AI and the Everything in the Whole Wide World Benchmark

Inioluwa Deborah Raji
Mozilla Foundation

Emily M. Bender
Department of Linguistics
University of Washington

Amandalynne Paullada
Department of Linguistics
University of Washington

Emily Denton
Google Research

Alex Hanna
Google Research

Abstract

There is a tendency across different subfields in AI to see value in a small collection of what we term "general" benchmarks, meant to operate as stand-ins or abstractions for a range of anointed common problems with a claim to indicate some general purpose performance. In this position paper, we discuss several reasons why these benchmarks are ultimately limited and do not in fact function as the broad measures of progress they are set up to be. We discuss how task formation for these benchmarks seems to happen independently of the intended and declared problem space, how such benchmarks tend to be de-contextualized and how performance on the benchmark is often inappropriately interpreted by the research community. As a result, these benchmarks consistently fall short of capturing meaningful abstractions of the declared motivations, present distorted data lenses of a specific worldview to be optimized for, and disguise key limitations in order to misrepresent the nature of actual "state of the art" (SOTA) performance of AI systems. We then present key considerations for the path towards more meaningful markers of progress in the field.

1 Introduction

In the 1974 Sesame Street children's storybook *Grover and the Everything in the Whole Wide World Museum* [Stiles and Wilcox, 1974], the Muppet monster Grover visits a museum claiming to showcase "everything in the whole wide world". Example objects representing certain categories fill each room. Several categories are arbitrary and subjective, including a showcase for "Things You Find On a Wall", and "The Long Thin Things You Can Write With Room". Some are oddly specific, such as "The Carrot Room", while others unhelpfully vague like "The Tall Hall". When he thinks that he has seen all that is there, Grover comes to a door that is labeled "Everything Else". He opens the door, only to find himself in the outside world.

As a children's story, the described situation is inherently silly, illogical and absurd. However, in this paper, we discuss how that faulty logic is inherent to recent trends of benchmarking in artificial intelligence (AI) — and specifically machine learning (ML) — evaluation, as some of the most popular benchmarks rely on the same false assumptions of the ridiculous "Everything in the Whole Wide World Museum" that Grover visits.

Broadly speaking, the goals of the development of artificial intelligence (AI) systems are highly under-specified. Some assess AI progress by asking scientific and philosophical questions on the notion of intelligence, thinking of AI as a method for modeling some broad cognitive function. Others evaluate AI performance on the success of a model's utility within some practical applications. In

NeurIPS 2020 Workshop: ML Retrospectives, Surveys Meta-analyses (ML-RSA)

both cases, we observe that a key set of influential benchmarks — which we describe as "general"[1] benchmarks — are developed and/or used in a manner that assumes a particular, closed sample of the world, categorized from one point of view, can capture everything in full generality and become more widely relevant.

In this work we begin by outlining the key characteristics of a "general" benchmark and provide examples of benchmark datasets that fall into this category. Despite their frequent presentation as general, universal, or exhaustive, we highlight that categories, implicit or explicit, in tasks or datasets, are inherently socially constructed and subjective; that these benchmarks, regardless of scale, represent a closed and finite scope of applications and cognitive tasks, in contradiction to their presentation as having a certain universal significance or relevance to open and infinite contexts; and that these limitations often go unacknowledged or misunderstood by those that come to make use of them. Using mainstream popular benchmarks to illustrate our point, we discuss how such assumptions about what these benchmarks represent leads to an inappropriate over-prioritization of achieving state-of-the-art (SOTA) results on a small set of less grounded "general" benchmark datasets, rather than more focused and explicitly defined problems. Inspired by similar lessons learnt in other disciplines, we conclude by challenging the ML community to be more precise and intentional about the nature of the experiments being run and the insights sought from the evaluation process.

2 Background

2.1 A Brief History of Benchmarking Practice in AI

In this paper we describe a *benchmark* as a particular combination of dataset or sets of datasets (at least test data, sometimes also training data), and a metric, conceptualized as representing one or more specific tasks or sets of abilities. The *task* is a particular specification of a problem, as represented in the dataset. A *metric* is a way to summarize system performance over some set or sets of tasks as a single number or score. Metrics can be a complex combination of scores on sub-tasks or an averaged amalgamation of other metrics, but can also represent a single concept such as accuracy across the entire benchmark. In either case, at the most basic level, the metric provides a means of counting success and failure at the level of individual system outputs (relative to a gold standard) and summarizing those counts over the full dataset. Such metrics allow for a stack-ranking of different models. Models obtaining the most favourable scores on the metrics for a benchmark are considered to be "state of the art" (SOTA) in terms of performance on the specified task.

Not all datasets achieve the status of benchmark. That status is conferred when the dataset — and the task it represents — are picked up by a community of researchers as a shared framework for the comparison of methods. While benchmark datasets, by design, encode limited representations of any given task, state-of-the-art performance is often taken to indicate progress towards a broader goal that is believed to be closely aligned with the benchmark.

Benchmarking as a practice represents the dominant paradigm of conducting research across multiple subfields of artificial intelligence and machine learning. The practice itself predates machine learning, but has parallel roots in several strands of measurement and assessment in sociotechnical systems. Even though natural language processing, computer vision, and reinforcement learning have converged into the current benchmarking practice, this framework can be traced back to previous iterations within their respective antecedent fields — for natural language processing that was information retrieval, speech transcription, and machine translation; for computer vision, it was signal processing, and pattern recognition; and for reinforcement learning, it was chess and game-playing. Here, we present a stylized history that's far from comprehensive, but attempts to provide the shape of the development of predecessors of current machine learning benchmarking practice.

Several linguists [e.g. Liberman, 2010, Church, 2018] have traced the influence of Fred Jelinek at IBM and Charles Payne at DARPA as crucial protagonists in the development of a common framework for the quantitative assessment of computational linguistic tasks, notably within speech recognition and

[1] Our use of quotations is meant to indicate that generality is a property being ascribed to these datasets (sometimes by dataset creators, other times by the communities leveraging the datasets) rather that a characteristic we as the authors believe them to hold.

machine translation. Jelinek had been a driver in developing what he had called the Common Task Framework (CTF). According to Donoho [2017], the Common Task Framework had three elements:

> (a) A publicly available training dataset involving, for each observation, a list of (possibly many) feature measurements, and a class label for that observation.
>
> (b) A set of enrolled competitors whose common task is to infer a class prediction rule from the training data.
>
> (c) A scoring referee, to which competitors can submit their prediction rule. The referee runs the prediction rule against a testing dataset which is sequestered behind a [wall]. The referee objectively and automatically reports the score (prediction accuracy) achieved by the submitted rule.
>
> [Donoho, 2017, p. 572]

Jelinek had played a role in pushing the CTF as a quantifiable mode of evaluation. Rather than experiments that were updated "twice a year", researchers could evaluate their experiments every hour [Church, 2018]. He was also keen on sharing of data outside of his organization. IBM had donated a copy of the Canadian Hansards, a corpus of government documents which had been translated into both English and French, to NIST as a neutral arbitrator, which could then be used for a machine translation task. Wayne, meanwhile, within DARPA, initiated a new speech recognition program within the agency, in an attempt to resuscitate the field of artificial intelligence research, after the publication of a withering report by John R. Pierce, an engineer and VP at Bell Labs. Pierce had written against AI more generally, and automated speech recognition in particular, deriding it as a waste of money and as impossible, given the contextual nature of speech [Pierce, 1969]. Therefore with DARPA's sponsorship and IBM's data, the CTF became a common mode of research in machine translation and speech recognition.

Shared and common tasks became a staple of other tasks within natural language processing. Belz and Kilgarriff [2006] have traced the practice of Shared Task Evaluation Campaigns (STECs) within information extraction to the information extraction from military messages around naval sightings and engagements, published by the US government. Grishman and Sundheim [1996] note that the first of these Message Understanding Conferences (MUCs) did not focus on optimization around particular metrics, but the task settled on a shared set of metrics in its second iteration. Later STECs within NLP coalesced around machine translation (NIST-MT) and word sense disambiguation (SENSEVAL).

Meanwhile, modern reinforcement learning algorithms are often evaluated by playing games like Atari, Starcraft, Dota2, and Go. However, one of the first games to widely capture the imagination of AI researchers was chess. [Ensmenger, 2012, p. 6] identifies that AI scientists of the 1950s and 1960s chose computer chess as a "relatively simple system that nevertheless could be used to explore larger, more complex phenomena." On its face, chess became a natural choice for a metric of general intelligence, given that Elo rating system — a system for calculating the relative rankings of players against each other — had been used for years and was seen as a tried-and-true metric for evaluation. Elo could then be adapted to evaluate non-human, machine players. However, Ensmenger also discusses how the game was chosen less for its applicability for actual artificial intelligence development, and more as a function of its popularity amongst engineers as a proxy for general intelligence, and its association with Cold War national competition between the US and Russia. In this sense, chess was meant to stand in for the whole of artificial intelligence based on these associations; as discussed below, it stood as the "model organism" or *drosophilia*, of AI research.

