Patterns

CellPress
OPEN ACCESS

Review

Data and its (dis)contents: A survey of dataset development and use in machine learning research

Amandalynne Paullada,[1,*] Inioluwa Deborah Raji,[3] Emily M. Bender,[1] Emily Denton,[2] and Alex Hanna[2,4]
[1]Department of Linguistics, University of Washington, Seattle, WA, USA
[2]Google Research, New York, NY, USA
[3]Mozilla Foundation, Mountain View, CA, USA
[4]Google Research, San Francisco, CA, USA
*Correspondence: paullada@uw.edu
https://doi.org/10.1016/j.patter.2021.100336

THE BIGGER PICTURE Datasets form the basis for training, evaluating, and benchmarking machine learning models and have played a foundational role in the advancement of the field. Furthermore, the ways in which we collect, construct, and share these datasets inform the kinds of problems the field pursues and the methods explored in algorithm development. In this work, we survey recent issues pertaining to data in machine learning research, focusing primarily on work in computer vision and natural language processing. We summarize concerns relating to the design, collection, maintenance, distribution, and use of machine learning datasets as well as broader disciplinary norms and cultures that pervade the field. We advocate a turn in the culture toward more careful practices of development, maintenance, and distribution of datasets that are attentive to limitations and societal impact while respecting the intellectual property and privacy rights of data creators and data subjects.

SUMMARY

In this work, we survey a breadth of literature that has revealed the limitations of predominant practices for dataset collection and use in the field of machine learning. We cover studies that critically review the design and development of datasets with a focus on negative societal impacts and poor outcomes for system performance. We also cover approaches to filtering and augmenting data and modeling techniques aimed at mitigating the impact of bias in datasets. Finally, we discuss works that have studied data practices, cultures, and disciplinary norms and discuss implications for the legal, ethical, and functional challenges the field continues to face. Based on these findings, we advocate for the use of both qualitative and quantitative approaches to more carefully document and analyze datasets during the creation and usage phases.

INTRODUCTION

The importance of datasets for machine learning research cannot be overstated. Datasets have been seen as the limiting factor for algorithmic development and scientific progress,[1,2] and a select few benchmark datasets, such as the ImageNet benchmark for visual object recognition[3] and the GLUE benchmark for English textual understanding,[4] have been the foundation for some of the most significant developments in the field. Benchmark datasets have also played a critical role in orienting the goals, values, and research agendas of the machine learning community.[5] In recent years, machine learning systems have been reported to achieve "super-human" performance when evaluated on such benchmark datasets. However, recent work from a variety of perspectives has surfaced not only the shortcomings of some machine learning datasets as meaningful tests of human-like reasoning ability, but also the troubling realities of the societal impact of how these datasets are developed and used. Together, these insights reveal how this apparent progress may rest on faulty foundations.

As the machine learning field turned to approaches with larger data requirements in the last decade, the sort of skilled and methodical annotation applied in dataset collection practices in earlier eras was spurned as "slow and expensive to acquire," and a turn toward unfettered collection of increasingly large amounts of data from the web, alongside increased reliance on crowdworkers, was seen as a boon to machine learning.[1,3,6,7] The enormous scale of such datasets has been mythologized as beneficial to the perceived generality of trained systems,[7] but they continue to be impacted by the limitations and biases that impact all datasets.[8] In particular, prevailing data practices tend to abstract away the human labor,

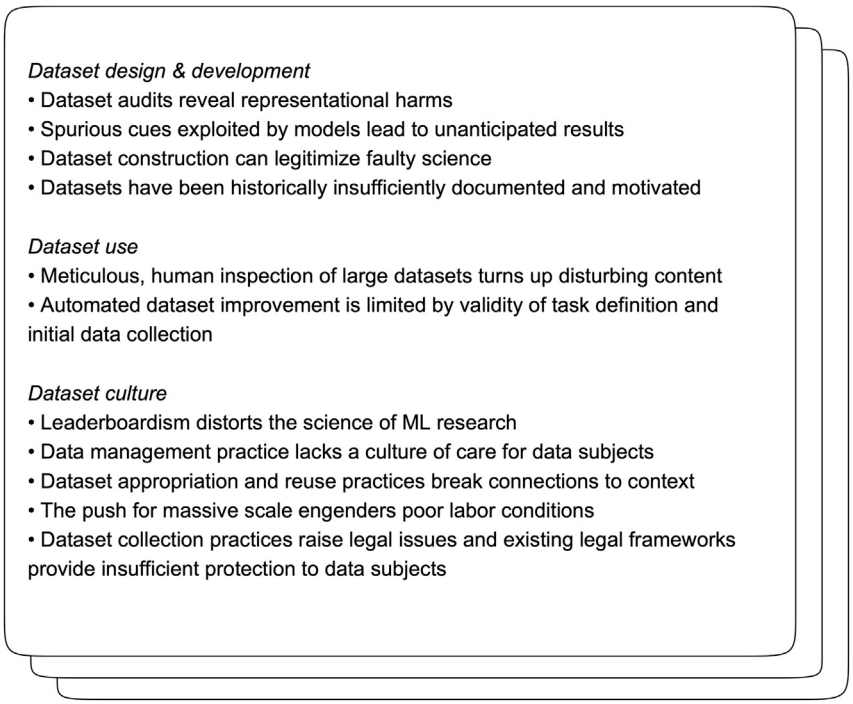

Figure 1. Key takeaways from a survey of perspectives on the challenges posed by recent trends in dataset use in machine learning

subjective judgments and biases, and contingent contexts involved in dataset production. However, such details are important for assessing whether and how a dataset might be useful for a particular application, for enabling more rigorous error analysis, and for acknowledging the significant difficulty required in constructing useful datasets.

The machine learning field has placed large-scale datasets at the center of model development and evaluation. As systems trained in this way are deployed in real-world contexts that affect the lives and livelihoods of real people, it is essential that researchers, advocacy groups, and the public at large understand both the contents of the datasets and how they affect system performance. In particular, as the field has focused on benchmarks as the primary tool for both measuring and driving research progress,[9] understanding what these benchmarks measure (and how well) becomes increasingly urgent.

We thus conduct a survey of the literature of recent issues pertaining to data in machine learning research, with a particular focus on work in computer vision and natural language processing (NLP). We structure our survey around three themes. The first, Dataset design and development, deals with studies that critically review the design of the datasets used as benchmarks. This includes studies that audit existing datasets for bias, those that examine existing datasets for spurious correlations which make the benchmarks gameable, those that critically analyze the framing of tasks , and work promoting better data collection and documentation practices. Next, in Dataset in(tro)spection, we review approaches to exploring and improving these aspects of datasets. In looking at approaches to filtering and augmenting data and modeling techniques aimed at mitigating the impact of bias in datasets, we see further critiques of the current state of datasets. However, we find that these approaches do not fully address the broader issues with data use. Finally, in Dataset culture, we survey work on dataset practices as a whole, including critiques of their use as performance targets, perspectives on data management and reuse, research into the precarious labor conditions that underpin much of dataset production, and papers raising legal issues pertaining to data collection and distribution . The key findings of the sections which form the body of the paper are summarized in Figure 1.

DEFINITIONS

We follow Schlangen[9] in distinguishing between *benchmarks*, *tasks*, *capabilities*, and *datasets*. While his work focused on NLP, we broaden these definitions to include aspects of other machine learning applications. In this context, a *task* is constituted of an input space and output space and an expected mapping between them. Schlangen notes that there are typically both *intensional* and *extensional* definitions of tasks. An intensional definition describes the relationship between input and output (e.g., the output in automatic speech recognition is a transcription of the audio signal in the input), where an extensional definition is simply the set of input-output pairs in the dataset. Thus tasks are exemplified by *datasets*, i.e., sets of input-output pairs that conform, if valid, to the intensional definition of the task. Tasks can be of interest for two (not mutually exclusive) reasons: either they map directly to a use case (e.g., automatic transcription of audio data) or they illustrate cognitive *capabilities*, typical of humans, that we are attempting to program into machines. In the former case, a task is suitable as a *benchmark* (for comparing competing systems to each other) if the task is well-aligned with its real-world use case and the dataset is sufficiently representative of the data the systems would encounter in production. In the latter case, establishing the value of the task as a benchmark is more involved: as Schlangen argues, success on the task has to be shown to rely on having some set of capabilities that are definable outside of the task itself and transferable to other tasks.

