The Benchmarking Epistemology: Construct Validity for Evaluating Machine Learning Models

Timo Freiesleben[1*†] Sebastian Zezulka[2*]

[1] LMU Munich, MCMP & MCML [2] University of Tübingen

Abstract

Predictive benchmarking, the evaluation of machine learning models based on predictive performance and competitive ranking, is a central epistemic practice in machine learning research and an increasingly prominent method for scientific inquiry. Yet, benchmark scores alone provide at best measurements of model performance relative to an evaluation dataset and a concrete learning problem. Drawing substantial scientific inferences from the results, say about theoretical tasks like image classification, requires additional assumptions about the theoretical structure of the learning problems, evaluation functions, and data distributions. We make these assumptions explicit by developing conditions of *construct validity* inspired by psychological measurement theory. We examine these assumptions in practice through three case studies, each exemplifying a typical intended inference: measuring engineering progress in computer vision with ImageNet; evaluating policy-relevant weather predictions with WeatherBench; and examining limitations of the predictability of life events with the Fragile Families Challenge. Our framework clarifies the conditions under which benchmark scores can support diverse scientific claims, bringing predictive benchmarking into perspective as an epistemological practice and a key site of conceptual and theoretical reasoning in machine learning.

Keywords: *Benchmarks; Construct Validity; Measurement; Machine Learning; Prediction*

1 Introduction

In the early days of artificial intelligence research, the Defence Advanced Research Projects Agency (DARPA) was one of its main funders. Not unlike today, researchers promised to solve complex and abstract engineering tasks like language translation or object recognition within just a few months or years, ultimately never coming close to delivering on their promises. A revealing episode played out

*Equal contribution.
†Corresponding authors: timo.freiesleben@lmu.de; sebastian.zezulka@uni-tuebingen.de.

in 1966, when Minsky and Papert, whose research was strongly supported by DARPA at the time, assigned a group of students a project aimed at artificially constructing "a significant portion of a visual system" over the course of just one summer (Papert, 1966). Following substantial governmental cuts in funding and disappointed expectations, DARPA employees were faced with the difficult question of assessing whether progress has been made on any of the problems.

DARPA answered this question in the late 1980s by adopting the *Common Task Framework* (Liberman, 2010, Donoho, 2017). Instead of the grand claims about solving artificial general intelligence that dominated the early days of AI, the Common Task Framework shifted the focus to concrete (usually supervised) learning problems. Data for building models to solve these problems was made publicly available and, crucially, performance was evaluated on unseen hold-out data. DARPA had both the authority and the resources to establish the Common Task Framework as the new norm in artificial intelligence research, thereby laying the groundwork for modern predictive benchmarking and fostering the shift from symbolic AI towards machine learning.[1]

A *predictive benchmark* has four components: (1) a *learning problem* specifying the prediction target, (2) a standardized *dataset* split in a training and an evaluation set, (3) an *evaluation metric*, and (4) a public *leaderboard*, ranking the models by the scores they obtained.[2] Take, as an example, the famous MNIST benchmark (Lecun et al., 1998). It consists in a seemingly simple learning problem: given a small greyscale image of a handwritten digit, identify which digit it represents from zero to nine (Yadav and Bottou, 2019). Performance of models is evaluated by their accuracy on the evaluation set. For some years, MNIST was a proving ground for new machine learning techniques and a public leaderboard still tracks model results. But times have changed and with current methods reaching about 99.87% accuracy, the benchmark is considered too easy and essentially "solved."

Predictive benchmarks have since evolved from a tool for making funding decisions into the dominant methodology for evaluating machine learning models and coordinating research programmes around shared problems.[3] They have become the "iron rule" of machine learning (Hardt, 2025, Ch. 1): justification for algorithmic innovation largely depends on outperforming the state of the art as measured by a relevant benchmark. With the growing influence of machine learning over the last decade, predictive benchmarks have transcended their engineering roots and are emerging as a relevant methodology in the empirical sciences. This became very clear in 2024, when benchmark-driven research earned its first Nobel Prize: John Jumper and Demis Hassabis were honoured for their protein structure prediction models, which substantially outperformed all previous methods on the CASP[4] benchmark challenge (Jumper et al., 2021). And it is not only structural biology–prominent predictive

[1] The Common Task Framework echoed ideas introduced as early as 1959, when Bill Highleyman outlined many of the principles in his PhD thesis on handwritten character recognition (Orr and Crawford, 2024). Highleyman also shared his dataset with other researchers – by mail.

[2] Recent benchmarks like SuperGLUE (Wang et al., 2019) or HELM (Bommasani et al., 2023) often encompass several learning problems, datasets, and metrics.

[3] Hardt and Recht (2022, Ch. 8), Paullada et al. (2021), and Koch and Peterson (2024) discuss the historical development of predictive benchmarks.

[4] The 'Critical Assessment of protein Structure Prediction' (CASP) challenge, dating back to 1994, is among the oldest scientific benchmark challenges (Donoho, 2024).

benchmarks now exist across many scientific fields, including meteorology (Rasp et al., 2020), medicine (Irvin et al., 2019), neuroscience (Schrimpf et al., 2018), materials science (Dunn et al., 2020), legal studies (Guha et al., 2023), and sociology (Sivak et al., 2024).

Despite the central role that benchmarks play in machine learning and in the sciences, they have received little attention from philosophers. Philosophers of machine learning have primarily concentrated on issues of epistemic opacity (Creel, 2020, Boge, 2022, Sullivan, 2022, Duede, 2023) and explainable artificial intelligence (Zednik and Boelsen, 2022, Watson, 2022, Buchholz and Grote, 2023, Freiesleben et al., 2024). A rich philosophical literature examines (big) data along its representational and relational dimensions (Bogen and Woodward, 1988, Leonelli, 2015, 2020), but not in the context of benchmarks. Formal epistemologists, providing a priori guarantees for learning algorithms, have typically adopted a theoretical perspective (Herrmann, 2020, Sterkenburg and Grünwald, 2021, Grote et al., 2024b, Sterkenburg, 2024), largely overlooking the empirically driven methodology of predictive benchmarking that underpins much of contemporary machine learning research. Yet, it is precisely because benchmarking structures entire research programs and informs scientific inquiry and policy decisions that it demands philosophical scrutiny. Without an epistemic basis, researchers risk drawing misleading or irreproducible inferences. This paper contributes such a basis, showing how predictive benchmarks can serve as epistemic tools capable of supporting a wide range of scientific inferences, if researchers explicitly consider the relevant condition of *construct validity*. Specifically, we address the following question:

> *How can benchmark scores support substantive scientific inferences within machine learning and across the sciences?*

We argue that, analogous to psychometric and educational tests, predictive benchmarks are best understood as measurement tools (Section 2). The resulting measurements, $M(F)$, quantify the predictive performance of a set of models F on a given learning problem, if conditions of *internal* validity, like the availability of independently and identically distributed (*i.i.d.*) samples, which are independent of the evaluated models, hold. As in psychometric and educational testing, drawing more substantial inferences I from the measurements requires additional assumptions A to render the inferences logically valid, that is,

$$M(F) \ \wedge \ A \ \rightarrow \ I.$$

The study of these assumptions is often referred to as *construct validity* (Kane, 2013, Messick, 1995). We adapt the *argument-based* framework for construct validity to the context of predictive benchmarks, comprising four steps: (1) defining the intended inference, (2) specifying necessary validity conditions, (3) providing evidence for and against the conditions, and (4) constraining the inference. We specify validity conditions under which substantial scientific inferences can be drawn from benchmark scores, establishing predictive benchmark as key sites of conceptual and theoretical reasoning in machine learning and the sciences more broadly (Section 3).

We illustrate this framework with three case studies that capture different stakeholders interpreting benchmark results, the diversity of inferences they intend to draw, and the plurality of construct

validity concerns they raise. The first case study examines the ImageNet benchmark from computer vision, where the goal is to map images to their assigned classes (Section 4). ImageNet serves as a proxy signal for engineers to measure methodological progress on the abstract task of image classification. However, we argue that to fulfill this inferential role, ImageNet must operationalize image classification as a theoretical task (content validity) and ensure that its measurements align with relevant alternative operationalizations of image classification (external validity). The second case study is about WeatherBench, where the target is forecasting global weather events (Section 5). Beyond serving as a proxy for methodological development, WeatherBench may also function as a criterion for policymakers when selecting between forecasting models for deployment. We contend that this interpretation of WeatherBench results is justified only if benchmark scores accurately reflect the consequences of downstream decisions in the respective deployment context (consequential validity). While we have substantive theory in meteorology that implies theoretical limitations to predictions, this is different in the third case study from the social sciences (Section 6). The goal of the Fragile Families Challenge is to predict life events of young people. The resulting low predictive performances of all models led researchers to consider whether life outcomes are fundamentally unpredictable. We argue that this interpretation requires ruling out alternative explanations for the results, such as a small sample size or unmeasured predictors (auxiliary validity). We conclude in Section 7 with an outlook on the relevance of construct validity–particularly convergent and divergent validity–in assessing the capabilities of large language models (LLMs) and the social epistemology of predictive benchmarking.

2 Predictive benchmarks are measurement tools

Predictive benchmarks are best understood as measurement tools that measure the predictive performance of machine learning models on specific learning problems.[5] In this sense, they can be seen analogously to psychological and educational tests (Blili-Hamelin and Hancox-Li, 2023, Cardoso et al., 2020). Psychological tests are designed to measure latent variables, often psychological attributes like intelligence, by using a set of test items. The idea is to systematically relate the observable performance on the test items to the hypothesised underlying latent psychological attribute using a statistical model (Millsap, 2012). The salient, albeit problematic, example is intelligence, which is commonly measured using a single-dimensional score obtained from an IQ test (Glymour, 1998, Cave, 2020). All components of psychological tests have a counterpart in predictive benchmarks: the data instances of the benchmark correspond to the test items and the test takers are the (machine learning) models of which scientists may want to infer latent properties. Consequently, as in psychological and educational testing, the resulting benchmark scores must be interpreted in light of the relevant validity conditions to support inferences.