Computer vision researchers engaged in shared task evaluation from at least as early as the 1990s, a time period which includes early NIST pattern recognition competitions in subfields such as facial recognition. The Facial Recognition Technology (FERET) program for automated facial recognition, inaugurated in 1993 [Phillips et al., 2000], was one of the first to do so. They report that, prior to the publication of the FERET database, "[o]nly a few of these algorithms reported results on images utilizing a common database let alone met the desirable goal of being evaluated on a standard testing protocol that included separate training and testing sets. As a consequence, there was no method to make informed comparisons among various algorithms" [Phillips et al., 2000, p. 2].

In this work we examine a particular style of benchmark dataset — and the accompanying practices of use — that has gained in popularity in recent years. These benchmarks, which we characterize in further detail in Section 2.3, embed notions of generality, both in their presentation by the creators and the manner in which they are adopted and used by the machine learning community. Before outlining

the characteristics of these "general" benchmarks, we first review the notion of general-purpose objective for AI models.

2.2 Striving for generality

The field of artificial intelligence and machine learning is driven, in large part, by ambitions to develop AI systems that are general-purpose in nature. For example, the development of Artificial General Intelligence (AGI) represents the pinnacle of success for a subset of the research community. While the precise definition of AGI remains underspecified and almost intentionally vague, researchers often use this term to refer to the development of systems that are flexible in nature, demonstrating intelligent abilities on a wide range of tasks and in a wide range of settings, hoping to mirror the adaptive cognitive abilities that humans are percieved to possess [Shevlin et al., 2019, Voss, 2007]. The development of AGI is sometimes contrasted with "narrow AI": systems that excel on a narrow set of tasks or domains [Pennachin and Goertzel, 2007].

However, even researchers indifferent to the goals of building AGI systems seem equally preoccupied by a desire for models to possess some level of general purpose performance. Within specific sub-fields and more pragmatic settings, researchers will often strive for the development of general-purpose systems that capture a breadth of functionality or knowledge within that domain. For example, natural language processing communities frequently position "general-purpose language understanding" as an ultimate goal of the field [Wang et al., 2019a]. Such general-purpose systems would have the capacity to perform multiple language tasks, mimicking the flexibility of human language capabilities. In computer vision, notions of generality often stem from the breadth of categories of concepts a system has the capacity to recognize coupled with the ability to recognize such concepts in a variety of scenes and contexts [Liu et al., 2020].

The development of general-purpose feature embeddings has also become a central research focus within natural language processing and computer vision communities. Here, the goal is the development of representations — through either unsupervised or supervised means — that can generalize (with minimal fine-tuning) to a wide range of other tasks they were not specifically developed for [e.g. Du et al., 2020, Huh et al., 2016].

2.3 "General" AI Benchmarks

Having reviewed benchmarking methods and the notion of general-purpose AI, we now turn our attention to the focus of this paper: benchmark datasets being presented as suitable for the measurement of general-purpose capabilities, such as general-purpose language understanding or general-purpose visual understanding. These "general" benchmarks operate as markers of progress towards long-term research objectives oriented around the development of generalized understanding or intelligence tasks. Moreover, these benchmarks play a critical role orienting research agendas and structuring incentives within major subfields in artificial intelligence. We believe this phenomenon is driven, in large part, by the presentation of the tasks as general in nature and by community excitement over leaderboards, rather than by scientifically rigorous and critical examinations of the extent to which the tasks in the benchmarks meaningfully capture the general cognitive abilities they are framed as exemplifying. This contrasts with more practically-oriented and tightly scoped AI tasks, such as machine translation (MT), where the required validation is whether the benchmark accurately reflects the practical task being asked of the computer in its real-world context.

Crucially, even when the creators of such benchmarks do not explicitly purport to be establishing benchmarks for "general intelligence", community practices and overall hype have dramatized what it means for a model to perform well on these benchmarks. As the inspiration for the setup of these benchmarks is often explicitly linked to the general nature of human knowledge, the ethos of these datasets often extend beyond a reasonable scope of interpretation and influence, especially given the limitations inherent in their construction.

In this paper, we focus our attention on two datasets, ImageNet [Deng et al., 2009] and GLUE/SuperGLUE [Wang et al., 2019a,b], as canonical examples of "general" benchmarks for computer vision and natural language understanding, respectively. We look at both how the benchmark creators present their datasets and how the community has taken them up, while noting that our critique is primarily focused on the culture of benchmarking that has grown up around this type

of dataset and how they are used in evaluating machine learning systems and supporting claims of generality.

2.4 ImageNet

The ImageNet dataset [Deng et al., 2009], which represents the inter-related tasks of object classification and object detection, was developed to advance research on artificial visual understanding. In the years since its development, ImageNet has become one of the most influential benchmark datasets within the field of computer vision, and arguably machine learning more broadly. Beginning in 2012, SOTA performance on the yearly ImageNet challenge was sufficient to launch a researcher's career and establish dominance of their methods.

We characterize ImageNet as a "general" benchmark due to several related observations. First, we observe that ImageNet is presented, and widely regarded, as a nearly all-encompassing and comprehensive representation of the visual world. For example, at the time of development, the creators describe ImageNet as representing "the most comprehensive and diverse coverage of the image world" [Deng et al., 2009] and, retrospectively, Li has described the project as an "attempt to map the entire world of objects" [Gershgorn, 2017]. With general-purpose object recognition framed as the ability to recognize a sizable breadth of objects in a manner that rivals human capabilities [Liu et al., 2020], we observe ImageNet's sheer size — both in terms of number of categories and number of images per category — further informs its perception as representing a general formulation of object recognition. Second, we observe community consensus that the task of large-scale object recognition, as formalized in the ImageNet dataset, represents a meaningful milestone towards longer term goals of artificial visual intelligence. Indeed, Li has explicitly characterized large scale object recognition as the "north star" of computer vision — a scientific quest that would define and guide the field towards the ultimate goal of artificial visual intelligence [Fei-Fei, 2019], with ImageNet operating as the canonical instantiation.

We believe these characteristics have contributed to the current research dynamics surrounding ImageNet whereby SOTA performance on ImageNet is frequently regarded as as marker of progress towards general-purpose image understanding and artificial general intelligence more broadly.

2.5 Glue and SuperGLUE

The creators of GLUE (General Language Understanding Evaluation) [Wang et al., 2019a] and SuperGLUE [Wang et al., 2019b] present these resources as "evaluation framework[s] for research towards general-purpose language understanding technologies" [Wang et al., 2019b, p.1], noting that unlike human language understanding most computer natural language understanding (NLU) systems are task or domain specific [Wang et al., 2019a]. When a human knows a language, they can use that knowledge across any task that involves that language. Thus a benchmark that tests whether linguistic knowledge acquired through training on one task can be applied to other tasks in principle tests for a specific and potentially well-defined kind of generalizability.[2] The GLUE and SuperGLUE benchmarks combine linguistic competence (the ability to model a linguistic system) and general commonsense and world reasoning as if they were equivalently scoped problems: "This dataset is designed to highlight common challenges, such as the use of world knowledge and logical operators, that we expect models must handle to robustly solve the tasks." [Wang et al., 2019a, p.1] Thus a benchmark designed to test for generalizability across different language understanding tasks comes to subsume not only the task of building up linguistic competence (e.g. logical operators) in the language in question (English) but also the ability to acquire and deploy world knowledge.

Furthermore, taken together, performance on these assessments can be interpreted as evidence of overall cognition or understanding. This logic elevates benchmarks like GLUE or ImageNet to become definitions of the essential common tasks to justify the performance of any given model. As a result, often the claims that are justified through these benchmark datasets extend far beyond the tasks they were designed for, and reach beyond even the initial ambitions for development. For example, OpenAI founder Ilya Sutskever uses ImageNet and GLUE benchmark performance to back up the claim that "short-term AGI (artificial general intelligence) is possible", also flagging advancement of

[2]Furthermore, the GLUE and SuperGLUE benchmarks include a subtask based on a diagnostic dataset which probes for specific linguistic structures of interest, to test whether systems show evidence of the linguistic competence thought to be required to handle the tasks.

reinforcement learning model performance on game benchmarks like Go, and the video game Dota2 [Sutskever, 2018].

3 Limits of the General Benchmark

Our goal in this paper is to illuminate the over-reliance on the Common Task Framework (CTF) for machine learning as it evolved into what we understand today to be these general benchmarks. We observe a culture of benchmarking within machine learning that takes up certain tasks and situates them as if they were canonical problems, appropriately representative of the real world, that if only computers could 'solve' then computers would have achieved (at least some aspect of) general intelligence. But, we argue, the imagined artifact of the general benchmark does not actually exist — real data is designed, subjective and limited in ways that necessitate a different framing from that of any claim to all-important or general knowledge.