In referring to dataset exemplars that pair instances (input) and labels (output), we follow a convention from machine learning of referring to the latter as *target labels*, which are those labels that are used as the learning target, and which have typically been produced by human annotators or, in some cases, automated labeling heuristics. These are also often referred to in the literature as "gold standard" or "ground truth" labels, but we wish to emphasize their role as training targets that are neither objective nor necessarily representative of reality.

DATASET DESIGN AND DEVELOPMENT

> "Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care."—Geoffrey Bowker (*Memory Practices in the Sciences*)[10]

In this section, we review papers that explore issues with the contents of datasets that arise due to the manner in which they were collected, the assumptions guiding the dataset construction process, and the set of questions guiding their development.

Representational harms in datasets

In recent years there has been growing concern regarding the degree and manner of representation of different sociodemographic groups within prominent machine learning datasets, constituting what Kate Crawford has called *representational harms.*[11] For example, a glaring under-representation of darker-skinned subjects, compared with lighter-skinned subjects, has been identified within prominent facial analysis datasets[6,12] and in image datasets used to train self-driving cars to detect pedestrians.[13] Meanwhile, the images in object recognition datasets have been overwhelmingly sourced from Western countries.[14] Zhao et al.[15] found a stark under-representation of female pronouns in the commonly used OntoNotes dataset for English coreference resolution; similarly, Lennon[16] found that feminine-coded names were vastly underrepresented in the CoNLL-2003 dataset used for named entity recognition. While the under-representation of marginalized groups in datasets has been met with calls for "inclusion," Hoffmann[17] provides a case for skepticism of this narrative, as it has the potential to merely uphold the very sort of power hierarchy that engenders such under-representation in the first place.

Stereotype-aligned correlations have also been identified in both computer vision and NLP datasets. For example, word co-occurrences in NLP datasets frequently reflect social biases and stereotypes relating to race, gender, (dis)ability, and more[18,19] and correlations between gender and activities depicted in computer vision datasets have been shown to reflect common gender stereotypes.[20–22] Dixon et al.[23] found that a dataset for toxicity classification contained a disproportionate association between words describing queer identities and text labeled as "toxic," while Park et al.[24] found evidence of gender bias against women in similar datasets. Such disparities in representation stem, in part, from the fact that particular, non-neutral viewpoints are routinely yet implicitly invoked in the design of tasks and labeling heuristics. For example, a survey of literature on computer vision systems for detecting pornography found that the task is largely framed around detecting the features of thin, nude, female-presenting bodies with little body hair, largely to the exclusion of other kinds of bodies—thereby implicitly assuming a relatively narrow and conservative view of pornography that happens to align with a straight male gaze.[25]

In an examination of the person categories within the ImageNet dataset,[3] Crawford and Paglen[26] uncovered millions of images of people that had been labeled with offensive categories, including racial slurs and derogatory phrases. In a similar vein, Birhane and Prabhu[27] examined a broader swath of image classification datasets that were constructed using the same categorical schema as ImageNet, finding a range of harmful and problematic representations, including non-consensual and pornographic imagery of women. In response to the work of Crawford and Paglen,[26] a large portion of the ImageNet dataset has been removed.[28] Similarly, Birhane and Prabhu's examination[27] prompted the complete removal of the TinyImages dataset.[29]

Spurious cues exploited by machine learning models

While deep learning models have seemed to achieve remarkable performance on challenging tasks in artificial intelligence, recent work has illustrated how these performance gains may be due largely to "cheap tricks" (to borrow a term from Levesque[30]) rather than human-like reasoning *capabilities*, as defined in Definitions. Geirhos et al.[31] illustrate how the performance of deep neural networks can rely on *shortcuts*, or decision rules that do not extrapolate well to out-of-distribution data and are often based on incidental associations. Oftentimes, these shortcuts arise due to artifacts in datasets that allow models to overfit to training data and to rely on nonsensical heuristics to "solve" the task—for example, detecting the presence of pneumonia in chest X-ray scans based on hospital-specific tokens that appear in the images.[31] That is, despite high predictive performance, models are not performing the task according to its *intensional* description, and thus the datasets may not be exemplary of reasoning *capabilities*. Recent work has revealed the presence of shortcuts in commonly used datasets that had been conceived of as proving grounds for particular competencies, such as reading comprehension and other "language understanding" capabilities. Experiments that illuminate such data artifacts, or "dataset ablations" as Heinzerling[32] calls them, involve simple or nonsensical baselines, such as training models on incomplete inputs and comparing performance to models trained on full inputs. Much recent work in NLP has revealed how these simple baselines are competitive, and that models trained on incomplete inputs for argument reasoning, natural language inference, fact verification, and reading comprehension—i.e., tasks restructured in such a way that there should be no information about the correct output in the input—perform quite well.[33–37] (Storks et al.[38] and Schlegel et al.[39] provide more comprehensive reviews of datasets and dataset ablations for natural language inference.) In many cases, this work has revealed how an over-representation of simple linguistic patterns (such as negation or presence of certain words) in dataset instances belonging to one label class can serve as a spurious signal for models to pick up on. Many of these issues result from the assumptions made in task design and in the underspecification of instructions given to human data labelers, and can thus can be addressed by rethinking the format that dataset collection takes. In light of this, recent work has proposed approaches to pre-empting spurious correlations by designing annotation frameworks that better leverage human "common sense"[40] and more critical approaches to dataset creation and use for tasks such as reading comprehension.[41]

How do datasets legitimize certain problems or goals?

As the previous sections have laid out, the mapping between inputs and target labels contained in datasets is not always a meaningful one, and the ways in which data are collected and

tasks are structured can lead models to rely on faulty heuristics for making predictions. The problems this raises are not limited to misleading conclusions based on benchmarking studies: when machine learning models can leverage spurious cues to make predictions well enough to beat a baseline on the test data, the resulting systems can appear to legitimize spurious tasks that do not map to real-world capabilities. More formally, there are some tasks that can be described *intensionally* but for which there is no possibility of a sufficient *extensional* realization, often because the underlying theory for the task is unsound.

Decisions about what data to collect in the first place and the problematization that guides data collection lead to the creation of datasets that formulate pseudoscientific tasks. For example, several studies in recent years that attempt to predict attributes such as sexuality and other fluid, subjective personal traits from photos of human faces presuppose that these predictions are possible and worthwhile to make. However, these datasets, like those discussed above, enable a reliance on meaningless shortcuts. These in turn support the apparent "learnability" of the personal traits in question. An audit by Agüera y Arcas et al.[42] found that a model trained to predict sexual orientation from images of faces harvested from online dating profiles was actually learning to spot stereotypical choices in grooming and self-expression, which are by no means universal, while Gelman et al. discuss how such a study strips away context and implies the existence of an "essential homosexual nature".[43] The task rests on a pseudoscientific essentialism of human traits. Another example, from NLP, is GermEval 2020 Task 1,[44] which asked systems to reproduce a ranking of students by IQ scores and grades using only German short answer texts produced by the students as input. By setting up this task as feasible (for machine models or otherwise), the task organizers suggested that short answer texts contain sufficient information to "predict" IQ scores and furthermore that IQ scores are a valid and relevant thing to measure about a person.[45] Jacobsen et al.[46] point out that shortcuts in deep learning, as described in Section 3.2, make ethically dubious questions seem answerable, and advise, "[W]hen assessing whether a task is solvable, we first need to ask: should it be solved? And if so, should it be solved by AI?" Not only are these task formulations problematic, but, as we describe in Section 5.3, once sensitive data has been collected, it can be misused.