But what are we measuring with benchmarks? Let us begin with what may be called the *standard case* in machine learning, where scientists use benchmark scores to infer the performance of models on

[5] Mussgnug (2022) analyses the outputs of individual machine learning models as measurements rather than the epistemological practice of predictive benchmarks.

randomly sampled, unseen data. In this setting, little measurement theory is required, as the problem can be formulated as a statistical estimation task: we estimate the expected error using the empirical error as its estimator. This procedure is commonly referred to as the *holdout method* (Grote et al., 2024b).

Evaluating a set of models on a benchmark yields numerical scores, the *empirical error* as defined by the evaluation metric.[6] For a machine learning model f, an evaluation dataset D_e with features x_i and labels y_i, and an evaluation metric ℓ, the empirical error is defined as

$$\hat{R}(f) := \frac{1}{\mid D_e \mid} \sum_{(x_i, y_i) \in D_e} \ell\left(f(x_i), y_i\right).$$

The empirical error is the average prediction error that model f makes on the items in the evaluation dataset. Benchmark scores for a set of models F are given by their respective empirical errors, from which one can also derive an ordinal ranking of models.

To *interpret* the scores, we must connect them to the theoretical quantity or latent variable we intended to measure. In the standard case, this measurement target is the *expected error*. For a model f and an evaluation metric ℓ, the expected error is defined as

$$R(f) := \mathbb{E}_{X,Y}\left[\ell\left(f(X), Y\right)\right],$$

where (X, Y) are random variables drawn from a fixed but unknown data distribution, from which the evaluation data D_e is assumed to be sampled. The expected error captures how the models would, on average, perform on new instances from this population, effectively representing the model's performance were we able to observe an infinite number of data points. This definition specifies the intended inference in the standard case.

Inferring the expected errors from benchmark scores requires supporting assumptions – we will call them conditions of *internal* validity – that connect the two. The validity conditions must provide a sound inferential bridge between the benchmark scores and the intended interpretation. In this case, necessary and sufficient conditions are that (i) the evaluation data is independent of the models, implying for example that one cannot use the evaluation data for model training; (ii) the evaluation data are statistically independent samples (x_i, y_i) drawn from one identical distribution over (X, Y), the *i.i.d.* assumption; and (iii) the evaluation data is sufficiently large so that the empirical error provides a reasonable approximation of the expected error.[7]

If the three assumptions hold, we can effectively interpret the benchmark scores as an implementation of the holdout method and, therefore, the resulting empirical errors as valid measures of the expected errors. This is a strong result and already allows us to draw important inferences from

[6] The evaluation metric does not need to coincide with the loss function used to train the model, nor must the model be trained for the learning problem on which it is evaluated.

[7] Under (i) and (ii), finite sample guarantees for the discrepancy between the empirical and expected error can be given for different sample sizes (Grote et al., 2024b). For a discussion on the interpretation of statistical assumptions, see Zhao (2025).

benchmark scores about the expected error of a model on the learning problem; how the models perform relative to each other on this learning problem; and even that certain learning algorithms, architecture designs, or hyper-parameter choices lead to models with lower expected errors – on this learning problem and in comparison to the competitors.

While these inferences are relevant, in science as well as in machine learning, we often want to draw more involved inferences from benchmark scores, which the three conditions of internal validity alone cannot support. The diverse inference targets go beyond statistical standards of estimation and require a richer structure of validity considerations, potentially changing also the estimation target. These conceptual, theoretical, and normative considerations are given a rich structure under the umbrella of *construct validity*. This literature provides a fruitful structure for the analysis of the diverse uses of predictive benchmarks and the required conditions of validity. As we have seen, *internal* validity allows us to accommodate the standard case in this framework. Analysing predictive benchmarks as measurement tools brings into focus the conditions of validity necessary so that the benchmark scores can support substantive inferences.

3 Construct validity for score interpretation

There are two main approaches to construct validity discussed in the literature (Alexandrova and Haybron, 2016, Zhao, 2023). The *correspondence account* treats validity as a property of the measurement tool or outcome itself, for example, establishing a causal relationship between the measurement and the latent variable (Borsboom et al., 2004). By contrast, the *argument-based account* understands validity as a property of the interpretation of scores rather than of the measurement, establishing a logical relationship between the measurement and the intended inference (Kane, 2013). Since we adopt an epistemological rather than an ontological perspective on benchmarks and investigate them as epistemic tools for drawing valid inferences, we follow the argument-based account of validity.

The argument-based account of validity allows for establishing inferences through a four-step procedure, which was first introduced by Messick (1995) and is now widely recognized as a standard in psychological and educational measurement (American Educational Research Association et al., 2014):

1. **Define the intended inference.** The intended inference to be drawn from the benchmark scores must be made specific. This includes a specification of the relevant theoretical constructs and the scope of the inference.

2. **Specify validity conditions.** Validity conditions must be specified that, together with the measured performance on the learning problem, make the intended inference valid. These secure, for example, the relevance of the operationalisation for the construct of interest or that the statistical assumptions are met.

3. **Provide evidence for and against the assumptions.** Researchers must provide evidence of whether and to what degree the validity conditions hold. Here, evidence is understood broadly

and may include benchmark scores for relevant subgroups, results from other studies, and conceptual or domain-specific theoretical insights.

4. **Constrain the inference.** In light of the evidence, the validity conditions and, consequently, the intended inference will only be supported to some degree. There will be evidence bolstering some and undermining other assumptions. Scientists must then constrain the scope and content of the intended inference accordingly.

Establishing an inference from benchmark scores within this framework comes with two requirements: (1) the intended inference must be formulated precisely enough to derive the relevant validity conditions and (2) given the measurement and the validity conditions, the intended inference must be logically valid. Both requirements can be difficult to meet. This is why we specify validity conditions for common types of inferences in Table 1. The following sections demonstrate the framework in action through case studies. Before that, we show how construct validity offers a unifying framework and shared language for discussions about predictive benchmarks and their limitations.

We are not the first to adopt a validity-based perspective on predictive benchmarks. Yee (2024) examines construct legitimacy, criterion validity, and construct validity for evaluating machine learning models in the case of counterterrorism. Closely related to our approach is Salaudeen et al. (2025), who have recently proposed a "claim-centred framework" for the evaluation of general-purpose AI systems. The attribution of *capabilities* to AI systems based on benchmark scores has generally spurred interest in issues of psychometric testing and validity (Wang et al., 2023, Sühr et al., 2025, Ye et al., 2025).

Researchers in machine learning have raised a number of concerns with the practice of benchmarking, in particular, whether the re-use of evaluation data over time leads to inconsistent scores (Blum and Hardt, 2015, Yadav and Bottou, 2019, Recht et al., 2019). This raises a problem of *internal validity* and ensures that benchmark scores are not sample-specific. We have already discussed the corresponding validity conditions in Section 2, which allow the interpretation of the empirical error as the expected error. Others have argued that performance improvements on specific benchmarks often fail to transfer to related prediction tasks, even within the same domain. That is, results depend on the particular learning setup, performance metric, and data distribution (Torralba and Efros, 2011, Herrmann et al., 2021, 2024). These issues concern *external validity*: we want benchmark scores to be *robust*[8] under *relevant* variations of the predictive benchmark. A set of relevant benchmarks $\mathcal{B}$, over which we expect models to perform robustly, extensionally specifies the domain of a task. *Content validity*, in turn, establishes the theoretical domain and intensional description of the task.

Any theoretical task can be operationalized in multiple ways by a benchmark (Liao et al., 2021). Schlangen (2021) introduces the key distinction between the *intensional* task description, like "language translation," and its *extensional* instantiation in a specific benchmark, such as English texts paired with their German translations. He argues that drawing inferences from benchmark scores requires validating that the extensional formulation adequately operationalizes the intensional task. Similarly,

[8]Robustness here is understood in causal terms: a benchmark score should remain sufficiently stable under changes to a modifier, such as an alternative operationalization of the task (Freiesleben and Grote, 2023).

Raji et al. (2021) maintain that only well-specified intensional tasks can be faithfully operationalized by benchmarks–a condition that many general-purpose language benchmarks, such as Super General Language Understanding Evaluation (SuperGLUE), fail to meet.[9] All these concerns can be understood as issues of *content validity*, which establishes the inferential link between measurements of a benchmark and claims about the respective intensionally defined theoretical task.

Beyond these issues, Table 1 also includes conditions for *consequential* and *auxiliary* validity. The former reflects the normative implications of using benchmark scores for decision-making, covering how the scores reflect the relevant utilities and potential harms arising from their use. The latter, a broader category, concerns the integration of score interpretations into the theoretical framework of the field. In Section 6, we elaborate on this with the example of inferences about the in-principle predictability of life events.

Under the argument-based view, validity is gradual and its assessment evolves over time (Kane, 2013). More substantial or wide-ranging inferences from benchmark scores also require stronger validity conditions. With more conceptual and theoretical implications of an inference, we are, for example, moving from concerns of internal validity to content or consequential validity. Nevertheless, this does not imply a strict hierarchy between types of validity and validity conditions. For instance, in policy contexts involving uniquely operationalizable tasks, external validity may be less relevant than content and consequential validity.

In the following sections, we apply the proposed framework to derive specific validity conditions for three intended inferences across three case studies: ImageNet in computer vision, WeatherBench in meteorology, and the Fragile Families Challenge in sociology.

4 ImageNet: Do we make progress on image classification?

The ImageNet Challenge[10] is one of the most influential predictive benchmarks in the history of computer vision and machine learning more broadly (Deng et al., 2009). It was at the centre of the deep learning revolution, when the convolutional neural network AlexNet (Krizhevsky et al., 2012) outperformed the competing models by about a 10% margin in 2012. Moreover, key algorithmic innovations in deep learning, such as dropout (Srivastava et al., 2014) and batch normalization (Ioffe, 2015), were adopted due to their effectiveness in improving performance on ImageNet.