In fact, we propose that setting up any datasets in this way is ultimately dangerous and deceptive, resulting in misguidance on task design and focus, underreporting of the many biases and subjective interpretations inherent in the data and enabling potential misuse as evidence for machine learning performance in inappropriate scenarios. The idea of general knowledge or intelligence is itself nebulously defined and questionable — the attempt to embed such ideas in a data-defined benchmark is, in actuality, inappropriate.

In this section, we walk through the details of key limitations of general benchmarks — *limited task design*, *de-contextualized data and performance reporting* as well as *inappropriate community use*. For each of these limitations, we break down the shortcoming and discuss the cited evidence and reasoning for our observations.

3.1 Limited Task Design

In the Grover story shared earlier, there is one question that is likely to immediately arise for readers: what is the purpose of an "Everything in the Whole Wide World" museum? A natural history museum, for instance, may be built to provide inspiration for wildlife conservation or an aviation museum constructed to perhaps spark academic interest in aerospace. What would be the point of a museum containing every object in the world?

Even in the situation where that objective of generality is coherent, the rooms Grover explores are clearly inconsistent and unsystematic in their collective design, leaving no room for a genuinely strategic attempt to achieve even the stated claim of containing the entirety of the world's objects. In the same way, many general benchmarks seem equally vague in purpose and unsystematic in their pursuit of an inadequately justified objective, at times – though not always – divorced from any specific and meaningful practical utility.

Schlangen [2020] defines a *task* as a mapping from an input space to an output space, defined both *intensionally* (via a description of the task) and *extensionally* (via a particular dataset, i.e. pairs of inputs and outputs matching the description). In machine learning, the tendency is for the latter scenario, where the benchmark task is defined by a dataset, often collected and curated with such a role in mind.

Task design and validity is questionable for several of these benchmarks, motivating researchers down an inappropriate or underspecified path. The task formation for these benchmarks seems to happen independently of the intended and declared problem space. More generally, dataset development practices within machine learning have been characterized by a "laissez-faire" attitude, rather than a more thoughtful, intentional, and curatorial approach [Jo and Gebru, 2020]. There is little justification provided for the exact nature of the tasks presented in these benchmarks, with many dataset design decisions arising from convenience. The following are the issues we've identified with respect to the task design for these "general" benchmarks.

3.1.1 Arbitrarily selected tasks and collections

If the so-called general benchmarks were legitimate tests of progress towards general artificial cognitive abilities, we would expect them to be chosen with reference to specific theories of the

cognitive abilities they model. Instead, what we observe looks more like samples of convenience: tasks and collections of tasks built out of what was handy.

The UC Irvine Machine Learning Repository (referred to as UCI; [Dua and Graff, 2017]), was one of the first foci of shared tasks in machine learning development. UCI was created in 1987 as a response to many calls for machine learning to have a centralized repository for data in machine learning [Radin, 2017]; it contains datasets which pertain to collection of individual subtasks, with benchmarks evolving to focus on the Iris, Adult, Wine, and Breast Cancer datasets. Kiri Wagstaff voices concern for a poor logic in the selection of this combination of subtasks, noting that none of tasks included in this collection represent any reasonable proxy or abstraction of real world problems even scientists in that field would care about [Wagstaff, 2012] — for example, the Iris dataset does not represent a task of any interest to botanists. Similarly, the subtasks of the GLUE benchmark were sourced with an informal survey of colleagues, not any comprehensive search [Wang et al., 2019a].

The challenge is that this haphazard selection of tasks directly results in attention towards unimportant and pseudo-random topics, implying that there is an extrapolation of performance on these tasks to more general situations, regardless of the fact that, unlike with software benchmarking [Lewis and Crews, 1985], the axiomatic subtasks were not actively designed as abstractions of any meaningful general function or sub-function, and are often not systematic in nature. Chollet [2019], for instance, speaks in detail about how current machine learning benchmarks fail to test for the logics of abstract thinking, though this is an expressed requirement for general intelligence. Similarly, it has long been shown that there are missing elements within the suite of NLP benchmark tasks [e.g. Palmer and Finn, 1990]. For instance, pragmatics is essential to dialogue and, although integration into chatbots or interactions with virtual assistants is often cited as a key incentive for the development of NLP models, there remains no inclusion of a suitable test for an abstraction of this linguistic skill included in any of the "general purpose language benchmarks".[3] It is clear that these benchmark tasks, from the overall objectives to the set of subtasks and categories, are most often defined by the nature of the data that happens to be the most available and easy to manipulate or distribute — rather than being determined by any intentional task design or considered, conscious focus on an articulated objective for dataset design.

3.1.2 Critical misunderstandings of domain knowledge and application problem space

If the so-called general benchmarks were legitimate tests of progress towards general artificial cognitive abilities, we would expect them to be grounded in the literature that looks at those cognitive abilities in humans or other animals. What we observe, however, is a culture in which not only is the input from domain experts not consistently sought out, it is often disregarded even when offered.

This is most evident in natural language processing, where the selected subtasks for "general" popular benchmarks such as GLUE have been repeatedly questioned by domain experts from the linguistics community about the limitations of the articulated test sets to capture any meaningful linguistic skill or an appropriate level of natural language understanding. For instance, several researchers have questioned the utility of the Question Answering benchmark SQuAD, not just in terms of practical relevance to any real world operations [Zellers et al., 2020], but also in terms of its ability to act as a reliable stand in as a task for "general language understanding". Bender and Koller mention the tendency of machine learning researchers to confuse advancement on certain benchmarks with any tangible evidence of capturing the meaning in language [Bender and Koller, 2020], arguing that benchmarks need to be constructed with care if they are to show evidence of "understanding" as opposed to merely the ability to manipulate linguistic form sufficiently to pass the test. Several experiments on question answering datasets including SQuAD reveal that with cosmetic changes (i.e. shuffled words in the question to make sentences nonsensical), the models optimized for performance on these benchmarks will answer questions correctly using pattern matching of form alone — indicating a lack of ability to truly "understand" the text at all [Rajpurkar et al., 2018, Sugawara et al., 2020]. Similar misconceptions have also been revealed around the limits of the conceptual design of other GLUE subtasks, such as those related to natural language inference [Schlegel et al., 2020].

[3] Qiu et al. [2020], in their summary of pre-trained language models, identify pragmatics as among the things a good representation of language should capture, yet none of the evaluations they report evaluate language models for this capability, nor is it represented in GLUE or SuperGLUE.

Similarly, there is a clear disconnect between the generalization of learned knowledge and the general language understanding task the GLUE benchmark has come to be viewed as being relevant to. Language understanding relies not only on linguistic competence but also world knowledge, common sense reasoning, and the ability to model the interlocutor's state of mind [Reddy, 1979, Clark, 1996], none of which are incorporated as considerations in GLUE. In addition, several researchers have raised the need to establish effective physical and social grounding as part of the process of moving towards robust and effective natural language understanding, warning against text-only learning as a limited approach [Bisk et al., 2020, Zellers et al., 2020]. However, despite repeated attempts for reform, the field has been mainly resistant to that direction, and most benchmarks have yet to evolve beyond purely text-based datasets.

We note here that in the ideal scenario, if the goal is to create benchmarks that can be used as legitimate tests for progress towards generalized language understanding, the benchmarks should be designed in the first instance based on what is currently known about how language understanding works in humans. To a certain extent, GLUE and SuperGLUE meet this criterion: they include a subtask based on a diagnostic dataset designed around specific linguistic phenomena that are both well-studied and at least plausibly cover a significant portion of a closed domain.[4] Similarly, the CommitmentBank [De Marneffe et al., 2019] task, included in SuperGLUE, explores a key component of natural language meaning, namely the ways in which lexical choices (such as *know* v. *believe* in *Kim knows/believes the sky is orange*) impact whether or not the speaker of a sentence is publicly committed to the truth of propositions embedded within it. The primary problem we see is in the framing and especially uptake of GLUE and SuperGLUE as testing general understanding of "language" (or even of English specifically), including so-called common sense reasoning and world knowledge, and the reluctance on the part of the community of users of the benchmark to revise their understanding of what the benchmark is measuring in light of critiques from domain experts.

Finally, we'd like to note that consulted experts can and should also include the affected population. A good example of this is the case of the VizWiz workshop, working with members of the blind community to co-design and validate benchmark datasets for the development of the automated captioning tools that impact them.[5]

3.1.3 Summary

In this section, we describe how Benchmarks claiming to represent general purpose objectives often make unstrategic attempts to address their motivating questions. Even if a general benchmark was something that could be designed and deployed, what we would expect to see around the scientific practice of developing and using such benchmarks falls short of the actual practice we observe in the field. Our goal here is not to explain how to better create and deploy general benchmarks, but rather to show how this gap between actual and ideal practice illuminates the infeasibility of true general benchmarks.