Collection, annotation, and documentation practices

A host of concerns regarding the practices of dataset collection, annotation, and documentation have been raised within recent years. In combination, these concerns reflect what Jo and Gebru[47] describe as a *laissez-faire* attitude regarding dataset development and the pervasive undervaluation of data work.[48] Rather than collecting and curating datasets with care and intentionality—as is more typical in other data-centric disciplines—machine learning practitioners often adopt an approach where anything goes. As one data scientist put it, "if it is available to us, we ingest it."[49]

The common practices of scraping data from internet search engines, social media platforms, and other publicly available online sources faced significant backlash in recent years. For example, facial analysis datasets have received push-back due to the inclusion of personal Flickr photos without data subjects' knowledge.[50] In many instances, the legality of the data usage has come into question, as we discuss further in Legal perspectives.

Dataset annotation practices have also come under great scrutiny in recent years. Much of this has focused on how subjective values, judgments, and biases of annotators contribute to undesirable or unintended dataset bias.[22,51–54] More generally, several researchers have identified a widespread failure to recognize annotation work as *interpretive work*, which in turn can result in a conflation of *target* labels in a collected dataset and *real-world* objects, for which there may be no single ground truth label.[55,56] As discussed further in Labor, data annotation tasks are often mediated through crowdwork platforms such as Amazon Mechanical Turk (AMT). These platforms, by design, position annotators as interchangeable workers, rather than individuals who bring to bear their own subjective experiences and interpretations to the task. Divergences in judgments across different annotator pools,[57] as well as between AMT annotators and other communities,[58] have been empirically explored.

Recent work by Tsipras et al.[59] has revealed that the annotation pipeline for ImageNet does not reflect the intention of its development for the purpose of object recognition in images. They note that ImageNet, constructed with the constraint of a single label per image, had its labels largely determined by crowdworkers indicating the visual presence of that object in the image. This has led to issues with how labels are applied, particularly to images with multiple objects, where the class of interest could include a background or obscured object that would be an unsuitable result for the image classification task of that particular photo. Furthermore, the nature of image retrieval for the annotation tasks biases the crowdworkers' response to the labeling prompt, making them much less effective at filtering out unsuitable examples for a class category. This is just one of several inconsistencies and biases in the data that hints at larger annotation patterns that mischaracterize the real-world tasks these datasets are meant to represent, and the broader impact of data curation design choices in determining the quality of the final dataset.

Dataset documentation practices have also been a central focus, especially as dataset development processes are increasingly being recognized as a source of algorithmic unfairness. A recent study of publications that leverage Twitter data found data decisions were heavily under-specified and inconsistent across publications.[60] Scheuerman et al.[61] found a widespread under-specification of annotation processes relating to gender and racial categories within facial analysis datasets. Several dataset documentation and development frameworks have been proposed in an effort to address these concerns, with certain frameworks looking to not just capture characteristics of the output dataset but also report details of the procedure of dataset creation for better transparency and accountability.[62–66]

The lack of rigorous and standardized dataset documentation practices has contributed to reproducibility concerns. For example, recent work by Recht et al.[67] undertook the laborious task of reconstructing ImageNet, following the original documented dataset construction process in an effort to test the generalization capabilities of ImageNet classifiers. Despite mirroring the original collection and annotation methods—including

leveraging images from the same time period—the newly constructed dataset was found to have different distributional properties. The differences were largely localized to variations in constructing target labels from multiple annotations. More specifically, different thresholds for inter-annotator agreement were found to produce vastly different datasets, indicating that so-called ground truth labels in datasets do not correspond to truth.

Summary

This section has centered on issues with dataset contents and structures, and the representational harms, spurious correlations, problem legitimization, and haphazard collection, annotation, and documentation practices that are endemic to many machine learning datasets. In the next section, we review methods which have been developed to address some of these issues.

DATASET IN(TRO)SPECTION

> "For any sociotechnical system, ask, 'what is being looked at, what good comes from seeing it, and what are we *not* able to see?'"—Mike Ananny and Kate Crawford (*Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability*)[68]

The massive sizes of contemporary machine learning datasets make it intractable to thoroughly scrutinize their contents,[69] and thus it is hard to know where to begin looking for the kinds of representational and statistical biases outlined in the previous section. Indeed, a culture characterized by a desire to harness large datasets without questioning what is in them or how it got there, no matter how unsavory the details might be, produces what machine learning researcher Vinay Prabhu calls the "abattoir effect."[70] While many of the dysfunctional contents discovered in datasets were found by using intuition and domain expertise to construct well-designed dataset ablations and audits, some of the most disturbing were found by manually combing through the data.

Further insight into issues with dataset contents can be found in work that attempts to identify and address some of the problems outlined in the previous section. In this section, we review a variety of methods for exploring the contents of datasets in support of discovering and mitigating the issues that lurk within datasets.

Inspection

Birhane and Prabhu,[27] summarized in the previous section, and Pipkin[71] show how meticulous manual audits of large datasets are compelling ways to discover the most surprising and disturbing contents therein. Pipkin spent hundreds of hours watching the entirety of MIT's "Moments in Time" video dataset,[72] finding shocking and unexpected footage of violence, assault, and death. They provocatively point out, in a reflection on the process of developing their artistic intervention *Lacework*, that the researchers who commission the curation of massive datasets may have less intimate familiarity with the contents of these datasets than those who are paid to look at and label individual instances, and, as we discuss in Labor, there is growing awareness of the need to better support the workers at the front lines of the often grim and under-valued work of data labeling. Caswell et al.[73] show the value of manual audits of multilingual corpora to highlight the dubious quality of many datasets used for language model training. Their team of human volunteers, with proficiency in about 70 languages altogether, found that several corpora scraped from the web are rife with examples of mistranslated text and mislabeled linguistic content (i.e., content in a particular language labeled erroneously as belonging to another language).

Introspection

While manual audits have provided invaluable insights into the contents of datasets, as datasets swell in size this technique is not scalable. Recent work has proposed algorithmic interventions that assist in the exploration and adjustment of datasets. Some methods leverage statistical properties of datasets to surface spurious cues and other possible issues with dataset contents. The AFLITE algorithm proposed by Sakaguchi et al.[74] provides a way to systematically identify dataset instances that are easily gamed by a model, but in ways that are not easily detected by humans. This algorithm is applied by Le Bras et al.[75] to a variety of NLP datasets, and they find that training models on adversarially filtered data leads to better generalization to out-of-distribution data. In addition, recent work by Swayamdipta et al.[76] proposes methods for performing exploratory data analyses based on training dynamics that reveal edge cases in the data, bringing to light labeling errors or ambiguous labels in datasets. Northcutt et al.[77] combine an algorithmic approach with human validation to surface labeling errors in the test set for ImageNet.

Han et al.[78] demonstrate the application of influence functions, originally introduced by Koh and Liang[79] as a way to identify the influence of particular training examples on model predictions, to the discovery of data artifacts. The REVISE tool by Wang et al.[80] can be used to identify unequal representation in image description datasets by leveraging features of the images and the corresponding texts. Using their tool, they spot that images of outdoor athletes are overwhelmingly labeled as men, and that in images where a person is too small for any sort of gender to be told at all, they are still labeled as men.

In response to a proliferation of challenging perturbations derived from existing datasets to improve generalization capabilities and lessen the ability for models to learn shortcuts, Liu et al.[81] propose "inoculation by fine-tuning" as a method for interpreting what model failures on perturbed inputs reveal about weaknesses of training data (or models). Recent papers also outline methodologies for leveraging human insight in the manual construction of counterfactual examples that complement instances in NLP datasets to promote better generalization.[82,83]

The case of VQA-CP[84] provides a cautionary tale of when a perturbed version of a dataset is, itself, prone to spurious cues. This complement to the original Visual Question Answering (VQA) dataset, consisting of VQA instances redistributed across train and test sets as an out-of-domain benchmark for the task, was found to be easy to "solve" with randomly generated answers. Cleverly designed sabotages that are meant to strengthen models' ability to generalize may ultimately follow

the same patterns as the original data, and are thus prone to the same kinds of artifacts. While this has prompted attempts to make models more robust to any kind of dataset artifact, it also suggests that there is a broader view to be taken with respect to rethinking how we construct datasets for tasks overall.