The ImageNet Challenge, and particularly its classification component, became so paradigmatic that it effectively served as a *measure of progress* in computer vision. Between 2010 and 2015, researchers were able to reduce the reported error rates from 28.2% in 2010 to 3.57% in 2015 (Nguyen et al., 2018). But how can performance gains on ImageNet be taken as evidence that models have improved on image classification more generally? The four-step approach introduced above allows us to evaluate when this inference is valid by analysing the underlying validity conditions.

[9] This limitation has been widely discussed in the literature (Dorner et al., 2023, Millière and Buckner, 2024, Zhang and Hardt, 2024, Grote et al., 2024a).

[10] Known as ILSVRC, the ImageNet Large Scale Visual Recognition Challenge.

Validity	Conditions	Inference
Internal (Sec 4)	(i) independence of D_e & f (ii) D_e is *i.i.d.* sample (iii) D_e sufficiently large	If f has empirical error/rank s on benchmark B, then its expected error/rank on B is also s.
External (Sec 4)	(i) robust performance across related learning problems & distributions (ii) scores informative across metrics	If f has expected error/rank s on B, then it also has s on other relevant benchmarks $B' \in \mathcal{B}$.
Content (Sec 4)	(i) learning problem describes task (ii) data represents task classes (iii) ℓ reflects task performance	If f has expected error/rank s across benchmarks $\mathcal{B}$, then it has s on the theoretical task T.
Consequential (Sec 5)	(i) ℓ reflects utility (ii) task meets application requirements	If f has performance/rank s on task T, then it has utility (rank) s in application A.
Auxiliary (Sec 6)	(i) models sufficiently diverse (ii) all relevant features measured	If a set of models F have minimal expected error s on task T, then s describes optimal performance on the task T.

more theory-laden inferences ↓

Table 1: The table summarizes the five types of validity discussed in this paper, specifying the associated validity conditions and exemplary inferences they support. Here, f denotes the model, D_e the evaluation data, and ℓ the evaluation metric. The arrow on the right indicates that, as the intended inferences become stronger and more theoretical, an increasing number of validity conditions must be satisfied. This does *not* establish a strict hierarchy between types of validity and validity conditions.

4.1 The intended inference

Computer vision researchers may aim to infer that *the error rates of models on ImageNet reflect their error rates on image classification tasks generally.* If this inference holds, improvements on ImageNet can legitimately be interpreted as advances in image classification and, by extension, in computer vision. This would also justify the community's intense concentration on ImageNet.

4.2 Specifying validity conditions & providing relevant evidence

Given the intended inference, step two now requires specifying the validity conditions for that inference. Step three involves supporting them with relevant evidence. To avoid jumps between the conditions and their respective justifications, we present each validity condition together with a critical discussion of its supporting and challenging evidence. Necessary and sufficient conditions for the intended inference are *internal*, *external*, and *content* validity.

Internal validity. Internal validity conditions concern inferences of expected errors from benchmark scores. As discussed in Section 2, to interpret the empirical error as the expected error, we require that

(i) the model is independent of the evaluation data,

(ii) statistically independent samples from an identical distribution (*i.i.d.*), and that

(iii) the evaluation dataset must be sufficiently large.

Critically, the ImageNet data was used over many years, during which researchers could adapt to prior results and successful model architectures. This sequential re-use violates the independence of the models and the evaluation data (Blum and Hardt, 2015). In an empirical study, Recht et al. (2019) therefore recreated ImageNet, mirroring the original data collection strategy, and re-evaluated successful models like AlexNet and ResNet on the new data. By construction, the new data is independent of the models.[11] They indeed found that models performed about $11 - 14\%$ worse on the re-created data, but that model rankings remained intact. Interestingly, the performance gap narrowed for more recent, higher-accuracy models. The authors, therefore, attributed the observed drops not to adaptivity, which should at least not decrease over time, and instead to subtle distribution shifts in data collection. We return to this when discussing external validity.[12]

Second, the *i.i.d.* assumption is not directly testable. It serves as a guideline for data collection and puts strong conceptual limitations on the population for which we get valid estimates. ImageNet

[11] Several studies have since then examined the empirical impact of this information leakage for other benchmarks (Yadav and Bottou, 2019, Roelofs et al., 2019, Miller et al., 2020).

[12] Beyer et al. (2020) and Shirali and Hardt (2025) observe similar performance drops of the ImageNet models on an alternative image dataset. However, by providing independent multi-label data from experts, Beyer et al. (2020) show that while model rankings are robust, the relative improvements with respect to the real labels decrease. This indicates that models trained and evaluated on ImageNet data may have problems when generalizing to even slightly different distributions (Tsipras et al., 2020). Supporting this result, Geirhos et al. (2018) and Hendrycks and Dietterich (2019) show that the predictive performance of ImageNet models drops strongly under various data corruptions and perturbations.

images were sourced from web search engines and labelled by at least ten Mechanical Turk workers per image (Deng et al., 2009, Denton et al., 2021). Duplicates and labels on which workers did not agree were filtered out to ensure high quality, but this introduces potential sampling biases (Torralba and Efros, 2011, Shirali and Hardt, 2025) which we discuss further under content validity.

Third, the evaluation set is large enough to yield reliable error estimates with roughly 100 images per class. Russakovsky et al. (2015) empirically confirm this by estimating bootstrap-based confidence intervals, showing low variation in the resulting scores.

External validity. External validity supports inference beyond the original context. For example, inferences from ImageNet scores to other data distributions, learning problems, or evaluation metrics. Consequently, the model performance must be *robust* to alternative operationalization of the task. This holds especially for benchmarks that share high content validity for the same task, as we discuss below. We distinguish two conditions for external validity.

(i) The performance scores from ImageNet are similar across relevant alternative learning problems and data distributions.

(ii) The performance scores for a given evaluation metric are informative about scores from other relevant evaluation metrics.

There is substantial evidence that ImageNet scores robustly *rank* the performance of models on a broad class of image classification problems, while absolute performance scores drop drastically even under small changes. Kornblith et al. (2019) evaluated 16 models, which were trained on ImageNet, on other classification tasks like CIFAR-10/100. The ImageNet scores showed near-perfect rank preservation and were predictive of the performance on the tasks.[13] These findings indicate stronger improvements in cross-task generalization compared to early results Torralba and Efros (2011).

New image datasets like LAIONet (Shirali and Hardt, 2025) and ImageNot (Salaudeen and Hardt, 2024) provide distributional alternatives to ImageNet. Like Recht et al. (2019), compared with ImageNet scores, they observe large absolute performance drops but stable rankings. In particular, Salaudeen and Hardt (2024) compare six eminent learning algorithms that have led to improved performances on ImageNet and show that they realize similar performance *gains* when trained on ImageNot. Nevertheless, small corruptions or perturbations in the data cause sharp performance drops, raising concerns about the domain on which they are robust (Goodfellow et al., 2014, Geirhos et al., 2018, Hendrycks and Dietterich, 2019, Tsipras et al., 2020).

Turning to the evaluation metric, we again find evidence that model rankings are largely consistent, especially across top-1, top-5, and multi-label accuracy (Russakovsky et al., 2015, Shankar et al., 2020). Naturally, the different metrics shift the absolute performance scores. However, when models were evaluated on expert-provided multi-labels, allowing for several labels per image, the relative *gains* between the models diminish (Beyer et al., 2020). This indicates metric sensitivity at finer levels of comparison than ordinal rankings.

[13] The correlation is $r = 0.55$, rising to $r = 0.86$ for large datasets (Kornblith et al., 2019).

Content validity. Content validity is crucial to draw inferences about the underlying theoretical construct. It refers to the extent to which measurements adequately cover and represent the domain of the perspective construct. In our context, it evaluates how the ImageNet benchmark operationalises image classification. This requires that

(i) the benchmark captures essential aspects of the theoretical task, e.g., image classification,

(ii) the dataset adequately represents the relevant classes, and

(iii) the evaluation metric meaningfully reflects predictive performance.

First, the learning problem of ImageNet is to map images that have been assigned a unique label to five out of a 1, 000 classes. This operationalisation of the task aligns with established practice in image classification (Torralba and Efros, 2011). The 1, 000 classes in the challenge were randomly chosen[14] from more than 5, 000 classes in the full ImageNet database, structured by the WordNet hierarchy, which provides a widely accepted ontological basis (Miller, 1995, Deng et al., 2009). Importantly, the class selection is very task-specific. For example, 120 of the 1, 000 classes are different dog breeds. The curators made this choice to increase between-image difficulty, requiring models to pick up on fine-grained structures. At the same time, the overrepresentation can distort inferences about general image classification performance.

Second, the curators explicitly aimed to capture visual diversity along dimensions like size, pose, and lighting (Deng et al., 2009). Nevertheless, web-data limitations, image exclusion criteria, annotator priming, and high-agreement filtering during labelling, which can limit diversity, remain as sources of task biases (Torralba and Efros, 2011, Tsipras et al., 2020, Shirali and Hardt, 2025). Probably the most fundamental limitation is that each image only gets a single label, even when multiple may apply. Shankar et al. (2020) estimate that approximately 20% of images of ImageNet have multiple plausible labels. This reflects practical annotation constraints, but reduces the benchmark's fidelity to general classification.

Third, the ImageNet Challenge uses the top-5 metric: a prediction is correct if the assigned label is among the model's five guesses. This compensates for the challenge that often multiple plausible labels exist, but can also lead to distortions. For instance, the prediction {"miniature schnauzer," "tiger," "pirate," "shovel," "fan"} scores perfectly for an image labeled as "miniature schnauzer," despite four very implausible guesses. In contrast, the prediction {"standard schnauzer," "giant schnauzer," "mailbox," "street-sign," "Irish terrier"} counts as an error, even though all labels are plausible for an image showing a dog sitting on a street. Multi-label accuracy would better reflect this classification performance and is standard in other benchmarks (Everingham et al., 2009, 2014).