3.2 De-contextualized Data and Performance Reporting

One of the most remarkable things about the "Everything in the Whole Wide World" museum is how subjectively defined the rooms are. Nothing about the museum is neutrally determined. Rooms are arbitrarily categorized, contingent on the personal taste of whoever built the museum to define how specific or vague a category is, and what is included or excluded from a particular room. For example, perhaps the museum owner loves carrots - creating a separate room for this vegetable, ie."The Carrot Room", as well as an "All the Vegetables in the Whole Wide World Besides Carrots" room. The inherent bias of selection becomes most clear when Grover sees himself as being a great fit for being in a room meant for "Things That Are Cute and Furry"!

Also, in the end, Grover himself wonders, "I have seen many things but have I seen everything?".This is what leads him to the exit of the closed system and into the outside world, indicating inherent limits to what can had been captured, in that case, within museum walls.

[4] In addition to items such as lexical entailment and factivity (lexical semantics), coordination scope and active/passive paraphrases (predicate-argument structure) and negation and conditionals (logic), however, this diagnostic dataset also includes the completely open-ended and ill-defined categories of "common sense" and "world knowledge".

[5] Details on the VizWiz workshop can be found here: https://vizwiz.org/workshops/2020-workshop/

General benchmarks operate as similarly closed and inherently subjective, localized constructions.

In this section, we explore one of the features that can lead researchers to mistakenly construe a benchmark as "general", namely the decontextualization of its component tasks and datasets. Just because a task is unpinned to a specific context, does not make it general. Furthermore, simply scaling the dataset to make it very large does not make this benchmark open-ended, neutral or accurate. A large closed problem is not an open problem — it is still of a limited scope. No dataset is neutral and there are inherent limits to what a benchmark can tell us. If anything, the claim to generality will often act as cover, allowing those developing the benchmarks to escape the responsibility of reporting details of these limitations. Part of the challenge of addressing this lack of context is proper documentation for these datasets, which is often underdeveloped [Gebru et al., 2020, Bender and Friedman, 2018].

3.2.1 Limited Scope

Just like the rooms in the museum that Grover visited, these "general" benchmarks do not represent an open context, but capture a fraction of the world, within a scope that is limited. As a dataset increases in size, it is more and more likely the dataset is to include a broader range of artifacts from the world. However, it will be impossible to include everything in the world for fairly practical reasons — the logistics of storage, the fact that the actual world is not static, the nature of our data sourcing and the simple fact that not every relevant aspect of the world is possible to even capture in computable media.

For instance, ImageNet [Deng et al., 2009], a general purpose vision benchmark, was amongst the largest image datasets of its time — including over 14 million images and over 10,000 classes. A dataset at that scale is difficult to overfit to, just given the range and variety of images it includes [Roelofs et al., 2019, Recht et al., 2019]. For the same reason, such large scale benchmarks are also quite effective as indicators of performance on other benchmarks, simply due to the high probability of representational overlap inevitable with a dataset of that size [bar]. However, even large datasets like ImageNet are ultimately closed systems, unable to cover media representations from the future. In fact, any images unlikely to appear in its source material of digital images posted online at a certain time and in a certain cultural context will be omitted, and remain vulnerable to any natural or synthetic image perturbations in the data. As Torralba and Efros demonstrate, images from the same class but different datasets are often distinguishable and embody a very specific style of capturing some segment of the real world, unable to ever get to full coverage of every potentially relevant edge case [Torralba and Efros, 2011]. Specifically, a well critiqued limitation of ImageNet is that the objects tend to be centered within the images, which does not reflect how "natural" images actually appear [bar].

Similarly, showing that a system can generalize from one or more tasks to a small set of others is not the same thing as showing that it can generalize to any language understanding task, as GLUE's design implies.

3.2.2 Benchmark Subjectivity

All datasets come with an embedded perspective — there is no neutral or universal dataset, just as there is no "view from nowhere" [Haraway, 1988, Stitzlein, 2004, Gebru, 2020]. If ImageNet was developed using the results of Hindi online image queries, rather than English ones, the result would be a dataset that looks drastically different, embodying completely new representations of concepts [de Vries et al., 2019].

Similarly, GLUE and SuperGLUE target one specific language (English), not "language" in the abstract. Perhaps a more apt expansion of the acronym would have been "General Linguistic Understanding of English." Building benchmarks with a very specific subset of American English text in natural language processing, and the built-in bias of annotation artifacts [Gururangan et al., 2018] results in inherently subjective outcomes [Waseem et al., 2020].[6] When the claim of the benchmark is to assess some general performance, any assumed objectivity makes it difficult to

[6]Waseem et al. [2020] are discussing NLP datasets and modeling in general, not GLUE, SuperGLUE or other similar benchmarks specifically. However, their general points apply: any given dataset represents the embodied viewpoint of its authors and annotators, and furthermore, datasets constructed without attention to whose viewpoints are being represented will likely over-represent hegemonic ones.

recognize biases and articulate them just as an assumed universality makes it difficult to set boundaries on the limits of scope.

Denying the existence of unacknowledged context does not make it disappear. In fact, this distorted data lens is often not limited in an arbitrary way, but limited in a way that hurts certain groups of people — those without the assumed power to define the data. For example, minority gender and race identities are unintentionally underrepresented in mainstream face datasets [Merler et al., 2019], and tagged with racial or ethnic slurs, even on large general use datasets such as MIT Tiny Images or ImageNet [Crawford and Paglen, 2019, Gehl et al., 2017, Prabhu and Birhane, 2020].

Making uncritical use of any dataset in evaluation is irresponsible, and to present a political and value-laden dataset as a completely neutral scientific artifact is dishonest. Often, for general benchmarks, such politics will remain undiscussed and hidden under the claim to broad relevance.

3.2.3 Summary

In this section, we have considered ways the decontextualization of benchmarks contributes to the mistaken perception that they can be "general": if we mistake sheer size for representativeness it is easy to be misled into seeing some large benchmark as including comprehensive or complete coverage of a real word situation when that is likely not be the case. Likewise, when labels taken from a specific, usually hegemonic, point of view are presented not as the perceptions of specific individuals who did the labeling but rather as "objective" ground truth, it is easy to believe that training a machine to reproduce those labels constitutes creating a system that can itself perceive ground truth. Finally, when the specifics of the metrics used are packed into simple headline numbers, it is correspondingly difficult to keep a sufficiently precise picture of what, exactly, is being measured.

3.3 Inappropriate Community Use

Grover came into the museum, excited and inspired by its claim to contain every object in the "Whole Wide World" in much the same way that we get excited by these benchmarks, encompassing sizable chunks of the "World Wide Web". Although this claim to generality is exactly what attracted Grover, once he enters, it is immediately clear that the rooms only really showcase a small handful of certain objects, and nothing close to the actual advertised scope. His focus is thus redirected to whatever is selected to be highlighted in the museum. Despite his initial interest being motivated by the pursuit of generality, it is only really the subset of included objects that actually captures his attention.

Similarly, performance on the benchmark is often inappropriately interpreted by the research community. When individual dataset creators and communities overstate the generality of these benchmark datasets, they elevate them to the status of a target the entire field should be aiming for. As a result, researchers may fall into the trap of uncritically chasing algorithmic improvement as measured by these datasets, losing sight of moments of performance mismatches with the real world. At times, this benchmark focus escalates to the point that the creators' initial intention gets lost. The authors of GLUE, for example, designed the benchmark to serve the role of a "playground" and "library", similar to UCI. It was meant to be a step towards enabling data accessibility for young scholars looking to test out ideas. However, the way it has been adopted by the machine learning community has almost compromised that intent, morphing its status within the community into that of a general benchmark.

The following are the issues with respect to the inappropriate community use for these "general" benchmarks.

3.3.1 Limits of competitive testing

The most clear social influence of general benchmarks has been their role as the ultimate version of the "Common Task" framework, pulling together many researchers of various interests to engage with a single benchmark to make rapid progress for the field. For the most part, this has been incredibly productive, leading to significant breakthroughs in the field — for instance, AlexNet and thus the subsequent deep learning movement we're experiencing in machine learning would not have been possible without ImageNet [Alom et al., 2018].