Considering that datasets will always be imperfect representations of real-world capabilities, recent work proposes methods of mitigating the impacts of noise in data on model performance. Teney et al.[85] propose an auxiliary training objective using counterfactually labeled data to guide models toward better decision boundaries. He et al.[86] propose the DRiFT algorithm for "unlearning" dataset bias. Sometimes, noise in datasets is not symptomatic of statistical anomalies or labeling errors, but rather, a reflection of variability in human judgment. Pavlick and Kwiatkowski[87] find that human judgment on natural language inference tasks is variable, and that machine evaluation on this task should reflect this variability.

Many of the methods outlined in this section crucially rely on statistical patterns in the data to surface problematic instances; it is up to human judgment to make sense of the nature of these problematic instances, whether they represent logical inconsistencies with the task at hand, cases of injustice, or both. In addition, while a variety of recent papers have proposed methods for removing spurious cues from training data or "de-biasing" models, recent work has shown that this can be damaging for model accuracy.[88]

In contrast to a focus on statistical properties of datasets as a site for addressing and mitigating harms, Denton et al.[89] propose a research agenda in the "data genealogy" paradigm that promotes critical assessment of the design choices with respect to the data sources, theoretical motivations, and methods used for constructing datasets. Prospective accounting for dataset contents using some of the methods discussed at the end of the previous section can offset the potential of post-hoc documentation debt that can be incurred otherwise.[69]

Summary

In this section we have reviewed a variety of works that address dataset content issues by providing lenses on data for inspection and introspection. We emphasize that procedural dataset modifications and bias mitigation techniques are only useful insofar as the dataset in question itself represents a well-designed task. In making lemonade from lemons, we must ensure the lemons are not ill-gotten or poorly formed.

DATASET CULTURE

> "Every data set involving people implies subjects and objects, those who collect and those who make up the collected. It is imperative to remember that on both sides we have human beings."—Mimi Ọnụọha (*The Point of Collection*)[90]

A final layer of critiques looks at the culture around dataset use in machine learning. In this section, we examine how common practices in dataset usage impact society at large by reviewing papers that ask: What are issues with the broader culture of dataset use? How do our dataset usage, storage, and re-usage practices wrench data away from their contexts of creation? What are the labor conditions under which large-scale crowdsourced datasets are produced? Finally, what can be learned from looking at machine learning dataset culture from a legal perspective?

Benchmarking practices

Benchmark datasets play a critical role in orienting the goals of machine learning communities and tracking progress within the field.[5,89] Yet, the near singular focus on improving benchmark metrics has been critiqued from a variety of perspectives. Indeed, the current benchmarking culture has been criticized as having the potential to stunt the development of new ideas.[91] NLP researchers have exhibited growing concern with the singular focus on benchmark metrics, with several calls to include more comprehensive evaluations—including reports of energy consumption, model size, fairness metrics, and more—in addition to standard top-line metrics.[92–94] Sculley et al.[95] examine the incentive structures that encourage singular focus on benchmark metrics—often at the expense of empirical rigor—and offer a range of suggestions including incentivizing detailed empirical evaluations, including negative results, and sharing additional experimental details. From a fairness perspective, researchers have called for the inclusion of disaggregated evaluation metrics, in addition to standard top-line metrics, when reporting and documenting model performance.[96]

The excitement surrounding leaderboards and challenges can also give rise to a misconstrual of what high performance on a benchmark actually entails. In response to the recent onslaught of publications misrepresenting the capabilities of BERT language models, Bender and Koller[97] encourage NLP researchers to be attentive to the limitations of tasks and include error analysis in addition to standard performance metrics.

Sen et al.[58] have questioned the legitimacy of the notion of a gold standard dataset for certain tasks, empirically demonstrating divergences between gold standards set by AMT workers and those from other communities. Other data-oriented fields have grappled with the politics inherent in quantification and measurement practices.[98–100] Jacobs and Wallach[101] locate dataset measurement concerns as a key factor underlying unfair outcomes of algorithmic systems and propose that machine learning practitioners adopt measurement modeling frameworks from the quantitative social sciences.

Data management and distribution

Secure storage and appropriate dissemination of human-derived data is a key component of data ethics.[102] To have a culture of care for the subjects of the datasets we make use of requires us to prioritize the well-being of the subjects in the dataset throughout collection, development, *and* distribution. To do so systematically, the machine learning community still has much to learn from other disciplines with respect to how they handle the data of human subjects. Unlike in the social sciences or medicine, the machine learning field has yet to develop the data management practices required to store and transmit sensitive human data.

Metcalf and Crawford[103] go so far as to suggest the re-framing of data science as human subjects research, indicating the need

for institutional review boards and informed consent as researchers make decisions about other people's personal information. Particularly in consideration of an international context, where privacy concerns may be less regulated in certain regions, the potential for data exploitation is a real threat to the safety and well-being of data subjects.[104] As a result, those that are the most vulnerable are at risk of losing control of the way in which their own personal information is handled. Without individual control of personal information, anyone who happens to be given the opportunity to access their unprotected data can act with little oversight, potentially against the interests or well-being of data subjects. This can become especially problematic and dangerous in the most sensitive contexts, such as personal finance information, medical data, or biometrics.[105]

However, machine learning researchers developing such datasets rarely pay attention to this necessary consideration. Researchers will regularly distribute biometric information—for example, face image data—without so much as a distribution request form or required privacy policy in place. Furthermore, the images are often collected without any level of informed consent or participation.[6,50,106] In the context of massive data collection projects, the potential harms extend beyond those that can be addressed with individual consent. Solove[107] provides a thoughtful overview of privacy as both a societal and individual value.

Even when these datasets are flagged for removal by the creators, researchers will still attempt to make use of that now illicit information through derivative versions and backchannels. For example, Peng[108] finds that, after certain problematic face datasets were removed, hundreds of researchers continued to cite and make use of copies of this dataset months later. Without any centralized structure of data governance for the research in the field, it becomes nearly impossible to take any kind of significant action to block or otherwise prevent the active dissemination of such harmful datasets.

Security concerns arise due to the manner in which large-scale datasets are curated and disseminated through a web-scraping paradigm. For example, it was recently discovered that one of the URLs in the ImageNet dataset that originally pointed to an image of a bat instead linked to malware, potentially making dataset users vulnerable to hacking.[109] Carlini et al.[110] also illustrate how large language models can be prodded to disgorge sensitive, personally identifying information they have picked up from their training data.

Best practices for sharing and managing datasets are a burgeoning area of research in NLP. In addition to a comprehensive accounting for the motivations and contents of abusive language datasets, Vidgen and Derczynski[111] provide several suggestions for the responsible dissemination of such data, including the establishment of data trusts, platform-supported datasets, and the use of synthetic data.

Use and reuse

Several scholars have written on the importance of reusable data and code for reproducibility and replicability in machine learning,[112,113] and the publication of scientific data is often seen as an unmitigated good, either in the pursuit of reproducibility[114] or as a means of focusing research effort and growing research communities (e.g., through shared task evaluations[115]). Here, we want to consider the potential pitfalls of taking data that had been collected for one purpose and using it for one in which it was not intended, particularly when this data reuse is morally and ethically objectionable to the original curators. Science and technology scholars have considered the potential incompatibilities and reconstructions needed in using data from one domain in another.[116] Indeed, Strasser and Edwards discuss several major questions for Big Data in science and engineering, asking critically "Who owns the data?" and "Who uses the data?".[117] Although in Legal perspectives we discuss ownership in a legal sense, ownership also suggests an inquiry into who the data have come from, such as the "literal […] DNA sequences" of individuals[117] or other biometric information. In this case, considering data reuse becomes a pivotal issue of benchmark datasets.

Instances of data reuse in benchmarks are often seen in the scraping and mining context, especially when it comes to Flickr, Wikipedia, and other openly licensed data instances. Many of the instances in which machine learning datasets drawn from these and other sources in ways that incur serious privacy violations are well-documented by Harvey and LaPlace,[106] who discuss instances of scraping Flickr and other image hosting services for human images without explicit user consent.