4.3 Constraining the inference

In light of the discussion, ImageNet benchmark scores cannot be interpreted as indicating the expected error rates of the models on image classification problems in general. Several validity conditions are

[14] With minor manual post-processing to exclude obscure classes (Russakovsky et al., 2015).

violated: empirical error rates risk not being *internally* valid estimates of the expected error due to potential adaptivity. More problematic is the lack of *external* validity due to the high sensitivity to the precise sampling domain, so that scores do not reliably transfer across different benchmarks. On the *content*-level, this is grounded in the fact that the theoretical task 'image classification' is vague beyond the extensional list of various operationalisations, putting strong limitations on valid interpretations of scores as expected errors.

Restricting the scope of the interpretation to model *rankings* instead allows for a more modest and better supported inference. Since models with higher ImageNet performance tend to perform better on other image classification tasks, gains in ImageNet accuracy can be taken to indicate improvements in the theoretical task of image classification. This interpretation has much stronger theoretical and empirical support and is, as Salaudeen and Hardt (2024) point out, sufficient to justify the use of a benchmark like ImageNet to track engineering progress in computer vision.

5 WeatherBench: What forecasting model should we deploy?

In recent years, machine learning and, with it, predictive benchmarking have become prevalent in weather forecasting. The current dominant benchmark is *WeatherBench* (Rasp et al., 2020), which recently got updated to *WeatherBench 2* (Rasp et al., 2024).

As with ImageNet, a key function of WeatherBench is to track algorithmic progress and provide a standard for comparing models in terms of their predictive accuracy, particularly across different modelling paradigms such as physical, hybrid,[15] and purely data-driven approaches. A key difference between meteorology and image classification is the availability of mechanistic theory and physical models about the underlying processes (Dueben et al., 2022). Weather forecasting is also deeply embedded in high-stakes decision-making: accurate forecasts guide critical decisions from issuing early warnings for extreme weather to planning energy market operations. Energy providers rely on weather models to forecast solar radiation and wind speeds in order to coordinate renewable energy supply with backup fossil fuel systems.

GraphCast, a graph neural network, was the first machine learning model to outperform the operational High Resolution Ensemble System (HRES) in several key WeatherBench metrics for medium-range deterministic forecasts (Lam et al., 2023). But can researchers conclude from these results that GraphCast is ready for deployment in high-stakes applications?

5.1 The intended inference

Meteorologists and other stakeholders may aim to infer that *the rankings of models on the application-relevant WeatherBench leaderboards reflect their usefulness in the corresponding practical applications,*

[15] Hybrid approaches use machine learning to find useful parametrizations of coarse-scale weather phenomena so that these data-driven schemes are better constrained by observations and physical principles than traditional, heuristic parametrizations (Jebeile et al., 2023, Kawamleh, 2021).

for example, for energy market planning. If this inference is valid, the benchmark scores can directly be interpreted by the relevant stakeholders as evidence in their decision-making.

5.2 Specifying validity conditions & providing relevant evidence

This inference requires, as before, establishing internal, external, and content validity. In addition, we also demand *consequential validity*. The performance score must reliably support its use as a basis for deployment decisions. To avoid redundancy, we will briefly revisit the first three conditions and focus on consequential validity.

Internal validity. We observe high internal validity. WeatherBench is based on the ERA5 reanalysis dataset curated by the European Centre for Medium-Range Weather Forecasts (ECMWF), which is recognized for its thorough curation and extensive coverage (Dueben et al., 2022). ERA5 integrates global hourly measurements from 1979 to 2024 with model forecasts from HRES to fill gaps and correct errors. The default evaluation set consists of data from the most recent year, which is sufficiently large and representative to provide robust performance estimates (Rasp et al., 2024). Importantly, the evaluation data changes yearly, increasing independence between models and evaluation data.[16] As is standard in time series settings, the data is temporally correlated, violating the *i.i.d.* assumption. For this reason, WeatherBench relies on time-based train-test splits of the data.

External validity. The external validity remains both empirically unclear and conceptually underdeveloped. ERA5 offers a comprehensive global picture of weather events with hourly resolution, $0.25° \times 0.25°$ horizontal resolution, and pressure levels ranging from 1hPa to 1, 000hPa. While the robustness of performance to lower-resolution data has been successfully tested, data is lacking for higher resolutions and other pressure levels (Rasp et al., 2024). Analyses have shown that high-performing machine learning models on WeatherBench fail to replicate the *butterfly effect*, implying that WeatherBench insufficiently reflects the sensitivity of weather states to small changes in initial conditions (Selz and Craig, 2023). External validity under drastically different geophysical or atmospheric conditions has, to the best of our knowledge, not been tested yet. Climate change, of course, is a large-scale distribution shift, and studies indicate that models generalize better if they incorporate climate knowledge (Beucler et al., 2024) or are trained on more recent data (Lam et al., 2023, Rasp et al., 2024).

Content validity. Due to the careful benchmark design, the content validity is high. WeatherBench's prediction tasks (e.g., surface temperature, precipitation, and extreme weather events), feature sets (e.g., atmospheric, surface, and oceanic variables), and forecasting time horizons are selected based on meteorological theory. ERA5 itself aims to provide an unbiased representation of *global* weather events.[17] By design, WeatherBench includes several metrics endorsed by the World Meteorological

[16] Unlike an annual competition, WeatherBench is a scientific benchmark. While teams could overfit on the public evaluation data, this would become evident in subsequent years.

[17] Regional differences in sensor coverage imply that some areas are more accurately represented than others. Additionally, ERA5 may be systematically biased for skewed distributions like precipitation (Bouallègue et al., 2024, Rasp et al., 2024).

Organization and operational weather centres, reflecting both the wide range of relevant aspects of weather events and the diversity of downstream applications that rely on forecasts (Dueben et al., 2022, Rasp et al., 2024).

Consequential validity. Consequential validity connects benchmark performance measurements to real-world deployment decisions. It can be characterized by the following two conditions:

(i) the evaluation metric reflects decision-makers *utility* in the application, i.e., the ethical and practical consequences of correct and incorrect predictions, and

(ii) the benchmark task adequately represents the setup and requirements of the intended application context.

The default evaluations of WeatherBench are not tailored to specific applications. They consist of common metrics that are computed for widely used weather prediction targets, such as 850 hPa temperature, 2 m surface temperature, and 24 h precipitation. While these evaluations may capture utility for some downstream tasks, they fall short in many respects. For instance, they do not assess the physical plausibility of predicted weather states (Bonavita, 2024). This is why high-performing models like GraphCast often produce overly smoothed, blurry forecasts at longer time horizons to hedge against uncertainty (Lam et al., 2023, Bouallègue et al., 2024). Rare and high-impact events like storms or skewed variables like precipitation are often *under*-penalized (Rasp et al., 2024, Bouallègue et al., 2024). Furthermore, models are usually evaluated on point predictions rather than uncertainty estimates, which are essential in many applications (Bradley and Steele, 2015, Moret, 2017, Dueben et al., 2022, Brenowitz et al., 2025).

However, as Rasp et al. (2024) emphasize, WeatherBench should not be seen merely as a single leaderboard competition but as a flexible tool for comparing different modelling approaches across multiple dimensions. WeatherBench allows users to create custom metrics tailored to their specific evaluation needs and to submit requests for extending the benchmark. For example, for tasks like energy market planning, stakeholders might design evaluation metrics that penalize underestimating solar radiation less than overestimating it, since the costs of potential blackouts caused by underestimation are greater than those from temporary oversupply.

Most application contexts differ significantly from the benchmark setup. WeatherBench uses reanalysis data, which is post-processed using physical models. This is an advantage not always available in deployment, where input data may be noisier or less complete (Dueben et al., 2022, Bouallègue et al., 2024). In addition, many real-world applications impose constraints not reflected in the benchmark evaluations. As noted, energy planning depends heavily on uncertainty estimates, which are provided by models like HRES or GenCast (Price et al., 2024), but not by GraphCast. Moreover, WeatherBench currently evaluates models only at ERA5 resolution or lower, limiting models to coarse-grained outputs. Higher-resolution predictions, which are often needed in practice, are currently available only with physical models like HRES IFS (Olivetti and Messori, 2024) and cannot be tested with WeatherBench (Bouallègue et al., 2024).

Legal and operational constraints may also demand interpretable models for accountability and human oversight (McGovern et al., 2019). GraphCast, like most deep learning models, remains opaque, but the same can be said of statistical weather forecasting and climate models (Jebeile et al., 2020). Computational efficiency and the time it takes to produce a forecast are also relevant factors in deployment, which are not tracked by WeatherBench.

5.3 Constraining the inference

WeatherBench scores cannot be used directly to reflect utility in downstream applications. While we observe strong internal and content validity, the consequential validity is highly dependent on the specific application context. In energy planning, for example, the default evaluation metrics of WeatherBench do not capture context-specific utilities or spatial scales. Requirements for deployment in regulated environments, such as uncertainty quantification or interpretability, are largely absent from current benchmark evaluations.

These limitations reflect current practice rather than in-principle constraints. WeatherBench can be adapted to contexts, for example, by incorporating application-specific and uncertainty-aware evaluation metrics (Brenowitz et al., 2025). Even in their current form, WeatherBench scores can inform decision-making. As in climate modelling (Bradley and Steele, 2015, Frigg et al., 2015), this requires decision-makers to acknowledge the limitations of model outputs rather than relying on leaderboard rankings as sufficient grounds for deployment decisions.

6 Fragile Families Challenge: Are life events predictable?

The Fragile Families Challenge (FFC) that took place in 2017 is a predictive benchmark in sociology (Salganik et al., 2019). Benchmarks have the potential to serve as a powerful evaluation tool in the quantitative social sciences, complementing commonly used explanatory in-sample methods (Verhagen, 2022, Shmueli, 2010, Hofman et al., 2017). They allow a comparison of radically different statistical models (Rocca and Yarkoni, 2021), testing model assumptions (Buchholz and Grote, 2023), but also to investigate the theoretical question of the *predictability* of life events. Unlike metrology, where mechanistic theory implies limits to predictability, theory in the social sciences usually lacks such constraints. Benchmark challenges like the FFC, therefore, have not only implications for public policy but also theoretical significance.