That being said, there are clear limits to the hyper-focus on benchmark performance we see. Chasing "state of the art" (SOTA) performance is a very peculiar way of doing science — one that focuses on

empirical and incremental work rather than hypothesis-based scientific inquiry [Hooker, 1995]. In 1995, at an ICML workshop, Lorenza Saitta criticized benchmark chasing, saying that "it allowed researchers to publish dull papers that proposed small variations of existing supervised learning algorithms and reported their small-but-significant incremental performance improvements in comparison studies" [Radin, 2017, p. 61]. Thomas and Uminsky go so far as to present metric chasing as an ethical issue that compromises the integrity of the field, stating that "overemphasizing metrics leads to manipulation, gaming, a myopic focus on short-term goals, and other unexpected negative consequences" [Thomas and Uminsky, 2020, p. 1]. More concretely, under this paradigm, researchers often only try to explain the logic of a "winning" model post-hoc — if they attempt to explain the performance at all. As Hooker states, "competitive testing tells us which algorithm is (better) but not why" [Hooker, 1995]. Given the other inherently limiting factors to algorithmic performance — such as hardware availability [Hooker, 2020] — it can be incredibly difficult to parse out the exact factors leading to a winning performance [see also Fokkens et al., 2013]. As a result, we often get to the SOTA model almost by chance, with at times little understanding of how this model won the contest and what happened for it to win, leaving few lessons for the community to leverage in its attempt to move forward.

3.3.2 Justification for practical out of context or unsafe applications

Most remarkably, general benchmarks also get mentioned in marketing copy for commercial machine learning products, with performance on the benchmark presented as evidence of real-world technical achievement. This context is when the significance of benchmarks is most severely distorted, when performance on benchmarks is not just the tool for algorithmic selection, but actually presented as some reliable marker of expected model achievement in deployment. Kiri Wagstaff points out some inherent flaws with this logic: Aside from the representational inconsistencies between benchmarks and real world scenarios, the aggregate performance evaluation format for many general benchmarks are inherently flawed. She notes that "80% accuracy on Iris classification might be sufficient for the botany world, but to classify as poisonous or edible a mushroom you intend to ingest, perhaps 99% (or higher) accuracy is required. The assumption of cross-domain comparability is a mirage created by the application of metrics that have the same range, but not the same meaning. Suites of experiments are often summarized by the average accuracy across all datasets. This tells us nothing at all useful about generalization or impact, since the meaning of an x% improvement may be very different for different data sets (or classes)." [Wagstaff, 2012]

Similarly, ImageNet claims to be "comprehensive" but classes are unbalanced, with a long tail of under-represented classes or some which are more homogenous than others. Recent studies reveal how that impacts performance on certain groups over others, though this disproportionate performance is often hidden. For instance, facial recognition models are less performant on darker skinned women than lighter skinned men, though this was obscured through aggregate performance metrics on biased benchmarks, disguising the models' consistent failure for a certain demographic [Buolamwini and Gebru, 2018, Merler et al., 2019]. In the same way, certain pruning methods have been revealed to decrease performance on certain underrepresented classes over others [Hooker et al., 2019, 2020]. This also applies to many collection datasets, such as GLUE, where poor performance on one subtask can be disguised by great performance on every other subtask to give the false impression of a performant model.

3.3.3 Redirection of focus for the field

General benchmarks are incredibly influential. Even today, the U.S. government invests millions of dollars in benchmark development projects as a move to incentivize academic interest in a particular problem domain that is relevant to their challenges. For example, the National Institute of Standards and Technology (NIST) invested in excess of $6.5 million in order to stir up research participation on the challenge of facial recognition, a technology of interest to the sponsoring agency — the Department of Defense Counterdrug Technology Development Program Office.[7]

However, outside of influencing community research focus, benchmarks are also greatly influential with respect to the development of other benchmarks. First created in 1990, WordNet [Miller, 1995] was a text database of semantic relationships between English words. Organized around synsets — sets of words that share at least one sense (definition) — this dataset has been an integral element

[7] https://www.nist.gov/programs-projects/face-recognition-technology-feret

of natural language processing, most notably inspiring peer databases but also direct derivative datasets such as WordSim-353 [Finkelstein et al., 2001] and now over 200 language translations.[8] Interestingly, WordNet also begat ImageNet, which is built from Image search query results from WordNet terms [Deng et al., 2009].

Benchmarks also influence the nature of the dominant algorithmic approaches attempted. For example, in the 1960s, chess caused the AI community to hyper-focus on deep-tree searching and the minimax algorithms, which were most effective on improving game performance. Both of these methods came to dominate the algorithmic development of this time, resulting in the neglect of alternate problems and approaches [Ensmenger, 2012]. Dotan and Milli discuss these benchmarks as value laden and definitive of the algorithmic approaches that end up defining the field [Dotan and Milli, 2020], with several scholars agreeing to the important role such benchmarks play in determining the nature of algorithmic development and in many cases setting the research agenda for decades to come.

Given that these benchmarks will often endure for a very long time — despite any known limitations, expiration and biases — this level of influence can easily become problematic, resulting in long term consequences for the discipline that are hard to trace. WordNet's focus on synonyms fed into the current paradigm of word similarity based heuristics in the assessment of meaning and the nature of word embeddings, as well as anchored an entire field to text-only based datasets presenting language form as a mechanism to learn meaning. In the same vein, the large scale of WordNet is what enables the large size of ImageNet. ImageNet's scale was a necessary feature to enable the appropriate testing necessary for the introduction of neural nets, however this also disguised problematic dictionary words inherited from WordNet, including several offensive terms and slurs that would later appear as labels in ImageNet [Crawford and Paglen, 2019]. And finally, the scale of the impact is one that can be almost impossible to roll back — even when a widely used benchmark is found to be inappropriate or otherwise harmful, and removed by the initial creator of the dataset, derivative versions and copies of the dataset will almost inevitably live on [Peng, 2020]. Given this pattern of long-lasting influence of benchmarks, the problematic characteristics of "general" benchmarks risk extending far beyond the lifetime of the benchmarks themselves.

3.4 Summary

In this section, we have critically examined the process of elevating given datasets to the status of "general" benchmark from several different angles. We argue that any task is inherently limited, such that treating a given task as a means of measuring general abilities is inherently misleading. Furthermore, we identify certain facets of the tasks elevated to "general" benchmark status which raise the risk that the research community will mistake them in this way: large size, which is erroneously conflated with representativeness; hegemonic viewpoint, which is erroneously conflated with objectivity; and simple metrics, which obscure the details of how system performance relates to the underlying task being measured. Finally, we explore the ways in which the practices around general benchmarking (competitive testing, use in marketing, and the influence accorded to the benchmarks) exacerbate the harms this approach can do, to the field and to the public at large.

4 Lessons from Other Fields

4.1 In Search of the Modern AI *Drosophila*

In discussing the concept of "model organisms" such as the bacterium *E. coli*, the fruit fly *Drosophila melanogaster* and the house mouse, *Mus musculus*, experimental biologists Levy and Currie note the difference between the framework of theoretical statistical or mechanistic modeling and the practice of executing experiments on such "model organisms". Model organisms are chosen for as a matter of convenience, ease of use and breeding. They also become institutions in and of themselves, part of a lab infrastructure which researchers use and reuse. Levy and Currie note that although both forms of models serve as a stand in for experimentation in the real world, "theoretical modeling is grounded in explicit and known analogies between model and target. By contrast, inferences from model organisms are empirical extrapolations." [Levy and Currie, 2015, p. 327]

[8] http://globalwordnet.org/resources/wordnets-in-the-world/

Biologists do experiments on fruit flies not because they believe fruit fly organs to be a reasonable abstraction of human organs, but because it is the available test bed for harmless exploration — for example, an ideal and relatively harmless milieu to investigate the dynamics of the interactions between chemical reactants in live tissue. The goal is not to cure a disease for the fly, believing that to be the cure for the same disease in a human — the goal is to explore safely and test hypotheses for the details of a broader plan for intervention, and aiming for performance on a more specific and representative target or evaluation [Kohler, 1994].

Even in the 1960s when chess dominated as AI's "general" benchmark, dataset development was a matter of convenience. As with Ensmenger's discussion of chess above, he notes that, "[A]s a technology embedded in systems of practice and networks of exchange" the selection of chess was a combination of more social rather than theoretical reasoning, since chess allowed researchers to "tap into a long tradition of popular chess culture, with its corresponding technical and theoretical literature, international networks of enthusiasts and competitions, and well-developed protocols for documenting, sharing, and analyzing data." [Ensmenger, 2012, p. 7]

Ensmenger also names a similar distinction between the role of a "model organism" and a target, with preference for the former. However, he notes that "unlike drosophila, however, and despite its apparent productivity as an experimental technology, computer chess ultimately produced little in terms of fundamental theoretical insights" despite "long-term implications for the research agenda of a discipline." [*Ibid.* p. 7] He argues that "the brute-force computational techniques that proved most suitable for winning computer chess tournaments distracted researchers from more generalizable and theoretically productive avenues of AI research" [p. 7], that " Deep Blue's brute force approach to computer chess – along with its narrowly specialized 'Kasporov Killer' techniques – was too single-minded to suggest any meaningful general intelligence [Aleksander, 2001, Ekbia, 2008]" [p. 22] and that "the way in which a minimax-based machine plays chess is not at all like the way a human plays chess" [p. 23].