The reuse of data can involve reusing data from one context and using this decontextualized data for machine learning applications. This dynamic is exemplified well by historian of science Joanna Radin's exploration of the peculiar history of the Pima Indians Diabetes Dataset (PIDD) and its introduction into the UCI Machine Learning Repository.[118] The PIDD has been used thousands of times as a "toy" classification task and currently lives in the UCI repository, a major repository for machine learning datasets. The data were collected by the National Institutes of Health from the Indigenous community living at the Gila River Indian Community Reservation, which had been extensively studied and restudied for their high prevalence of diabetes. In her history of this dataset, Radin is attentive to the politics of the creation and processing of the data itself. The fact that "data was used to refine algorithms that had nothing to do with diabetes or even to do with bodies, is exemplary of the history of Big Data writ large". Moreover, the residents of the Reservation, who refer to themselves as the Akimel O'odham, had been the subject of intense anthropological and biomedical research, especially due to a high prevalence of diabetes, which in and of itself stemmed from a history of displacement and settler-colonialism. However, their participation in research had not yielded any significant decreases in obesity or diabetes among community members.

Another concerning example of data reuse occurs when derivative versions of an original dataset are distributed—beyond the control of its curators—without any actionable recourse for removal. The DukeMTMC (Duke Multi-Target, Multi-Camera) dataset was collected from surveillance video footage from eight cameras on the Duke campus in 2014, used without consent of the individuals in the images and distributed openly to researchers in the US, Europe, and China. After reporting in the *Financial Times*[119] and research by Harvey and LaPlace,[106] the dataset was taken down on June 2, 2019. However, Peng[108] has recently highlighted how the dataset and its derivatives are still freely available for download and used in scientific

publications. It is nearly impossible for researchers to maintain control of datasets once they are released openly or if they are not closely supervised by institutional data repositories.

Across all of these instances of data use and reuse we observe that, when datasets are not created specifically and only for the use of machine learning research, there is the potential for a culture clash between the data practices of machine learning and the data practices of the field where the data comes from. Currently, machine learning (and computer science more generally) is relatively powerful compared with many other academic disciplines and risks exporting its data practices, whereas we argue that, as a field, we should be looking to learn from other fields' approaches to appropriate and situated data handling (see, e.g., Jo and Gebru[47]). We further note that a culture change around data use and reuse is a field-level problem that will require community buy-in and field-level allocation of resources to address. For example, to address the ways in which deprecated datasets, such as DukeMTMC, continue to circulate it is not enough to create a central repository that holds information about dataset retraction and other updates; researchers must also be incentivized and trained to consult such repositories.

Labor

As the machine learning community has increasingly turned to the cheap and scalable work forces offered by crowd sourcing platforms, there has been growing concern regarding the working conditions of those laboring on machine learning datasets. Data annotation is often cast as unskilled work—work *anyone* can perform—which in turn contributes to a dehumanizing and alienating work experience. For example, Irani[120] describes how crowdwork platforms, such as AMT, create a hierarchy of data labor, positioning crowdwork as menial work relative to the innovative work of those leading dataset development. Miceli et al.[55] discuss how, in commercial data annotation companies, power asymmetries and company hierarchies affect the work output of data annotation teams. Framing data annotation as unskilled work frames crowdworkers as essentially interchangeable, and creates the infrastructural conditions of precarity and invisibility.[121–123] For example, crowd-sourced data annotation is typically mediated through digital interfaces that distance the crowdworkers from the dataset developers constructing annotation tasks, rendering both the workers and the labor concerns they might face invisible.[124,125] Such labor concerns include low and unstable wages, unfair treatment by task requesters, and barriers to worker solidarity and collective action.[126–128]

In response to these growing concerns, guidelines and tools for task creators have been developed to help facilitate fair pay[129,130] and interventions oriented at crowdworkers directly have been developed to support worker solidarity[124,131] and fair pay.[132] Gray and Suri[128] also discuss corporate interventions, such as providing collaborative online discussion spaces, offline shared workspaces, and portable reputation systems, as well as governmental responses, such as the construction of worker guilds, unions, and platform cooperatives, and the provision of a social safety network for these precarious workers.

As personal data are increasingly commodified by technology companies and harvested at scale to improve proprietary machine learning systems, often in ways that are by turns inscrutable or distasteful to the general public,[133] recent proposals call for not only re-framing data as labor,[134] but also for "data strikes" in which users collectively withhold their data as a means to shift the power imbalance back toward subjects who are not compensated for the ambient collection of their data.[135]

Legal perspectives

The above subsections surveyed a range of literature critiquing different aspects of dataset culture in machine learning. In this section, we review literature that looks at the collection and use of datasets from a legal perspective, considering both the legal risks that dataset collectors incur and the extent to which existing legal frameworks protect data subjects. Benchmark datasets are often mined from the internet, collecting data instances that have various levels of licensing attached and storing them into a single repository. Different legal issues arise at each stage in the data-processing pipeline, from collection to annotation, from training to evaluation, from inference and the reuse of downstream representations, such as word embeddings and convolutional features.[136] Legal issues also arise that impact a host of different people in the process, including dataset curators, AI researchers, copyright holders, data subjects (those people whose likenesses, utterances, or representations are in the data), and consumers (those who are not in the data but are impacted by the inferences of the AI system). Different areas of law can protect (and also possibly harm) each of the different actors in turn.[137]

Benchmark datasets are drawn from a number of different sources, each with a different configuration of copyright holders and permissions for their use in training and evaluation in machine learning models. For instance, ImageNet was collected through several image search engines where licensing/copyright restrictions on data instances in those images are unknown.[138] The ImageNet project does not host the images on their website, and therefore sidesteps the copyright question by claiming that they operate like a search engine[139] (fn. 36). PASCAL VOC was collected via the Flickr API, meaning that the images were all held through the Creative Commons license.[140] Open licenses, such as Creative Commons, allow for training of machine learning models under fair use doctrine.[141] Faces in the Wild and Labeled Faces in the Wild were collected through Yahoo News, and via an investigation of the captions on the images we can see that the major copyright holders of those images are news wire services, including the Associated Press and Reuters.[142] Other datasets are collected in a studio environment, where images were taken by dataset curators and therefore are copyright holders, which avoids potential copyright issues.

US copyright law is not well-suited to cover the range of uses of benchmark datasets, and there is limited case law establishing precedent in this area. Legal scholars have defended the use of copyrighted material for data science and machine learning by suggesting that this material's usage is protected by fair use, since it entails the non-expressive use of expressive materials.[143] In contrast, Levendowski[139] has argued that copyright is actually a useful tool for battling algorithmic bias by offering a larger pool of works from which machine learning practitioners can draw from. She argues that, given that pre-trained

representations, such as `word2vec` and other word embeddings, suffer from gender and racial bias,[144,145] and other public domain datasets are older or obtained through means likely to result in amplified representation of stereotypes and other biases in the data (e.g., the Enron text dataset), that using copyrighted data can battle biased datasets and their use would fall under copyright's fair use exception.

Even in cases in which all data were collected legally from a copyright perspective—such as through open licenses, like Creative Commons—many downstream questions remain, including issues about privacy, informed consent, and procedures of opt-out.[141] O'Sullivan[109] discusses how technically legal uses of personal data that are not anticipated by or fully disclosed to the original owners of the data, e.g., the use of images scraped from the web to train facial recognition algorithms, constitute the ethical equivalent of data theft. Copyright guarantees are not sufficient protections for safeguarding privacy rights of individuals, as seen in the collection of images for the Diversity in Faces and MegaFace datasets.[50,119] Potential privacy violations arise when datasets contain biometric information that can be used to identify individuals, including faces, fingerprints, gait, and voice among others. However, at least in the US, there is no national-level privacy law that deals with biometric privacy. A patchwork of laws exist in Illinois, California, and Virginia that have the potential to safeguard the privacy of data subjects and consumers. However, only the Illinois Biometric Privacy law requires corporate entities to provide notice to data subjects and obtain their written consent.[137]

The machine learning and AI research communities have responded to this crisis by attempting to outline alternatives to licensing which make sense for research and benchmarking practices more broadly. The Montreal Data License (https://montrealdatalicense.com/) outlines different contingencies for a particular dataset, including whether the dataset will be used in commercial versus non-commercial settings, whether representations will be generated from the dataset, whether users can annotate the label or use subsets of it, and more.[136] This is a step forward in clarifying the different ways in which the dataset can be used once it has been collected, and therefore is a clear boon for AI researchers who create their own data instances, such as photos developed in a studio or text or captions written by crowdworkers. However, this does not deal with the larger issue of the copyright status of data instances scraped from the web, or the privacy implications of those data instances.