The challenge was to predict at least one out of six individual life outcomes of 15-year-old teenagers, such as their grade point average (GPA), a psychological construct of perseverance called *grit*, or whether their household experienced material hardship in the past year (Salganik et al., 2019). It was built on a 15-year longitudinal survey for which more than 4, 200 families were interviewed six times. In the end, 160 models were submitted, ranging from conventional statistical methods to various machine learning approaches.

The results, however, were sobering: all the models predicted all the outcomes poorly. On the

evaluation metric, where 1 indicates perfect accuracy and 0 corresponds to predictions that are no better than predicting simply the respective average in the training data, even the best models achieve scores of only 0.23 for material hardship, 0.19 for GPA, and down to 0.03 for lay-off.

6.1 The intended inference

Social scientists may want to interpret the results as suggesting that *the best scores achieved in the FFC represent the upper bound of what can be expected when predicting individual life events*. Under this interpretation, the FFC results would indicate that life events are inherently difficult to predict, rather than attributing the failure to the survey data or the available modelling techniques or resources.

6.2 Specifying validity conditions & providing relevant evidence

We briefly discuss the internal, external, and content validity in the context of the FFC. Given the policy implications, it is also important to consider consequential validity. In this case, however, the interpretation is not primarily concerned with evaluating prediction models, but with the prediction targets themselves: the extent to which individual life events are *in-principle* predictable. *Auxiliary validity* specifies conditions under which benchmark scores support such theoretical inferences.

Internal validity. The internal validity of the FFC results is high due to the careful study design. The final scores are based on data from the last survey wave, not publicly available at the time. This ensures independence between the models and the evaluation data.[18] The challenge data was collected as part of the Future Families and Child Wellbeing Study, a representative survey of non-marital births in large U.S. cities (Reichman et al., 2001). The main threat to internal validity lies in the limited sample size. Although 4, 242 families constitute a large cohort for a longitudinal survey, considering the costs of a 15-year study, the dataset is small relative to the 12, 942 features and additional splits into training, leaderboard, and evaluation data. This results in considerable uncertainty in identifying a "winning" model. When bootstrapping several test sets from the evaluation data, the top-ranked model changes in more than half the cases for GPA and nearly a quarter for children's grit (Salganik et al., 2020). Thus, the scores rankings partly reflect sampling artifacts rather than genuine performance differences between the models. Nevertheless, changes in the rankings are not a major obstacle to the intended inference since all models performed poorly.

External validity. The external validity of the FFC results is mixed: across a number of studies, most life outcomes show overall low predictability, while many can be predicted at least better than chance. Simple models using only a few variables often seem competitive. The FFC covers six different targets, indicating a certain robustness of the results across different life events. Despite slightly higher scores for *material hardship* and *GPA*, overall predictability is low for all targets. Results from other

[18] The reported scores may still be overestimated due to the finite samples. Score estimates that are corrected for this bias are qualitatively similar, even smaller but also less precise (Salganik et al., 2020).

studies also support this pattern. Dressel and Farid (2018) and Breen and Seltzer (2023) both report very limited predictability for the likelihood of *recidivism* and individual-level *mortality*, respectively. Zheng and Cheng (2025) find more mixed results: while *log total income* and *employment status* in the $40 - 50$ year age group were poorly predictable, *log hourly wages* reached a substantially higher score of 0.53 on the same metric used in the FFC. *Long-term unemployment* has also been found to be not very well predictable, with documented accuracies between $60 - 86\%$ for different statistical methods (Kuikka, 2024, Kern et al., 2021, Desiere and Struyven, 2020, Desiere et al., 2019). Building on extensive administrative labour and health records from Denmark, recently Savcisens et al. (2023) trained a transformer model to predict individual mortality. The model outperforms simple baseline models and has an accuracy of 78%, in line with the relatively weak predictive performance reported across life events.

Content validity. We observe strong content validity through FFC's close connection to social science theory. The data is drawn from the Fragile Families and Child Wellbeing Study, a well-documented survey study with extensive theoretical grounding (Reichman et al., 2001). The prediction targets cover diverse aspects and are established in the field. A central motivation of the survey underlying the FFC was to examine the role of unwed fathers and their impact on child wellbeing, a group that is often underrepresented in surveys because they are harder to identify (Reichman et al., 2001). This intended research focus on so-called "fragile families" in urban contexts introduces a systematic bias in its data coverage: on average, participating families are poorer and more marginalized than the general U.S. population. Predicting outcomes at age 15 additionally is particularly challenging, since adolescence is a volatile period of life. In contrast, outcomes in later adulthood are considered more predictable (Zheng and Cheng, 2025). The evaluation metric underlying the reported normalized out-of-sample R^2 scores is the mean squared error (MSE), which disproportionately penalizes large deviations. The choice reflects the research interest to identify families with "particularly unexpected outcomes" (Salganik et al., 2020).

Life events are not a uniform construct, but highly heterogeneous, and their predictability will vary with cultural and societal conditions. As such is the level of predictability informative as a measure of the social *rigidity* of a society (Zheng and Cheng, 2025, Blau and Duncan, 1967). Studying the predictability of life events, therefore, requires making explicit the social conditions under which predictability can be expected (Liou et al., 2023).

Consequential validity. The prediction of individual life events and the construction of datasets like the Fragile Families and Child Wellbeing Study are closely intertwined with policy interests. Predictions inform the design of targeted interventions and resource allocation or serve to evaluate existing programs (Perdomo et al., 2025, Zezulka and Genin, 2024, Coston et al., 2020). Despite the poor predictive performance, the results may nevertheless shape policy. In some settings, even modest improvements in accuracy may improve decision-making (Kleinberg et al., 2017), while in others, high accuracy is nevertheless insufficient (Shirali et al., 2024). While not directly applied in policy contexts,

the FFC's results carry strong implications for the feasibility and legitimacy of algorithmic governance (Wang et al., 2024).

Non-epistemic values play a constitutive role when defining and operationalising prediction targets and determining data collection procedures. Child wellbeing, central to the FFC, is a *thick* concept that intertwines descriptive and evaluative dimensions (Thoma, 2023, Blili-Hamelin and Hancox-Li, 2023, Alexandrova, 2018). A telling example is the study's original name "fragile families", chosen because children of unmarried parents are, on average, poorer and expected to spend less time with their fathers (Reichman et al., 2001, Garfinkel and McLanahan, 2003). The term has strong evaluative impetus and was later replaced by "future families."

Auxiliary validity. Auxiliary validity specifies the conditions that allow one to draw theoretical inferences about the target phenomenon. Following, we introduce two conditions that allow to infer limitations on the *predictability* of life events by ruling out relevant alternative explanations. The first concerns the best theoretically achievable performance given the predictor and target variables. Formally, this can be defined as the expected error of the *Bayes optimal model* (Hastie et al., 2009). Internal validity only implies that the benchmark scores are reliable estimates of the expected error of the models on the leaderboard, but these models might still perform far worse than the theoretically optimal. The second condition provides even stronger limitations on predictability by extending the evaluation to changes to the specified learning problem.

(i) The set of evaluated models is sufficiently diverse and the data size is large enough that the best model(s) approximate the Bayes optimal model.

(ii) The poor predictability is not a result of unmeasured predictors or the operationalisation of the target variable.

First, with 160 teams independently submitting models based on different approaches, the FFC takes a substantial step in this direction. The submissions cover the relevant modelling approaches available at the time of the challenge.[19] The relatively small data size, as discussed under internal validity, nevertheless puts strong limitations on the intended interpretation of the scores.

Second, the poor predictive performance may also result from inaccurate or incomplete data, a possibility investigated by Lundberg et al. (2024).[20] To identify limitations beyond the sample size that may contribute to the prediction errors, they interviewed 40 children and their families from the FFC, focusing primarily on cases where GPA predictions were inaccurate. From the interviews, they identified three important sources of errors: (a) imperfectly measured features, like coarse measures that hide meaningful variation or simple recording errors; (b) unmeasured features that often reflect constraints in time and budget. For example, one child's better-than-predicted GPA was explained

[19] Details on all submissions are provided in the Appendix of Salganik et al. (2020).

[20] Systematically investigating the errors made on a benchmark is a valuable approach to learn about the underlying phenomenon of interest and to improve predictions. Roß et al. (2023), for example, develop methods using hierarchical models in the context of medical imaging benchmarks.

by a strong social network outside her immediate family, a factor not covered by the data; and (c) unmeasurable features such as events that influence outcomes but occurred only after the survey was conducted.

6.3 Constraining the inference

The uniformly low scores of all models across the targets in the FFC cannot support the strong interpretation that life outcomes are *inherently* unpredictable. Evidence for the external validity is mixed, with some outcomes being more predictable than the targets in the FFC. The *auxiliary* conditions of data coverage and sample size are also important: using larger datasets and more informative predictors, some life outcomes may prove to be better predictable.

In light of the conceptual heterogeneity of life events and their theoretical meaning as the *rigidity* of a society, a weaker inference is better supported. The benchmark scores and the associated validity conditions support the inference that, despite the fact that many individual life events can be predicted better than chance, the overall predictive performance is likely to be low. This holds particularly for young and marginalized social groups. The challenge also provides important evidence that simple models using only a few variables are often competitive with more complex ones.

7 Towards a benchmarking epistemology

As demonstrated by the three case studies, predictive benchmarks are a powerful methodology for the empirical sciences. Benchmark scores, when considered together with conditions of validity, can support a wide range of inferences: measuring research progress in machine learning, informing deployment decisions, or assessing the predictability of life events. Drawing on the analogy with psychological testing, we argued that any such inference requires specifying and evaluating its conditions of validity. The four-step procedure, together with the case studies, offers a practical guide for interpreting benchmark scores. Developing the epistemology of benchmarking further opens several avenues for further inquiry, including (1) the assessment of model capabilities, (2) a social epistemology of benchmarks, and (3) examining the limits of benchmarking.