AI researcher John McCarthy goes so far as to comment, "Computer chess has developed much as genetics might have if the geneticists had concentrated their efforts starting in 1910 on breeding racing Drosophila. We would have some science, but mainly we would have very fast fruit flies" [McCarthy, 1997, p. 1518]. In other words, as Ensmenger puts it, "computers got much better at chess, but increasingly no one much cared." [p. 7]

In a word, we've been down this road before. AI researchers focusing on chess-playing algorithms substituted the game for the whole, the minimax-based algorithms for general intelligence. In the same way, "general" benchmarks should not be thought of as abstractions of our world. They are an inappropriate analogy — limited in scope, perspective and scale.

There is no dataset that will be able to capture the full complexity of the details of existence, in the same way that there can be no museum to contain the full catalog of everything in the whole wide world. Rather than engaging in the futile attempt of attempting to swallow the whole world in a benchmark, we need to adjust to this limitation and begin to explore what characteristics of a dataset or challenge problem would be most useful for tinkering.

Unlike in the realm of early software development, where the subroutines used to test the speed of candidate computer machines were defined in full by the test engineer, machine learning is working with data. Data is messy and inherently subjective. Data is dynamic and unpredictable, finite and compact. Data is human. Therefore, we cannot just uncritically follow in the footsteps of prior computing practices when it comes to benchmarking, acknowledging that our own benchmarks do not mean the same thing. Our datasets and collections are not analogies, but rather a playing ground; not definitive as markers of progress but rather an opportunity to try something out and understand. In such a case, the significance of "state of the art" loses its appeal — it means little to win for the sake of it in this case, as there is no guaranteed practical reward, since an analogy to the complete real world situation is revealed to require ridiculous abstractions the data cannot uphold. More important than winning is discovery; more important than the leaderboard is testing for a hypothesis. Reframing how we report performance on benchmarks, and how we develop and communicate about current as well as future "general" benchmarks, will be essential to the machine learning community being able to take this crucial step forward.

4.2 Gross Domestic Product & the Vague Measurement of "Progress"

We also see an analogy between general AI benchmarks and gross domestic product (GDP) in economics. GDP was originally conceptualized as a way to get information about a specific aspect of a nation's economy (the capacity it had for being taxed, or for producing arms and other goods during war time) and was then standardized so that it could be compared across countries (and across time) [Ivković, 2016]. From there, it was reappropriated as a measurement of "progress" without sufficient grounding in either what "progress" means in economic terms nor study of the connection between GDP and some (suitably specified) notion of "progress" [*Ibid.*]. Ivković [2016] notes that even as it was initially adopted as a measure of "progress", there were critiques from Nobel laureates and others, identifying its anchor to Western, capitalistic interpretations of economic performance and pointing with reluctance for the potential impossibility of genuinely defining such a measure.

The discussion around GDP also reflects the tendency of AI's "general" benchmarks to be influential without precision. If the goal of research in AI is to build systems which manifest intelligence, we must be stringent in our tests for what counts as evidence of intelligence. If the goal of research in AI is to build practical systems that work well and equitably in their deployed contexts, we must stringent in our tests of alignment between system evaluation and deployment context. To accept particular benchmarks as our means of measuring progress towards these goals, without fully contextualizing them within a larger understanding of the object of study or on-the-ground deployment contexts, entails the same narrowing of vision as does accepting GDP as a measurement of "progress" in society. Furthermore, the details of what gets overlooked in these two narrowings of vision resonate with each other strongly: both taking GDP to be the primary measure of social progress and taking increases in SOTA on purportedly general, decontextualized benchmarks to be the primary measure of progress towards AI takes the focus away from people and their concerns, including such things as privacy and other fundamental rights, environmental considerations, and the way the product development summarized in that number can harm individuals.

5 Alternative Roles for Benchmarking and Alternative Evaluation Methods

In preceding sections, we have explored the ways in which "general" benchmarks fail to serve as effective measures of progress in machine learning, at best, and misdirect effort, at worst. Now we ask: What can be done instead? It is not a question of fixing or improving the general benchmarks, as general benchmarks will always suffer from Grover's problem: a benchmark can measure performance on a specific task, and perhaps generalization from one domain to another, but it can't represent everything in the whole wide world. Instead, we argue that benchmarks should be developed and understood as specific, rather than general, and viewed as just one tool among many for understanding how systems do and don't work. In this section, we consider how benchmarks can be used appropriately and what other techniques they can be paired with.

The purpose of benchmarks can be understood, in the best light, as a means of stimulating and measuring progress in a field. When thinking about alternatives to benchmarks for these purposes, it is worthwhile to step back and consider what we mean by "progress in the field". Looking back at scientific achievements, it is easy to conceive of progress as linear in the sense that if it were optimally efficient, it would follow a certain set of steps from some starting position to the current state of the art (understood broadly). Applying that same viewpoint to future progress, it is easy to fall into the assumption that there is some specific future goal that is the inevitable end point of the next set of scientific progress, and the best way to hasten our arrival at that end point is to find the quickest path from here to there.

We would encourage, however, a different mode of thinking, one in which science, engineering, and scholarship more broadly are about exploration and the construction of ever larger bodies of knowledge that are strengthened by interconnections. This conceptualization of the goals and process of scholarship highlights the value of bringing in a diverse range of methods for understanding the effectiveness and limitations of different technological approaches. But even if we conceive of scholarship as progress to a predetermined goal, we argue that we will get there faster if our methods for carrying out and reporting on empirical studies take a broader view of experiments than as simply measurements of progress towards (a proxy for) that goal.

Figure 1: Photo of a benchmark in Edinburgh, by Jeremy Atherton, used without modification according to license CC-BY-SA-2.5

Fortunately, we do not need to start from scratch. There are many existing traditions of scholarship, even within fields under the AI umbrella, that provide methodologies for understanding what is working and why and thereby shed more light on the objects of study (which we take to include both systems and tasks). We take as inspiration the metaphor embedded in the term "benchmark" itself. This term has as its etymology a mark that was added to buildings to indicate the position of a surveyor's bench (see Fig. 1), itself a tool for creating a level surface on which to put a leveling rod, used in the process of surveying. This etymological source contrasts with the use of benchmarks for charting the state of the art: surveyors are not in the business of measuring the furthest anyone has gone along some particular trail but rather in understanding the shape of the landscape and how it changes over time [Kahmen and Faig, 1988].[9]

If we return to a notion of benchmarks as way points that help us understand the landscape, what is their role in scientific discourse and what else can/should they be paired with? We propose that their role should be relegated to (i) providing facile points of comparison between differing systems and (ii) serving as sanity checks that a given system is performing better than a naïve baseline, i.e. that it has captured and is able to apply some regularity in the data. Other methodologies that can be deployed to fill in our picture of the landscape include testsuites, audits and adversarial testing; system output analysis; ablation testing; analysis of model properties; and evaluations of interpretability. In the remainder of this section, we will briefly consider each of these in turn.

5.1 Testsuites, audits and adversarial testing

Typical benchmark evaluation datasets are sampled from some larger dataset (e.g. via a train/test split) such that the frequency distribution of test item *types* in the test data is influenced by their distribution in that underlying dataset. In contrast, testsuites and audits specifically design their test sets to map out some space of test item types and evaluate systems in terms of the extent to which they can handle them.[10] The testsuite-based approach has a long history in NLP, with notable early publications including Lehmann et al. [1996] and recent work such as Ribeiro et al. [2020]. The audit methodology, exemplified by Buolamwini and Gebru [2018], creates test sets balanced for sensitive categories so as to be able to test for differential performance across those categories. Finally, adversarial testing seeks to explore the edges of a system's competence by finding minimally contrasting pairs of examples where the system being tested succeeds on one member of the pair and fails on the other [Ettinger et al., 2017].[11] Importantly, these evaluation approaches are designed around diagnosing particular areas of system failure: the point of the tests isn't to show which systems can "solve" them but to understand which aspects of the problem space remain challenging.

5.2 System Output Analysis

System output analysis is another way to explore the system output in detail. This can take the form of error analysis, disaggregated analysis, and counterfactual analysis.

[9] We can take further inspiration from the field of surveying, as we think about how the measurements we take relate to the terrain we wish to understand: "A surveyor is not only charged with providing results derived from [their] measurements, but also has to give an indication of the quality and reliability of these. This requires a clear understanding of the functional and stochastic relationships between measured quantities and derived results, as well as a solid understanding of the external factors that influence the measurements." [Kahmen and Faig, 1988, p.1]

[10] A notable exception to the trend of benchmarks to not include testsuites is GLUE, which includes among its component tests a testsuite mapping out various linguistic constructions in English [Wang et al., 2019a].