Summary

In this section, we have shed light on issues around benchmarking practices, dataset use and reuse, and the legal status of benchmark datasets. These issues are more about the peculiar practices of data in machine learning culture, rather than the technical challenges associated with benchmark datasets. In this way, we want to highlight how datasets work *as* culture—that is, "not [as] singular technical objects that enter into many different cultural interactions, but … rather [as] unstable objects, culturally enacted by the practices people use to engage with them".[146] Interrogating benchmark datasets from this view requires us to expand our frame from simply technical aspects of the system, to thinking how datasets intersect with communities of practice, communities of data subjects, and legal institutions.[147]

CONCLUSION

> "Not all speed is movement."—Toni Cade Bambara (*On the Issue of Roles*)[148]

In this paper, we present a survey of issues in dataset design and development, as well as reflections on the current broader culture of dataset use in machine learning. A viewpoint internal to this culture values rapid and massive progress: ever larger training datasets, used to train ever larger models, which post ever higher scores on ever harder benchmark tasks developed at a quicker and quicker pace. What emerges from the papers we survey, however, is a viewpoint, largely external to the current culture of dataset use, which reveals intertwined scientific and ethical concerns appealing to a more careful, systems-level and detail-oriented strategy.

The critiques of dataset design and development we survey in this paper highlight various different kinds of pitfalls: first, there are challenges with representation wherein datasets are biased both in terms of which data subjects are predominantly included and whose gaze is represented. Second, we find issues with the artifacts in the data, which machine learning models can easily leverage to "game" the tasks. Third, we find evidence of whole tasks which are spurious, where success is only possible given artifacts because the tasks themselves do not correspond to reasonable real-world correlations or capabilities. Finally, we find critiques of insufficiently careful data annotation and documentation practice, which erode the foundations of any scientific inquiry based on these datasets.

A variety of methods have been applied to examining dataset contents to surface some of the quality issues and harmful contents within. Attempts to rehabilitate datasets or models starting from the flawed datasets themselves further reinforce the problems outlined in the critiques of dataset design and development. The development of adversarial datasets or challenge sets, while possibly removing some spurious cues, does not address most of the other issues with either the original datasets or the research paradigm.

Critiques of the dataset culture itself focus on the overemphasis on benchmarking to the exclusion of other evaluation practices, legal and ethical issues in data management, distribution, and reuse, and labor practices in data curation. Hyperfocus on benchmarking pushes out work that connects models more carefully to their modeling domain and approaches not optimized for the available crop of benchmarks. The papers we surveyed suggest a need for work that takes a broader view than is afforded by the one-dimensional comparison of systems typical of benchmarks. Furthermore, critiques of data management and distribution show the need for growing a culture of care for the subjects of datasets in machine learning, i.e., to keep in mind that "data are people" and behave appropriately toward the people from whom we collect data.[149] Reflections on issues of data reuse emphasize the connection between data and its context, and the risks of harm (to data subjects and others) that arise when data is disconnected from its context and carried to and recontextualized in new domains. Legal

vulnerabilities inherent to current data collection and distribution practices in machine learning as well as the often precarious and under-compensated nature of dataset work reveal the complexities of data development and use within the context of society. Overall, these critiques shed light on the need for the kind of broader systems-level thinking required to navigate an undervalued although clearly necessary aspect of machine learning development.

What paths forward are visible from this broader viewpoint? We argue that fixes that focus narrowly on improving datasets by making them more representative or more challenging might miss the more general point raised by these critiques—namely that data challenges are in general under-considered in the field. If such data issues are left unaddressed, the field will be trapped in the Sisyphean task of finding and fixing dataset flaws rather than taking the necessary step back to address the more systematic issues at play. This renewed focus is essential to making progress as a field, so long as notions of progress are largely defined by performance on datasets. At the same time, we wish to recognize and honor the liberatory potential of datasets, when carefully designed, to make visible patterns of injustice in the world such that they may be addressed (see, for example, the work of Data for Black Lives [https://d4bl.org/]). Recent work by Register and Ko[150] illustrates how educational interventions that guide students through the process of collecting their own personal data and running it through machine learning pipelines can equip them with skills and technical literacy toward self-advocacy—a promising lesson for the next generation of machine learning practitioners and for those impacted by machine learning systems. We also recognize the need for a fundamental shift in the incentive structures that guide how machine learning practitioners prioritize dataset-related tasks. The introduction of a "Datasets and Benchmarks Track"[151] at the Neural Information Processing Systems Conference 2021, which will incentivize data-focused research, indicates a positive step in this direction.

We hope the response to this work goes beyond optimizing datasets to be "bigger" and "better"—a goal that does nothing to challenge the current paradigm of techniques idolizing speed and scale. Instead, we aspire for this survey to also prompt a more cautious and complex view of the considerations involved with data in the machine learning field. We advocate for a turn in the culture toward carefully collected datasets that are rooted in their original contexts, distributed in ways that respect the intellectual property and privacy rights of data creators and data subjects, and constructed in conversation with impacted stakeholders or domain experts. This is how we hope to arrive at datasets that faithfully embody tasks targeting realistic capabilities and that acknowledge the humanity of those represented within the data, in addition to those participating in the process of its creation. Such datasets will undoubtedly be more expensive to create, in time, money, and effort, but this is small price to pay for the consideration of the human lives at stake.

ACKNOWLEDGMENTS

A version of this paper appeared at the NeurIPS 2020 Workshop on Machine Learning Retrospectives, Surveys, and Meta-analyses (ML-RSA). We thank the anonymous reviewers for their comments and suggestions, as well as the organizers and reviewers at ML-RSA for feedback on an earlier version of this paper.

AUTHOR CONTRIBUTIONS

Conceptualization and project administration, A.P. All authors contributed to writing, editing, and investigation.

DECLARATION OF INTERESTS

I.D.R. serves on the advisory board *Patterns*. The authors declare no other competing interests.

REFERENCES

1. Halevy, A., Norvig, P., and Pereira, F. (2009). The unreasonable effectiveness of data. IEEE Intell. Syst. *24*, 8–12.

2. Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE International Conference on Computer Vision, pp. 843–852.

3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L. (2009). ImageNet: a large-scale hierarchical image database (CVPR).

4. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (Association for Computational Linguistics), pp. 353–355.

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

6. Raji, I.D., and Fried, G. (2021). About face: a survey of facial recognition evaluation. arXiv, arXiv:210200813.

7. Scheuerman, M.K., Denton, E., and Hanna, A. (2021). Do datasets have politics? Disciplinary values in computer vision dataset development. Computer Supported Cooperative Work (CSCW)).

8. boyd, d., and Crawford, K. (2012). Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inf. Commun. Soc. *15*, 662–679.

9. Schlangen, D. (2020). Targeting the benchmark: on methodology in current natural language processing research. ArXiv, abs/2007.04792.

10. Bowker, G.C. (2005). Memory Practices in the Sciences, *vol. 205* (Mit Press Cambridge).

11. Crawford, K. (2017). The trouble with bias. https://www.youtube.com/watch?v=fMym_BKWQzk; neurIPS keynote.

12. Buolamwini, J., and Gebru, T. (2018). Gender shades: intersectional accuracy disparities in commercial gender classification. In Proceedings of Machine Learning Research, *vol. 81*, S.A. Friedler and C. Wilson, eds. (PMLR), pp. 77–91.

