Erkenntnis*, 1-21.

Corfield, D., Schölkopf, B., & Vapnik V. (2009), Falsification and Statistical Learning Theory: Comparing the Popper and Vapnik-Chervonenkis dimensions. *Journal for General Philosophy of Science, (40)*, 51-58.

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science*, *87*(4), 568-589.

Mayo-Wilson, C. (2018). Epistemic closure in science. *Philosophical Review*, *127*(1), 73-114.

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

Duede, E. (2022b). Deep learning opacity in scientific discovery. *Philosophy of Science*, 1-13.

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

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, *47*(5), 329-335.

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., ... & Saria, S. (2021). The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, *385*(3), 283-286.

Fleisher, W. (2022). Understanding, idealization, and explainable AI. *Episteme*, *19*(4), 534-560.

Freiesleben, T. (2022). The intriguing relation between counterfactual explanations and adversarial examples. *Minds and Machines*, *32*(1), 77-109.

Freiesleben, T., & Grote, T. (2023). Beyond generalization: a theory of robustness in machine learning. *Synthese*, *202*(4), 109.

Garg, S., Balakrishnan, S., & Lipton, Z. (2022). Domain adaptation under open set label shift. *Advances in Neural Information Processing Systems*, *35*, 22531-22546.

Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665-673.

Genin, K., & Grote, T. (2021). Randomized controlled trials in medical AI: A methodological critique. *Philosophy of Medicine*, *2*(1), 1-15.

Goldman, A. (1979). What Is Justified Belief?. *Justification and Knowledge: New Studies in Epistemology*, George S. Pappas (ed.), Dordrecht: Reidel, 1–25

Goldman, A., & Beddor, B. (2021). Reliabilist Epistemology. *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2021/entries/reliabilism/>.

Günther, M., & Kasirzadeh, A. (2022). Algorithmic and human decision making: for a double standard of transparency. *AI & SOCIETY*, 1-7.

Harman, G., & Kulkarni, S. (2012). *Reliable reasoning: Induction and Statistical Learning Theory*. MIT Press.

Holm, S. (2023). Statistical evidence and algorithmic decision-making. *Synthese*, *202*(1), 28.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, *32*.

Khosrowi, D. & van Basshuysen, P. (forthcoming). Making a murderer. How risk assessment tools may produce rather than predict criminal behavior. *American Philosophical Quarterly*.

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*.

Krishnan, M. (2020). Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, *33*(3), 487-502.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., ... & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473.

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press.

Parker, W. S. (2020). Model evaluation: An adequacy-for-purpose view. *Philosophy of Science*, *87*(3), 457-477.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., & Hardt, M. (2020). Performative prediction. In *International Conference on Machine Learning* (pp. 7599-7609). PMLR.

Peters, U. (2023). Explainable AI lacks regulative reasons: why AI and human decision-making are not equally opaque. *AI and Ethics*, *3*(3), 963-974.#

Räz, T., & Beisbart, C. (2022). The importance of understanding deep learning. *Erkenntnis*, 1-18.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206-215.

Schmidt, E., Sesing-Wagenpfeil, A., & Köhl, M. A. (2023). Bare statistical evidence and the legitimacy of software-based judicial decisions. *Synthese*, *201*(4), 134.

Shalev-Shwartz, S., & Ben-David, S. (2014) *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172-180.

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180-186).

Slack, D., Hilgard, A., Lakkaraju, H., & Singh, S. (2021). Counterfactual explanations can be manipulated. *Advances in neural information processing systems*, *34*, 62-75.

Steel, D. (2009). Testability and Ockham's Razor: How Formal and Statistical Learning Theory Converge in the New Riddle of Induction. *Journal of Philosophical Logic, (38)*, 471-489.

Sterkenburg, T., & Grünwald, P. (2021). The no-free-lunch theorems of supervised learning. *Synthese, 199(3-4)*, 9979-10015.

Sterkenburg, T. (2023). Statistical Learning Theory and Occam's Razor: The Argument from Empirical Risk Minimization. Perprint. https://philsci-archive.pitt.edu/22259/

Sullivan, E. (2022a). Understanding from machine learning models. *The British Journal for the Philosophy of Science*.

Sullivan, E. (2022b). Inductive Risk, Understanding, and Opaque Machine Learning Models. *Philosophy of Science*, *89*(5), 1065-1074.

Sullivan, E. (2023). Do ML models represent their targets?. *Philosophy of Science*, 1-14.
Zednik, C., & Boelsen, H. (2022). Scientific exploration and explainable artificial intelligence. *Minds and Machines*, *32*(1), 219-239.

Vandenburgh, J. (2023). Machine Learning and Knowledge: Why Robustness Matters. *arXiv preprint arXiv:2310.19819*.

Von Luxburg, U., & Schölkopf, B. (2011) Statistical learning theory: Models, concepts, and results. *Handbook of the History of Logic, 10*. North-Holland, 651-706.

Watson, D. S. (2022). Conceptual challenges for interpretable machine learning. *Synthese*, *200*(2), 65.

Winsberg, E. (2003). Simulated experiments: Methodology for a virtual world. *Philosophy of Science*, *70*, 105–125.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: is there a double standard?. *Philosophy & Technology*, *32*, 661-683.