7.1 Assessing model capabilities

Although the four-step framework offers a general structure, the proposed validity conditions remain incomplete. The evaluation of large language models (LLMs), in particular, calls for additional conditions. With the development of LLMs, the role of benchmarks in machine learning has changed. Although these models are typically trained as next-token predictors, some researchers aim to attribute general *capabilities* to them and have begun to develop benchmarks to test these more complex constructs. For example, benchmarks have been designed to assess constructs such as morality (Jiang et al., 2021), theory of mind (Strachan et al., 2024), or legal reasoning (Masry et al., 2022).

The difficulty with the capability constructs is that they are complex, interconnected, and consequently hard to accurately measure (Strauss and Smith, 2009). Even the choice of the metric can have drastic consequences for the validity of measures (Schaeffer et al., 2023). Further, with the ability of moral reasoning, we typically also ascribe to some extent a theory of mind and a form of ability for legal reasoning. To ensure that a benchmark meaningfully measures the intended construct, one might verify that the measurements correlate with theoretically related constructs (*convergent validity*) and are uncorrelated with unrelated constructs (*discriminant validity*). Lacking a solid theoretical grounding and careful validation, inferences from benchmark scores risk *construct underrepresentation*, the failure to measure important aspects of the intended construct, or *construct irrelevance*, where the measures are influenced by factors unrelated to it (Strauss and Smith, 2009). Both could imply that a model appears to possess a capability according to one benchmark but not another, even though they aim to measure the same construct (Wiggins and Christopherson, 2019, Sühr et al., 2025). Such issues can only be avoided by rigorously evaluating the construct validity of inferences drawn from benchmark scores.

7.2 The social dimensions of benchmarking

In the present paper, we focused primarily on the epistemic role of predictive benchmarks, leaving aside their equally important social role. Who holds the power to establish benchmarks within a community? And how do benchmarks shape research practices and communities?

A growing body of work has begun to address these questions. Koch et al. (2021) show that research is increasingly concentrated on a small set of benchmark datasets—often repurposed from other tasks and produced by a few powerful institutions, raising concerns about biased data and task selection. Dehghani et al. (2021) argue that when a single benchmark dominates, it favours methods that are already well-aligned with the benchmark's idiosyncrasies; conversely, when multiple benchmarks coexist, researchers may simply choose those that best showcase their models–a phenomenon they call the "Benchmark Lottery." Taken together, these findings point to a fragility of benchmark-based evaluation, especially when model assessment is reduced to benchmarking alone. Hooker (1995) further cautions that framing scientific research as a competition fosters a race for marginal numerical improvements at the expense of more substantive qualitative progress.

In the context of our work, we observe another important social dimension. Machine learners and scientists typically employ benchmarks that they did not curate themselves. As a consequence, many of the validity conditions discussed above, including internal and content validity, lie outside their control. This underscores the need for *benchmark curators* to maintain close communication with the users of their benchmarks, ensuring that the benchmarks are designed to support relevant inferences and that the underlying design choices are clearly conveyed in the corresponding publications.

7.3 Construct validity can prevent "thinning the world"

In the beginning, we noted that predictive benchmarks were originally developed to shift the focus of evaluation away from broad claims about complex constructs such as artificial general intelligence and toward measuring predictive performance on concrete, narrowly defined learning problems. Ironically, with modern LLMs, researchers have returned to making claims about general capabilities—this time grounded in benchmark scores. This raises the question of the limits of benchmarks: what can, and what cannot, be meaningfully measured by them?

We maintain that predictive benchmarks are best suited for assessing model performance on specific, well-defined learning problems. The more complex and multifaceted the constructs we attempt to capture through benchmarks become, the greater the risk of construct underrepresentation or construct irrelevance (Raji et al., 2021). This aligns with Denton et al. (2021) and Orr and Kang (2024), who argue that emphasizing a small set of metrics within competitive benchmarking structures risks "thinning the world," neglecting aspects of a phenomenon that resist easy quantification. This danger becomes particularly acute when benchmarks become authoritative proxies for entire tasks, as ImageNet once did for image classification, gaining legitimacy through their widespread adoption within the research community (Orr and Crawford, 2024).

The multifaceted nature of validity conditions anchored in intended inferences offers a way to address this concern by systematically evaluating the interpretations of scores and the validity claims that support them. In other words, drawing reliable and well-supported inferences from benchmark scores requires *construct validity*.

Acknowledgements

We thank Konstantin Genin, Thomas Grote, Jan-Willem Romeijn, Markus Ahlers, Katharina Holzhey, Tom Sterkenburg, Moritz Hardt, Wolfgang Spohn, Bob Williamson, Luis Winckelmann, Tim Schreier, and Alex Braun for helpful discussions and feedback.

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. The authors acknowledge support for Timo Freiesleben from the Carl Zeiss Stiftung (Project: Certification and Foundations of Safe Machine Learning Systems in Healthcare) and thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Sebastian Zezulka.

References

Anna Alexandrova. Can the Science of Well-Being Be Objective? *The British Journal for the Philosophy of Science*, 69(2):421–445, 2018. doi: 10.1093/bjps/axw027.

Anna Alexandrova and Daniel M. Haybron. Is Construct Validation Valid? *Philosophy of Science*, 83 (5):1098–1109, 2016. doi: 10.1086/687941.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. American Educational Research Association, 2014.

Tom Beucler, Pierre Gentine, Janni Yuval, Ankitesh Gupta, Liran Peng, Jerry Lin, Sungduk Yu, Stephan Rasp, Fiaz Ahmed, Paul A O'Gorman, et al. Climate-invariant machine learning. *Science Advances*, 10(6), 2024.

Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet?, 2020. arXiv.

Peter M. Blau and Otis Dudley Duncan. *The American Occupational Structure*. John Wiley & Sons, New York, 1967.

Borhane Blili-Hamelin and Leif Hancox-Li. Making Intelligence: Ethical Values in IQ and ML Benchmarks. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 271–284. ACM, 2023. doi: 10.1145/3593013.3593996.

Avrim Blum and Moritz Hardt. The Ladder: A Reliable Leaderboard for Machine Learning Competitions. In *International Conference on Machine Learning*, volume 37, pages 1006–1014. PMLR, 2015.

Florian J. Boge. Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines*, 32(1):43–75, 2022.

James Bogen and James Woodward. Saving the Phenomena. *The Philosophical Review*, 97(3):303–352, 1988.

Rishi Bommasani, Percy Liang, and Tony Lee. Holistic Evaluation of Language Models. *Annals of the New York Academy of Sciences*, 1525(1):140–146, 2023. doi: 10.1111/nyas.15007.

Massimo Bonavita. On Some Limitations of Current Machine Learning Weather Prediction Models. *Geophysical Research Letters*, 51(12), 2024. doi: 10.1029/2023gl107377.

Denny Borsboom, Gideon J. Mellenbergh, and Jaap van Heerden. The Concept of Validity. *Psychological Review*, 111(4):1061–1071, 2004. doi: 10.1037/0033-295x.111.4.1061.

Zied Ben Bouallègue, Mariana C. A. Clare, Linus Magnusson, Estibaliz Gascón, Michael Maier-Gerber, Martin Janoušek, Mark Rodwell, Florian Pinault, Jesper S. Dramsch, Simon T. K. Lang, Baudouin Raoult, Florence Rabier, Matthieu Chevallier, Irina Sandu, Peter Dueben, Matthew Chantry, and Florian Pappenberger. The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning-Based Weather Forecasts in an Operational-Like Context. *Bulletin of the American Meteorological Society*, 105(6):E864–E883, 2024. doi: 10.1175/bams-d-23-0162.1.

Richard Bradley and Katie Steele. Making Climate Decisions. *Philosophy Compass*, 10(11):799–810, 2015. doi: 10.1111/phc3.12259.

Casey Breen and Nathan Seltzer. Demographic Perspectives on Predicting Individual-level Mortality. SocArXiv, 2023.

Noah D. Brenowitz, Yair Cohen, Jaideep Pathak, Ankur Mahesh, Boris Bonev, Thorsten Kurth, Dale R. Durran, Peter Harrington, and Michael S. Pritchard. A Practical Probabilistic Benchmark for AI Weather Models. *Geophysical Research Letters*, 52(7), 2025. doi: 10.1029/2024gl113656.

Oliver Buchholz and Thomas Grote. Predicting and explaining with machine learning models: Social science as a touchstone. *Studies in History and Philosophy of Science*, 102:60–69, 2023. doi: 10.1016/j.shpsa.2023.10.004.

Lucas F. F. Cardoso, Vitor C. A. Santos, Regiane S. Kawasaki Francês, Ricardo B. C. Prudêncio, and Ronnie C. O. Alves. *Decoding Machine Learning Benchmarks*, pages 412–425. Springer International Publishing, 2020. doi: 10.1007/978-3-030-61380-8_28.

Stephen Cave. The Problem with Intelligence: Its Value-Laden History and the Future of AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 29–35. ACM, 2020. doi: 10.1145/3375627.3375813.

Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 582–593, New York, NY, USA, 2020. Association for Computing Machinery. doi: 10.1145/3351095.3372851.

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

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The Benchmark Lottery, 2021. arXiv.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Li Kai, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009. doi: 10.1109/cvpr.2009.5206848.

Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, 8(2): 20539517211035955, 2021. doi: 10.1177/20539517211035955.

Sam Desiere and Ludo Struyven. Using Artificial Intelligence to classify Jobseekers: The Accuracy-Equity Trade-off. *Journal of Social Policy*, 50(2):367–385, 2020. doi: 10.1017/s0047279420000203.

Sam Desiere, Kristine Langenbucher, and Ludo Struyven. Statistical profiling in public employment services. Technical Report 224, OECD, 2019.

David Donoho. 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, 26(4): 745–766, 2017. doi: 10.1080/10618600.2017.1384734.

David Donoho. Data Science at the Singularity. *Harvard Data Science Review*, 6(1), 2024. doi: 10.1162/99608f92.b91339ef.