13. Wilson, B., Hoffman, J., and Morgenstern, J. (2019). Predictive inequity in object detection. arXiv, arXiv:190211097.

14. DeVries, T., Misra, I., Wang, C., and van der Maaten, L. (2019). Does object recognition work for everyone? In IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019 (Computer Vision Foundation/IEEE), pp. 52–59. http://openaccess.thecvf.com/content_CVPRW_2019/html/cv4gc/de_Vries_Does_Object_Recognition_Work_for_Everyone_CVPRW_2019_paper.html.

15. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.W. (2018). Gender bias in coreference resolution: evaluation and debiasing methods. . Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, *Volume 2* (Association for Computational

Linguistics), pp. 15–20. Short Papers. https://doi.org/10.18653/v1/N18-2003. https://www.aclweb.org/anthology/N18-2003.

16. Lennon, J. (2020). If you're de-biasing the model, it's too late. https://scale.com/blog/if-youre-de-biasing-the-model-its-too-late.

17. Hoffmann, A.L. (2020). Terms of Inclusion: Data, Discourse, Violence (New media & society), 1461444820958725.

18. Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc. Natl. Acad. Sci. *115*, E3635–E3644.

19. Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., and Denuyl, S.C. (2020). Social biases in NLP models as barriers for persons with disabilities. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5491–5501.

20. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.W. (2017). Men also like shopping: reducing gender bias amplification using corpus-level constraints. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 2979–2989. https://doi.org/10.18653/v1/D17-1323. https://www.aclweb.org/anthology/D17-1323.

21. Burns, K., Hendricks, L.A., Darrell, T., and Rohrbach, A. (2018). Women Also Snowboard: Overcoming Bias in Captioning Models (ECCV).

22. van Miltenburg, E. (2016). Stereotyping and bias in the flickr30k dataset. In Proceedings of Multimodal Corpora: Computer Vision and Language Processing, *2016* (MMC), pp. 1–4.

23. Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 67–73.

24. Park, J.H., Shin, J., and Fung, P. (2018). Reducing gender bias in abusive language detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2799–2804.

25. Gehl, R.W., Moyer-Horner, L., and Yeo, S.K. (2017). Training computers to see internet pornography: gender and sexual discrimination in computer vision science. Television & New Media *18*, 529–547.

26. Crawford, K., Paglen, T., and Excavating, A.I. (2019). he politics of images in machine learning training sets. https://www.excavating.ai/.

27. Birhane, A., and Prabhu, V.U. (2021). Large image datasets: a pyrrhic win for computer vision? In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1537–1547.

28. Yang, K., Qinami, K., Fei-Fei, L., Deng, J., and Russakovsky, O. (2020). Towards fairer datasets: filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20 (Association for Computing Machinery), pp. 547–558, ISBN 9781450369367.

29. Torralba, A., Fergus, R., and Freeman, W.T. (2008). 80 million tiny images: a large data set for nonparametric object and scene recognition. IEEE Trans. Pattern Anal Mach Intell. *30*, 1958–1970. https://doi.org/10.1109/TPAMI.2008.128.

30. Levesque, H.J. (2014). On our best behaviour. Artif. Intelligence *212*, 27–35.

31. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., et al. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence *2* (11), 665–673.

32. Heinzerling, B. (2019). NLP's Clever Hans Moment Has Arrived (The Gradient).

33. Niven, T., and Kao, H.Y. (2019). Probing neural network comprehension of natural language arguments. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy (Association for Computational Linguistics), pp. 4658–4664. https://doi.org/10.18653/v1/P19-1459. https://www.aclweb.org/anthology/P19-1459.

34. Schuster, T., Shah, D., Yeo, Y.J.S., Ortiz, D.R.F., Santus, E., and Barzilay, R. (2019). Towards debiasing fact verification models. 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), pp. 3410–3416.

35. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N.A. (2018). 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) (Association for Computational Linguistics), pp. 107–112.

36. Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., and Van Durme, B. (2018). Hypothesis only baselines in natural language inference. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (Association for Computational Linguistics), pp. 180–191. https://doi.org/10.18653/v1/S18-2023. https://www.aclweb.org/anthology/S18-2023.

37. Kaushik, D., and Lipton, Z.C. (2018). How much reading does reading comprehension require? A critical investigation of popular benchmarks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 5010–5015. https://doi.org/10.18653/v1/D18-1546. https://www.aclweb.org/anthology/D18-1546.

38. Storks, S., Gao, Q., and Chai, J.Y. (2019). Recent advances in natural language inference: a survey of benchmarks, resources, and approaches. arXiv, arXiv:190401172.

39. Schlegel, V., Nenadic, G., and Batista-Navarro, R. (2020). Beyond leaderboards: a survey of methods for revealing weaknesses in natural language inference data and models. arXiv, arXiv:200514709.

40. Srivastava, M., Hashimoto, T., and Liang, P. (2020). Robustness to spurious correlations via human annotations. In International Conference on Machine Learning (ICML).

41. Gardner, M., Berant, J., Hajishirzi, H., Talmor, A., and Min, S. (2019). On making reading comprehension more comprehensive. In Proceedings of the 2nd Workshop on Machine Reading for Question Answering (Association for Computational Linguistics), pp. 105–112. https://doi.org/10.18653/v1/D19-5815. https://www.aclweb.org/anthology/D19-5815.

42. Agüera y Arcas, B., Todorov, A., and Mitchell, M. (2018). Do Algorithms Reveal Sexual Orientation or Just Expose Our Stereotypes? (Medium). https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477.

43. Gelman, A., Mattson, G., and Simpson, D. (2018). Gaydar and the fallacy of decontextualized measurement. Sociological Sci. *5*, 270–280. https://doi.org/10.15195/v5.a12.

44. Johannßen, D., Biemann, C., Remus, S., Baumann, T., and Sheffer, D. (2020). Germeval 2020 task 1 on the classification and regression of cognitive and emotional style from text: companion paper. . Proceedings of the 5th SwissText & 16th KONVENS Joint Conference, *vol. 2624*.

45. Bender, E.M. (2020). Is there research that shouldn't be done? Is there research that shouldn't be encouraged? mediumcom. https://medium.com/@emilymenonbender/is-there-research-that-shouldnt-be-done-is-there-research-that-shouldn-t-be-encouraged-b1bf7d321bb6.

46. Jacobsen, J.H., Geirhos, R., and Michaelis, C. (2020). Shortcuts: Neural Networks Love to Cheat (The Gradient).

47. Jo, E.S., and Gebru, T. (2020). Lessons from archives: strategies for collecting sociocultural data in machine learning. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20 (Association for Computing Machinery), pp. 306–316. ISBN 9781450369367. https://doi.org/10.1145/3351095.3372829.

48. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P.K., and Aroyo, L.M. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Association for Computing Machinery), pp. 1–15, 39.

49. Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., and Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Association for Computing Machinery), pp. 1–16.

50. Solon, O. (2019). Facial recognition's 'dirty little secret': millions of online photos scraped without consent. https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921.

51. Misra, I., Zitnick, C., Mitchell, M., and Girshick, R. (2016). Seeing through the human reporting bias: visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), pp. 2930–2939.

52. Ghai, B., Liao, Q.V., Zhang, Y., and Mueller, K. (2020). Measuring social biases of crowd workers using counterfactual queries. arXiv, arXiv:200402028.

53. Hube, C., Fetahu, B., and Gadiraju, U. (2019). Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19 (Association for Computing Machinery), pp. 1–12. ISBN 9781450359702. https://doi.org/10.1145/3290605.3300637.

54. Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N.A. (2019). The risk of racial bias in hate speech detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 1668–1678. https://www.aclweb.org/anthology/P19-1163.

55. Miceli, M., Schuessler, M., and Yang, T. (2020). Between subjectivity and imposition: power dynamics in data annotation for computer vision. Proc. ACM Hum-comput Interact *4* (CSCW2). https://doi.org/10.1145/3415186.