[11] Adversarial testing also includes work like Niven and Kao [2019] that discovers what kind of spurious cues systems are leveraging to effectively 'cheat' on a particular benchmark and creates alternate versions of the testsets that neutralize those cues.

Error analysis Error analysis involves the detailed analysis of system errors, either mechanically or by manual inspection of system inputs. Mechanical analysis includes such simple methodologies as confusion matrices in labeling tasks. A confusion matrix compares gold labels to system output labels and provides a summary of which categories are most reliably labeled and which are most frequently confused for each other. Other mechanical analyses include looking at errors by easily automatically measurable properties of system input such as sentence length or presence or number of out of vocabulary items.

More detailed error analysis digs into the specific system inputs to look for patterns that can't necessarily be measured automatically and might not be modeled in any way by the system being evaluated. For example, error analysis of a sentiment analysis system might find that it is frequently tripped up by sarcasm or in a simpler case by sentences with phenomena such as coordination or subordinate clauses. Error analysis of a machine translation (MT) system might find that it frequently fails on examples with negation [Wetzel and Bond, 2012, Fancellu and Webber, 2015, Hossain et al., 2020]. As can be seen in the MT example, error analysis can turn up important problems that don't have a large effect on the metric. Metrics for MT systems (including but not limited to BLEU) are also not good at measuring the impact of negation errors [Hossain et al., 2020].

Disaggregated analysis Disaggregated analysis can reveal disparate patterns of performance that may not be visible through aggregate metrics alone. This method has been leveraged within the ground-breaking audit of facial analysis systems, performed by Buolamwini and Gebru [2018], that evaluated performance across unitary and intersectional subgroups defined by gender and Fitzpatrick skin type. This analysis revealed significant disparities in model performance — with darker skinned female subjects experiencing the highest error rates — that was not visible through examination of aggregate performance metrics alone. The method of disaggregated analysis has since been adopted by a myriad of auditing and evaluation works [e.g. Raji and Buolamwini, 2019] and integrated into frameworks of standardized model reporting [Mitchell et al., 2019]. Following these works, we encourage researchers to report performance metrics on socially salient slices of their dataset, in addition to the full test set.

Counterfactual analysis Counterfactual analysis is another technique of model evaluation and assessment that has gained in popularity in recent years. At a high level, these methods evaluate how a model's output changes in response to a counterfactual change in the input. This method has been leveraged for fairness-informed analysis of natural language processing systems by comparing model performance on paired inputs that differ only in a reference to a sensitive identity group [Garg et al., 2019, Hutchinson et al., 2020]. While both counterfactual analysis and disaggregated analysis have been leveraged to disparities in model performance for different sensitive groups, counterfactual analysis can additionally provide insight into causal mechanisms underlying particular patterns in performance. Counterfactual analysis can also be leveraged to assess model robustness to small distribution shifts [Christensen and Connault, 2019].

The results of system output analysis tend to be rich and detailed and not amenable for quick cross-system comparison. But this is a feature, in our view, and not a bug: the goal, after all, is not anoint one system the winner (until some new system claims that spot), but rather to understand how aspects of system design map onto different aspects of the problem space so as to inform the next iteration of system development.

5.3 Ablation testing

Another well-established technique for understanding system performance is ablation testing. Ablation testing involves isolating the contributions of different system components by removing them, one by one, and evaluating the modified system. In statistical NLP prior to deep learning approaches, ablation testing was commonly applied to different feature sets to explore the extent to which systems were using information captured by different aspects of input representation. Ablation testing can also be performed on subsets of training data to explore the effect of e.g. in-domain v. out-of-domain training data or on components of system architecture, to the extent that these can be removed without completely disabling the system. Heinzerling [2019], in a discussion inspired by Niven and Kao [2019] suggests various data ablations that can be applied to test data as well, to investigate what cues a system might be relying on.

5.4 Analysis of model properties

System performance on test items (in aggregate as in test sets in standard benchmarks or in detail as in error analysis or testsuites or audits) is only one facet of systems to consider, especially when trying to gauge which approaches are most feasible for practical applications. Other dimensions include energy consumption (both for development and testing) [Strubell et al., 2019, Henderson et al., 2020, Schwartz et al., 2019, Ethayarajh and Jurafsky, 2020]; memory and compute requirements, which may be more or less constrained depending on the deployment context [Ethayarajh and Jurafsky, 2020, Dodge et al., 2019]; and stability in the face of perturbations to the training data [Sculley et al., 2018]. This latter is especially important for systems that need to be continually retrained in order to handle changing types of input data, such as named entity recognition system that needs to keep up with different political figures of note.

6 Conclusion

The days when the AI research community could plausibly claim that its evaluation practices were only of import to research community-internal concerns are long gone. If the measures and conceptualization of "progress" we adopt are faulty, the harm extends far beyond just the research program. Given the recent wide-spread deployment of technology built on AI research, there are now larger consequences for the decisions we make about our evaluation practices. National standards bodies are currently engaged in creating standards around AI. For example, the US National Institute of Standards and Technology prepared a planning document entitled *U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools* in response to US Executive Order 13859. Regarding metrics, this document states, "Develop metrics and data sets to assess reliability, robustness and other trustworthy attributes of AI systems, focusing on approaches that are readily understandable, available, and can be put on a path to standardization." [National Institute of Standards and Technology, 2019, p.5] When governmental standards bodies look to AI research for input on how to evaluate AI technology, it is imperative that we provide them with clear-eyed, realistic proposals for assessment.

As a result, we need to be careful which datasets we set as the vanguards of progress in this field. Open-world, universal, neutral problems don't exist for such data-driven technologies, and the notion of generality, at least in this moment, seems misguided by our current method of benchmarking. "General" benchmarks do little to appropriately test for the cognitive and practical functions we wish to evaluate, they obscure contextual details of performance in a way that can become harmful and they are overly enthusiastically interpreted by researchers in a way that legitimately hurts the field.

The effective development of benchmarks is critical to progress in machine learning, but what makes a benchmark effective is not the strength of its arbitrary and false claim to "generality" but its effectiveness in how it helps us understand as researchers how certain systems work — and how they don't. Benchmarking is not about winning a contest but more about surveying a landscape — the more we can re-frame, contextualize and appropriately scope these datasets, the more useful they will become as an informative dimension to more impactful algorithmic development. At minimum, given the alternative roles and interpretations for evaluation we could explore in this space, it is essential that we move quickly beyond the narrow-yet-totalizing lens of the "Everything in the Whole Wide World" benchmark we remain anchored to.

References

Igor Aleksander. *How to Build a Mind: Toward Machines with Imagination*. Columbia University Press, 2001.

Md Zahangir Alom, Tarek M Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S Awwal, and Vijayan K Asari. The history began from AlexNet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*, 2018.

Anja Belz and Adam Kilgarriff. Shared-task evaluations in HLT: Lessons for NLG. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 133–135, 2006.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://www.aclweb.org/anthology/Q18-1041.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL https://www.aclweb.org/anthology/2020.acl-main.463.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-main.703.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Timothy Christensen and Benjamin Connault. Counterfactual sensitivity and robustness, 2019. https://arxiv.org/abs/1904.00989.

Kenneth Ward Church. Emerging trends: A tribute to Charles Wayne. *Natural Language Engineering*, 24(1):155–160, January 2018. ISSN 1351-3249, 1469-8110. doi: 10.1017/S1351324917000389.

Herbert H. Clark. *Using Language*. Cambridge University Press, Cambridge, 1996.

Kate Crawford and Trevor Paglen. Excavating AI: The politics of images in machine learning training sets. *Excavating AI*, 2019.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. The CommitmentBank: Investigating projection in naturally occurring discourse. In *Proceedings of Sinn und Bedeutung*, volume 23, pages 107–124, 2019.

Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1224. URL https://www.aclweb.org/anthology/D19-1224.

David Donoho. 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, 26(4): 745–766, October 2017. ISSN 1061-8600. doi: 10.1080/10618600.2017.1384734.

Ravit Dotan and Smitha Milli. Value-laden disciplinary shifts in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 294, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3373157. URL https://doi.org/10.1145/3351095.3373157.

Jingfei Du, Myle Ott, Haoran Li, Xing Zhou, and Veselin Stoyanov. General purpose text embeddings from pre-trained language models for scalable inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3018–3030, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.findings-emnlp.271.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. http://archive.ics.uci.edu/ml.

Hamid Reza Ekbia. *Artificial dreams: The quest for non-biological intelligence*, volume 200. Cambridge University Press Cambridge, 2008.

Nathan Ensmenger. Is chess the drosophila of artificial intelligence? A social history of an algorithm. *Social Studies of Science*, 42(1):5–30, 2012. doi: 10.1177/0306312711424596. URL https://doi.org/10.1177/0306312711424596. PMID: 22530382.

Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboard design. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-main.393.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. Towards linguistically generalizable NLP systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5401. URL https://www.aclweb.org/anthology/W17-5401.

Federico Fancellu and Bonnie Webber. Translating negation: A manual error analysis. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 2–11, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-1301. URL https://www.aclweb.org/anthology/W15-1301.

Li Fei-Fei. Fei-Fei Li - Where Did ImageNet Come From?, 2019. https://www.youtube.com/watch?v=Z7naK1uq1F8.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414, 2001.

Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P13-1166.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 219–226, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3317950. URL https://doi.org/10.1145/3306618.3317950.

Timnit Gebru. Race and gender. In Markus D. Dubber, Frank Pasquale, and Sunit Das, editors, *The Oxford Handbook of Ethics of AI*. Oxford University Press, Oxford, 2020.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets, 2020. https://arxiv.org/abs/1803.09010.

Robert W Gehl, Lucas Moyer-Horner, and Sara K Yeo. Training computers to see internet pornography: Gender and sexual discrimination in computer vision science. *Television & New Media*, 18 (6):529–547, 2017.

Dave Gershgorn. The data that transformed AI research——and possibly the world. *Quartz*, 2017. https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/.

Ralph Grishman and Beth Sundheim. Message understanding conference-6: A brief history. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, pages 466–471, USA, 1996. Association for Computational Linguistics. doi: 10.3115/992628.992709.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://www.aclweb.org/anthology/N18-2017.

Donna Haraway. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies*, 14(3):575–599, 1988.

Benjamin Heinzerling. NLP's Clever Hans moment has arrived. Blog post, available at https://bheinzerling.github.io/post/clever-hans/, accessed July 25, 2019, 2019.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning, 2020. https://arxiv.org/abs/2002.05651.

John N Hooker. Testing heuristics: We have it all wrong. *Journal of heuristics*, 1(1):33–42, 1995.

Sara Hooker. The hardware lottery, 2020. https://arxiv.org/abs/2009.06489.

Sara Hooker, Aaron Courville, Yann Dauphin, and Andrea Frome. Selective brain damage: Measuring the disparate impact of model pruning. 2019. https://arxiv.org/abs/1911.05248.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models, 2020. https://arxiv.org/abs/2010.03058.

Md Mosharaf Hossain, Antonios Anastasopoulos, Eduardo Blanco, and Alexis Palmer. It's not a non-issue: Negation as a source of error in machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3869–3885, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.findings-emnlp.345.

Mi-Young Huh, Pulkit Agrawal, and A. Alexei Efros. What makes imagenet good for transfer learning? *NIPS Workshop on Large Scale Computer Vision Systems*, 2016.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.487. URL https://www.aclweb.org/anthology/2020.acl-main.487.

Anita Frajman Ivković. Limitations of the GDP as a measure of progress and well-being. *Ekonomski Vjesnik/Econviews-Review of Contemporary Business, Entrepreneurship and Economic Issues*, 29 (1):257–272, 2016.

Eun Seo Jo and Timnit Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 306–316, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372829. URL https://doi.org/10.1145/3351095.3372829.

Heribert Kahmen and Wolfgang Faig. *Surveying*. Walter de Gruyter, Berlin, 1988.

Robert E Kohler. *Lords of the fly: Drosophila genetics and the experimental life*. University of Chicago Press, 1994.

Sabine Lehmann, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Hervé Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold. TSNLP — Test Suites for Natural Language Processing. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 711 – 716, Kopenhagen, Denmark, 1996.

Arnon Levy and Adrian Currie. Model organisms are not (theoretical) models. *The British Journal for the Philosophy of Science*, 66(2):327–348, 2015.

Byron C Lewis and Albert E Crews. The evolution of benchmarking as a computer performance evaluation technique. *MIS Quarterly*, pages 7–16, 1985.

Mark Liberman. Fred Jelinek. *Computational Linguistics*, 36(4):595–599, 2010.

Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, 128(2):261–318, 2020.

John McCarthy. Al as sport. *Science*, 276(5318):1518, 1997.

Michele Merler, Nalini Ratha, Rogerio S. Feris, and John R. Smith. Diversity in faces, 2019. https://arxiv.org/abs/1901.10436.

George A Miller. Wordnet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596.

National Institute of Standards and Technology. U.S. leadership in AI: A plan for federal engagement in developing technical standards and related tools, 2019. URL https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf.

Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy, 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1459.

Martha Palmer and Tim Finn. Workshop on the evaluation of natural language processing systems. In *Computational Linguistics*, pages 175–181, 1990.

Kenny Peng. Facial recognition datasets are being widely used despite being taken down due to ethical concerns. Here's how. *Freedom to Tinker*, 2020.

Cassio Pennachin and Ben Goertzel. *Contemporary Approaches to Artificial General Intelligence*, pages 1–30. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-68677-4. doi: 10.1007/978-3-540-68677-4_1. URL https://doi.org/10.1007/978-3-540-68677-4_1.

P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1090–1104, 2000.

John R Pierce. Whither speech recognition? *The journal of the acoustical society of america*, 46 (4B):1049–1051, 1969.

Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision?, 2020. https://arxiv.org/abs/2006.16923.

XiPeng Qiu, TianXiang Sun, YiGe Xu, YiGe Shao, Ning Dai, and XuanJing Huang. Pre-trained models for natural language processing: A survey. *SCIENCE CHINA Technological Sciences*, 63 (10):1872–1897, 2020. URL https://engine.scichina.com/publisher/ScienceChinaPress/journal/SCIENCECHINATechnologicalSciences/63/10/10.1007/s11431-020-1647-3,doi=.

Joanna Radin. "Digital Natives": How Medical and Indigenous Histories Matter for Big Data. *Osiris*, 32(1):43–64, September 2017. ISSN 0369-7827, 1933-8287. doi: 10.1086/693853.

Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 429–435, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314244. URL https://doi.org/10.1145/3306618.3314244.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL https://www.aclweb.org/anthology/P18-2124.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/recht19a.html.

Michael J Reddy. The conduit metaphor: A case of frame conflict in our language about language. In *Metaphor and Thought*, pages 164–201. 1979.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://www.aclweb.org/anthology/2020.acl-main.442.

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 *Advances in Neural Information Processing Systems*, pages 9179–9189, 2019.

David Schlangen. Targeting the benchmark: On methodology in current natural language processing research. *ArXiv*, abs/2007.04792, 2020.

Viktor Schlegel, Goran Nenadic, and Riza Batista-Navarro. Beyond leaderboards: A survey of methods for revealing weaknesses in natural language inference data and models, 2020. https://arxiv.org/abs/2005.14709.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *CoRR*, abs/1907.10597, 2019. URL http://arxiv.org/abs/1907.10597.

D. Sculley, Jasper Snoek, Alexander B. Wiltschko, and A. Rahimi. Winner's curse? On pace, progress, and empirical rigor. In *ICLR*, 2018.

Henry Shevlin, Karina Vold, Matthew Crosby, and Marta Halina. The limits of machine intelligence. *Science & Society*, 20(10), 2019.

Norman Stiles and Daniel Wilcox. *Grover and the Everything in the Whole Wide World Museum*. Random House, New York, 1974. Illustrations by Joe Mathieu.

Sarah M Stitzlein. Replacing the 'view from nowhere': A pragmatist-feminist science classroom. *Electronic Journal of Science Education*, 9(2), 2004.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL https://www.aclweb.org/anthology/P19-1355.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. Assessing the benchmarking capacity of machine reading comprehension datasets. In *AAAI*, pages 8918–8927, 2020.

Ilya Sutskever. Recent advances in deep learning and AI from OpenAI, 2018. Keynote talk at AI Frontiers Conference.

Rachel Thomas and David Uminsky. The problem with metrics is a fundamental problem for AI, 2020. https://arxiv.org/abs/2002.08512.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

Peter Voss. *Essentials of General Intelligence: The Direct Path to Artificial General Intelligence*, pages 131–157. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

Kiri L. Wagstaff. Machine learning that matters. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1851–1856, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*, 2019a.

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 *Advances in Neural Information Processing Systems*, pages 3266–3280, 2019b.

Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. Disembodied machine learning: On the illusion of objectivity in NLP, 2020. https://openreview.net/forum?id=fkAxTMzy3fs.

Dominikus Wetzel and Francis Bond. Enriching parallel corpora for statistical machine translation with semantic negation rephrasing. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 20–29, Jeju, Republic of Korea, July 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W12-4203.

Rowan Zellers, Ari Holtzman, Elizabeth Clark, Lianhui Qin, Ali Farhadi, and Yejin Choi. Evaluating machines by their real-world language use, 2020. https://arxiv.org/abs/2004.03607.