Florian Dorner, Tom Sühr, Samira Samadi, and Augustin Kelava. Do Personality Tests Generalize to Large Language Models? In *Socially Responsible Language Modelling Research*, Workshop at Advances in Neural Information Processing Systems, 2023.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), 2018. doi: 10.1126/sciadv.aao5580.

Peter D. Dueben, Martin G. Schultz, Matthew Chantry, David John Gagne, David Matthew Hall, and Amy McGovern. Challenges and Benchmark Datasets for Machine Learning in the Atmospheric Sciences: Definition, Status, and Outlook. *Artificial Intelligence for the Earth Systems*, 1(3), 2022. doi: 10.1175/aies-d-21-0002.1.

Eamon Duede. Deep Learning Opacity in Scientific Discovery. *Philosophy of Science*, 90(5):1089–1099, 2023. doi: 10.1017/psa.2023.8.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Computational Materials*, 6(1), 2020. doi: 10.1038/s41524-020-00406-3.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88 (2):303–338, 2009. doi: 10.1007/s11263-009-0275-4.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2014. doi: 10.1007/s11263-014-0733-5.

Timo Freiesleben and Thomas Grote. Beyond generalization: a theory of robustness in machine learning. *Synthese*, 202(4), 2023. doi: 10.1007/s11229-023-04334-9.

Timo Freiesleben, Gunnar König, Christoph Molnar, and Álvaro Tejero-Cantero. Scientific Inference with Interpretable Machine Learning: Analyzing Models to Learn About Real-World Phenomena. *Minds and Machines*, 34(3), 2024. doi: 10.1007/s11023-024-09691-z.

Roman Frigg, Leonard A. Smith, and David A. Stainforth. An assessment of the foundational assumptions in high-resolution climate projections: the case of UKCP09. *Synthese*, 192(12):3979–4008, 2015. doi: 10.1007/s11229-015-0739-8.

Irwin Garfinkel and Sara McLanahan. Unwed Parents in the US: Myths, Realities, and Policy Making. *Social Policy and Society*, 2(2):143–150, 2003. doi: 10.1017/s1474746403001209.

Robert Geirhos, Carlos R. M. Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Clark Glymour. What Went Wrong? Reflections on Science by Observation and The Bell Curve. *Philosophy of Science*, 65(1):1–32, 1998. doi: 10.1086/392624.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. ICLR 2015, 2014. arXiv.

Thomas Grote, Timo Freiesleben, and Philipp Berens. Foundation models in healthcare require rethinking reliability. *Nature Machine Intelligence*, 6(12):1421–1423, 2024a. doi: 10.1038/s42256-024-00924-5.

Thomas Grote, Konstantin Genin, and Emily Sullivan. Reliability in Machine Learning. *Philosophy Compass*, 19(5), 2024b. doi: 10.1111/phc3.12974.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 44123–44279. Curran Associates, Inc., 2023.

Moritz Hardt. The Emerging Science of Machine Learning Benchmarks. Online at `https://mlbenchmarks.org`, 2025. Manuscript.

Moritz Hardt and Benjamin Recht. *Patterns, Predictions, and Actions: Foundations of Machine Learning*. Princeton University Press, 2022.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=HJz6tiCqYm`.

Daniel A. Herrmann. PAC Learning and Occam's Razor: Probably Approximately Incorrect. *Philosophy of Science*, 87(4):685–703, 2020. doi: 10.1086/709786.

Moritz Herrmann, Philipp Probst, Roman Hornung, Vindi Jurinovic, and Anne-Laure Boulesteix. Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in Bioinformatics*, 22(3), 2021. doi: 10.1093/bib/bbaa167.

Moritz Herrmann, F. Julian D. Lange, Katharina Eggensperger, Giuseppe Casalicchio, Marcel Wever, Matthias Feurer, David Rügamer, Eyke Hüllermeier, Anne-Laure Boulesteix, and Bernd Bischl. Position: Why We Must Rethink Empirical Research in Machine Learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 18228–18247. PMLR, 2024.

Jake M. Hofman, Amit Sharma, and Duncan J. Watts. Prediction and explanation in social systems. *Science*, 355(6324):486–488, 2017. doi: 10.1126/science.aal3856.

J. N. Hooker. Testing heuristics: We have it all wrong. *Journal of Heuristics*, 1(1):33–42, 1995. doi: 10.1007/bf02430364.

Sergey Ioffe. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, 2019. doi: 10.1609/aaai.v33i01.3301590.

Julie Jebeile, Vincent Lam, and Tim Räz. Understanding climate change with statistical downscaling and machine learning. *Synthese*, 199(1–2):1877–1897, 2020. doi: 10.1007/s11229-020-02865-z.

Julie Jebeile, Vincent Lam, Mason Majszak, and Tim Räz. Machine learning and the quest for objectivity in climate model parameterization. *Climatic Change*, 176(8), 2023. doi: 10.1007/s10584-023-03532-1.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten

Sap, Regina Rini, and Yejin Choi. Can Machines Learn Morality? the Delphi Experiment, 2021. arXiv.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Michael T. Kane. Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50(1):1–73, 2013. doi: 10.1111/jedm.12000.

Suzanne Kawamleh. Can Machines Learn How Clouds Work? the Epistemic Implications of Machine Learning Methods in Climate Science. *Philosophy of Science*, 88(5):1008–1020, 2021. doi: 10.1086/714877.

Christoph Kern, Ruben L. Bach, Hannah Mautner, and Frauke Kreuter. Fairness in Algorithmic Profiling: A German Case Study, 2021. arXiv.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 2017. doi: 10.1093/qje/qjx032.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G. Foster. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Bernard J. Koch and David Peterson. From Protoscience to Epistemic Monoculture: How Benchmarking Set the Stage for the Deep Learning Revolution, 2024. arXiv.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do Better ImageNet Models Transfer Better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2661–2671, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Sanni Kuikka. The (Un)Predictability of Early (Un)Employment: A Machine Learning Approach. *Socius: Sociological Research for a Dynamic World*, 10, 2024. doi: 10.1177/23780231241286655.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677): 1416–1421, 2023. doi: 10.1126/science.adi2336.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Sabina Leonelli. What Counts as Scientific Data? A Relational Framework. *Philosophy of Science*, 82 (5):810–821, 2015. doi: 10.1086/684083.

Sabina Leonelli. Scientific Research and Big Data. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2020 edition, 2020.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. In *Advances in Neural Information Processing Systems: Track on Datasets and Benchmarks*, 2021.

Mark Liberman. Obituary: Fred Jelinek. *Computational Linguistics*, 36(4):595–599, 2010. doi: 10.1162/coli_a_00032.

Gloria Liou, Drew H. Bailey, Chayce R. Baldwin, Angela Lee Duckworth, and Louis Tay. Why Life Outcomes are Hard to Predict. PsyArXiv, 2023.

Ian Lundberg, Rachel Brown-Weinstock, Susan Clampet-Lundquist, Sarah Pachman, Timothy J. Nelson, Vicki Yang, Kathryn Edin, and Matthew J. Salganik. The origins of unpredictability in life outcome prediction tasks. *Proceedings of the National Academy of Sciences*, 121(24), 2024. doi: 10.1073/pnas.2322973121.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, page 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.177.

Amy McGovern, Ryan Lagerquist, David John Gagne, G. Eli Jergensen, Kimberly L. Elmore, Cameron R. Homeyer, and Travis Smith. Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. *Bulletin of the American Meteorological Society*, 100 (11):2175–2199, 2019. doi: 10.1175/bams-d-18-0195.1.

Samuel Messick. Standards of Validity and the Validity of Standards in Performance Asessment. *Educational Measurement: Issues and Practice*, 14(4):5–8, 1995. doi: 10.1111/j.1745-3992.1995.tb00881.x.

George A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995. doi: 10.1145/219717.219748.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The Effect of Natural Distribution Shift on Question Answering Models. In *International Conference on Machine Learning*, volume 119, pages 6905–6916. PMLR, 2020.

Raphaël Millière and Cameron Buckner. A Philosophical Introduction to Language Models - Part II: The Way Forward, 2024. arXiv.

Roger E. Millsap. *Statistical Approaches to Measurement Invariance*. Routledge, 2012. doi: 10.4324/9780203821961.

Stefano Moret. *Strategic energy planning under uncertainty*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2017.

Alexander Martin Mussgnug. The predictive reframing of machine learning applications: good predictions and bad measurements. *European Journal for Philosophy of Science*, 12(3), 2022. doi: 10.1007/s13194-022-00484-8.

Kien Nguyen, Clinton Fookes, Arun Ross, and Sridha Sridharan. Iris Recognition With Off-the-Shelf CNN Features: A Deep Learning Perspective. *IEEE Access*, 6:18848–18855, 2018. doi: 10.1109/access.2017.2784352.

Leonardo Olivetti and Gabriele Messori. Do data-driven models beat numerical models in forecasting weather extremes? a comparison of ifs hres, pangu-weather, and graphcast. *Geoscientific Model Development*, 17(21):7915–7962, 2024.

Will Orr and Kate Crawford. The social construction of datasets: On the practices, processes, and challenges of dataset creation for machine learning. *New Media & Society*, 26(9):4955–4972, 2024. doi: 10.1177/14614448241251797.

Will Orr and Edward B. Kang. AI as a Sport: On the Competitive Epistemologies of Benchmarking. In *Conference on Fairness, Accountability, and Transparency*, FAccT '24. ACM, 2024. doi: 10.1145/3630106.3659012.

Seymour A. Papert. The Summer Vision Project. AI Memo AIM-100, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, 1966. `http://hdl.handle.net/1721.1/6125`.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021. doi: 10.1016/j.patter.2021.100336.

Juan Carlos Perdomo, Tolani Britton, Moritz Hardt, and Rediet Abebe. Difficult Lessons on Social Prediction from Wisconsin Public Schools. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 2682–2704. ACM, 2025. doi: 10.1145/3715275.3732175.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2024. doi: 10.1038/s41586-024-08252-9.

Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. AI and the Everything in the Whole Wide World Benchmark. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11), 2020. doi: 10.1029/2020ms002203.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models. *Journal of Advances in Modeling Earth Systems*, 16(6), 2024. doi: 10.1029/2023ms004019.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 09–15 Jun 2019.

Nancy E. Reichman, Julien O. Teitler, Irwin Garfinkel, and Sara S. McLanahan. Fragile Families: sample and design. *Children and Youth Services Review*, 23(4–5):303–326, 2001. doi: 10.1016/s0190-7409(01)00141-4.

Roberta Rocca and Tal Yarkoni. Putting Psychology to the Test: Rethinking Model Evaluation Through Benchmarking and Prediction. *Advances in Methods and Practices in Psychological Science*, 4(3):2515245921102 68, 2021. doi: 10.1177/25152459211026864.

Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A Meta-Analysis of Overfitting in Machine Learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

Tobias Roß, Pierangela Bruno, Annika Reinke, Manuel Wiesenfarth, Lisa Koeppel, Peter M. Full, Bünyamin Pekdemir, Patrick Godau, Darya Trofimova, Fabian Isensee, Tim J. Adler, Thuy N. Tran, Sara Moccia, Francesco Calimeri, Beat P. Müller-Stich, Annette Kopp-Schneider, and Lena Maier-Hein. Beyond rankings: Learning (more) from algorithm validation. *Medical Image Analysis*, 86: 102765, 2023. doi: 10.1016/j.media.2023.102765.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Olawale Salaudeen and Moritz Hardt. ImageNot: A contrast with ImageNet preserves model rankings, 2024. arXiv.

Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation, 2025. arXiv.

Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, and Sara McLanahan. Introduction to the Special Collection on the Fragile Families Challenge. *Socius: Sociological Research for a Dynamic World*, 5:2378023119871580, 2019. doi: 10.1177/2378023119871580.

Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, Jennie E. Brand, Nicole Bohme Carnegie, Ryan James Compton, Debanjan Datta, Thomas Davidson, Anna Filippova, Connor Gilroy, Brian J. Goode, Eaman Jahani, Ridhi Kashyap, Antje Kirchner, Stephen McKay, Allison C. Morgan, Alex Pentland, Kivan Polimis, Louis Raes, Daniel E. Rigobon, Claudia V. Roberts, Diana M. Stanescu, Yoshihiko Suhara, Adaner Usmani, Erik H. Wang, Muna Adem, Abdulla Alhajri, Bedoor AlShebli, Redwane Amin, Ryan B. Amos, Lisa P. Argyle, Livia Baer-Bositis, Moritz Büchi, Bo-Ryehn Chung, William Eggert, Gregory Faletto, Zhilin Fan, Jeremy Freese, Tejomay Gadgil, Josh Gagné, Yue Gao, Andrew Halpern-Manners, Sonia P. Hashim, Sonia Hausen, Guanhua He, Kimberly Higuera, Bernie Hogan, Ilana M. Horwitz, Lisa M. Hummel, Naman Jain, Kun Jin, David Jurgens, Patrick Kaminski, Areg Karapetyan, E. H. Kim, Ben Leizman, Naijia Liu, Malte Möser, Andrew E. Mack, Mayank Mahajan, Noah Mandell, Helge Marahrens, Diana Mercado-Garcia, Viola Mocz, Katariina Mueller-Gastell, Ahmed Musse, Qiankun Niu, William Nowak, Hamidreza Omidvar, Andrew Or, Karen Ouyang, Katy M. Pinto, Ethan Porter, Kristin E. Porter, Crystal Qian, Tamkinat Rauf, Anahit Sargsyan, Thomas Schaffner, Landon Schnabel, Bryan Schonfeld, Ben Sender, Jonathan D. Tang, Emma Tsurkov, Austin van Loon, Onur Varol, Xiafei Wang, Zhi Wang, Julia Wang, Flora Wang, Samantha Weissman, Kirstie Whitaker, Maria K. Wolters, Wei Lee Woon, James Wu, Catherine Wu, Kengran Yang, Jingwen Yin, Bingyu Zhao, Chenyun Zhu, Jeanne Brooks-Gunn, Barbara E. Engelhardt, Moritz Hardt, Dean Knox, Karen Levy, Arvind Narayanan, Brandon M. Stewart,

Duncan J. Watts, and Sara McLanahan. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403, 2020. doi: 10.1073/pnas.1915006117.

Germans Savcisens, Tina Eliassi-Rad, Lars Kai Hansen, Laust Hvas Mortensen, Lau Lilleholt, Anna Rogers, Ingo Zettler, and Sune Lehmann. Using sequences of life-events to predict human lives. *Nature Computational Science*, 4(1):43–56, 2023. doi: 10.1038/s43588-023-00573-5.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are Emergent Abilities of Large Language Models a Mirage? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 55565–55581, 2023.

David Schlangen. Targeting the Benchmark: On Methodology in Current Natural Language Processing Research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-short.85.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? BioRxiv, 2018.

T. Selz and G. C. Craig. Can Artificial Intelligence-Based Weather Prediction Models Simulate the Butterfly Effect? *Geophysical Research Letters*, 50(20), 2023. doi: 10.1029/2023gl105747.

Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating Machine Accuracy on ImageNet. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8634–8644. PMLR, 2020.

Ali Shirali and Moritz Hardt. What Makes ImageNet Look Unlike LAION. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=IrBYuh9W3T`.

Ali Shirali, Rediet Abebe, and Moritz Hardt. Allocation Requires Prediction Only if Inequality Is Low. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 45114–45153. PMLR, 2024.

Galit Shmueli. To Explain or to Predict? *Statistical Science*, 25(3):289–310, 2010. URL `http://www.jstor.org/stable/41058949`.

Elizaveta Sivak, Paulina Pankowska, Adriënne Mendrik, Tom Emery, Javier Garcia-Bernardo, Seyit Höcük, Kasia Karpinska, Angelica Maineri, Joris Mulder, Malvina Nissim, and Gert Stulp. Combining the strengths of Dutch survey and register data in a data challenge to predict fertility (PreFer). *Journal of Computational Social Science*, 2024. doi: 10.1007/s42001-024-00275-6.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Tom F. Sterkenburg. Statistical Learning Theory and Occam's Razor: The Core Argument. *Minds and Machines*, 35(1), 2024. doi: 10.1007/s11023-024-09703-y.

Tom F. Sterkenburg and Peter D. Grünwald. The no-free-lunch theorems of supervised learning. *Synthese*, 199(3-4):9979–10015, 2021. doi: 10.1007/s11229-021-03233-1.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, 2024. doi: 10.1038/s41562-024-01882-z.

Milton E. Strauss and Gregory T. Smith. Construct Validity: Advances in Theory and Methodology. *Annual Review of Clinical Psychology*, 5(1):1–25, 2009. doi: 10.1146/annurev.clinpsy.032408.153639.

Emily Sullivan. Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*, 73(1):109–133, 2022. doi: 10.1093/bjps/axz035.

Tom Sühr, Florian E. Dorner, Olawale Salaudeen, Augustin Kelava, and Samira Samadi. Stop Evaluating AI with Human Tests, Develop Principled, AI-specific Tests instead, 2025. arXiv.

Johanna Thoma. Social Science, Policy and Democracy. *Philosophy & Public Affairs*, 2023. doi: 10.1111/papa.12250.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, Colorado Springs, CO, USA, 2011. IEEE. doi: 10.1109/cvpr.2011.5995347.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From ImageNet to Image Classification: Contextualizing Progress on Benchmarks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9625–9635. PMLR, 2020.

Mark D. Verhagen. A Pragmatist's Guide to Using Prediction in the Social Sciences. *Socius: Sociological Research for a Dynamic World*, 8, 2022. doi: 10.1177/23780231221081702.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. Against Predictive Optimization: On the Legitimacy of Decision-making Algorithms That Optimize Predictive Accuracy. *ACM Journal on Responsible Computing*, 1(1):1–45, 2024. doi: 10.1145/3636509.

Xiting Wang, Liming Jiang, Jose Hernandez-Orallo, David Stillwell, Luning Sun, Fang Luo, and Xing Xie. Evaluating General-Purpose AI with Psychometrics, 2023. arXiv.

David S. Watson. Conceptual challenges for interpretable machine learning. *Synthese*, 200(2), 2022. doi: 10.1007/s11229-022-03485-5.

Bradford J. Wiggins and Cody D. Christopherson. The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, 39 (4):202–217, 2019. doi: 10.1037/teo0000137.

Chhavi Yadav and Leon Bottou. Cold Case: The Lost MNIST Digits. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. Large Language Model Psychometrics: A Systematic Review of Evaluation, Validation, and Enhancement, 2025. arXiv.

Adrian K. Yee. Construct Validity in Automated Counterterrorism Analysis. *Philosophy of Science*, pages 1–18, 2024. doi: 10.1017/psa.2024.65.

Carlos Zednik and Hannes Boelsen. Scientific Exploration and Explainable Artificial Intelligence. *Minds and Machines*, 32(1):219–239, 2022. doi: 10.1007/s11023-021-09583-6.

Sebastian Zezulka and Konstantin Genin. From the Fair Distribution of Predictions to the Fair Distribution of Social Goods: Evaluating the Impact of Fair Machine Learning on Long-Term Unemployment. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT'24. Association for Computing Machinery, 2024. doi: https://doi.org/10.1145/3630106.3659020.

Guanhua Zhang and Moritz Hardt. Inherent Trade-Offs between Diversity and Stability in Multi-Task Benchmarks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 58984–59002. PMLR, 2024.

Kino Zhao. Measuring the Nonexistent: Validity before Measurement. *Philosophy of Science*, 90(2): 227–244, 2023. doi: 10.1017/psa.2023.3.

Kino Zhao. What Are Statistical Assumptions About? An Answer From Perspectivism. *Harvard Data Science Review*, 7(1), 2025. doi: 10.1162/99608f92.7fcd521d.

Haowen Zheng and Siwei Cheng. Social Rigidity Across and Within Generations: A Predictive Approach. *Sociological Methods & Research*, 2025. doi: 10.1177/00491241251347984.