56. Aroyo, L., and Welty, C. (2015). Truth is a lie: crowd truth and the seven myths of human annotation. AI Mag. *36*, 15–24. https://doi.org/10.1609/aimag.v36i1.2564. https://aaai.org/ojs/index.php/aimagazine/article/view/2564.

57. Geva, M., Goldberg, Y., and Berant, J. (2019). Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. 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) (Association for Computational Linguistics), pp. 1161–1166. https://doi.org/10.18653/v1/D19-1107. https://www.aclweb.org/anthology/D19-1107.

58. Sen, S., Giesel, M.E., Gold, R., Hillmann, B., Lesicko, M., Naden, S., et al. (2015). Turkers, scholars, "Arafat" and "peace": cultural communities and algorithmic gold standards. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. CSCW '15 (Association for Computing Machinery), pp. 826–838, ISBN 9781450329224.

59. Tsipras, D., Santurkar, S., Engstrom, L., Ilyas, A., and Madry, A. (2020). From imagenet to image classification: contextualizing progress on benchmarks. In International Conference on Machine Learning (ICML), pp. 9625–9635.

60. Geiger, R.S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., et al. (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20 (Association for Computing Machinery), pp. 325–336. ISBN 9781450369367. https://doi.org/10.1145/3351095.3372862.

61. Scheuerman, M.K., Wade, K., Lustig, C., and Brubaker, J.R. (2020). How we've taught algorithms to see identity: constructing race and gender in image databases for facial analysis. Proc. ACM Hum-comput Interact *4* (CSCW1). https://doi.org/10.1145/3392866.

62. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé, H., III, et al. (2018). Datasheets for datasets. arXiv, arXiv:180309010.

63. Bender, E.M., and Friedman, B. (2018). Data statements for natural language processing: toward mitigating system bias and enabling better science. Trans. Assoc. Comput. Linguistics *6*, 587–604. https://doi.org/10.1162/tacl_a_00041. https://www.aclweb.org/anthology/Q18-1041.

64. Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. (2018). The dataset nutrition label: a framework to drive higher data quality standards. arXiv, arXiv:180503677.

65. Chmielinski, K.S., Newman, S., Taylor, M., Joseph, J., Thomas, K., Yurkofsky, J., et al. (2020). The dataset nutrition label (2nd gen): leveraging context to mitigate harms in artificial intelligence, neurIPS Workshop on Dataset Curation and Security http://securedata.lol/.

66. Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., et al. (2021). Towards accountability for machine learning datasets: practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery), pp. 560–575.

67. Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do ImageNet classifiers generalize to ImageNet? In Proceedings of Machine Learning Research, *vol. 97*, K. Chaudhuri and R. Salakhutdinov, eds. (PMLR), pp. 5389–5400.

68. Ananny, M., and Crawford, K. (2018). Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. New media Soc. *20*, 973–989.

69. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.. On the dangers of stochastic parrots: can language models be too big?. In: Proc. FAccT 2021. 2021,.

70. Raji, D., Hoffmann, A.L., Moorosi, N., Prabhu, V., Metcalf, J., and Stanley, S. (2020). Panel discussion: harms from ai research, neurIPS Workshop on Navigating the Broader Impacts of AI Research https://nbiair.com/.

71. Pipkin, E. (2020). On Lacework: Watching an Entire Machine-Learning Dataset (unthinkingphotography).

72. Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, T., et al. (2019). Moments in time dataset: one million videos for event understanding. IEEE Trans. Pattern Anal. Machine Intelligence, 1–8. https://doi.org/10.1109/TPAMI.2019.2901464.

73. Caswell, I., Kreutzer, J., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., et al. (2021). Quality at a glance: an audit of web-crawled multilingual datasets. arXiv, arXiv:210312028.

74. Sakaguchi, K., Bras, R.L., Bhagavatula, C., and Choi, Y. (2020). Winogrande: An Adversarial Winograd Schema Challenge at Scale (AAAI).

75. Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M.E., Sabharwal, A., et al. (2020). Adversarial filters of dataset biases. In International Conference on Machine Learning (PMLR), pp. 1078–1088.

76. Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N.A., et al. (2020). Dataset cartography: mapping and diagnosing datasets with training dynamics. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (Association for Computational Linguistics), pp. 9275–9293.

77. Northcutt, C.G., Athalye, A., and Mueller, J. (2021). Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks (ICLR).

78. Han, X., Wallace, B.C., and Tsvetkov, Y. (2020). Explaining black box predictions and unveiling data artifacts through influence functions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5553–5563.

79. Koh, P.W., and Liang, P. (2017). Understanding black-box predictions via influence functions. In International Conference on Machine Learning (PMLR), pp. 1885–1894.

80. Wang, A., Narayanan, A., and Russakovsky, O. (2020). REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets (European Conference on Computer Vision (ECCV)).

81. Liu, N.F., Schwartz, R., and Smith, N.A. (2019). Inoculation by fine-tuning: a method for analyzing challenge datasets. . Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, *Volume 1* (Association for Computational Linguistics), pp. 2171–2179, Long and Short Papers. https://www.aclweb.org/anthology/N19-1225.

82. Gardner, M., Artzi, Y., Basmov, V., Berant, J., Bogin, B., Chen, S., et al. (2020). Evaluating models' local decision boundaries via contrast sets. In Findings of the Association for Computational Linguistics: EMNLP

2020 (Association for Computational Linguistics), pp. 1307–1323. https://www.aclweb.org/anthology/2020.findings-emnlp.117.

83. Kaushik, D., Hovy, E., and Lipton, Z.C. (2020). Learning the Difference that Makes a Difference with Counterfactually-Augmented Data (ICLR).

84. Teney, D., Kafle, K., Shrestha, R., Abbasnejad, E., Kanan, C., and Hengel, A.v.d. (2020). On the value of out-of-distribution testing: an example of Goodhart's law. arXiv, arXiv:200509241.

85. Teney, D., Abbasnedjad, E., and Hengel, A.v.d. (2020). Learning what makes a difference from counterfactual examples and gradient supervision. In European Conference on Computer Vision.

86. He, H., Zha, S., and Wang, H. (2019). Unlearn dataset bias in natural language inference by fitting the residual. In Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019) (Association for Computational Linguistics), pp. 132–142. https://doi.org/10.18653/v1/D19-6115. https://www.aclweb.org/anthology/D19-6115.

87. Pavlick, E., and Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. Trans. Assoc. Comput. Linguistics *7*, 677–694.

88. Khani, F., and Liang, P. (2021). Removing spurious features can hurt accuracy and affect groups disproportionately. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery), pp. 196–205.

89. Denton, E.L., Hanna, A., Amironesei, R., Smart, A., Nicole, H., and Scheuerman, M.K. (2020). Bringing the people back in: contesting benchmark machine learning datasets (arXiv preprint arXiv:2007.07399).

90. Ọnụọha, M. (2016). The Point of Collection (Points). https://points.datasociety.net/the-point-of-collection-8ee44ad7c2fa.

91. Simonite, T. (2018). Google's AI guru wants computers to think more like brains. https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/.

92. Ethayarajh, K., and Jurafsky, D. (2020). Utility is in the eye of the user: a critique of NLP leaderboards. arXiv, arXiv:2009.13888.

93. Dodge, J., Gururangan, S., Card, D., Schwartz, R., and Smith, N.A. (2019). 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) (Association for Computational Linguistics), pp. 2185–2194. https://doi.org/10.18653/v1/D19-1224. https://www.aclweb.org/anthology/D19-1224.

94. Schwartz, R., Dodge, J., Smith, N.A., and Etzioni, O. (2019). Green AI. arXiv http://arxiv.org/abs/1907.10597.

95. Sculley, D., Snoek, J., Wiltschko, A.B., and Rahimi, A. (2018). Winner's Curse? on Pace, Progress, and Empirical Rigor (ICLR).

96. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., et al. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19 (Association for Computing Machinery), pp. 220–229. ISBN 9781450361255. https://doi.org/10.1145/3287560.3287596.

97. Bender, E.M., and Koller, A. (2020). 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 (Association for Computational Linguistics), pp. 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463. https://www.aclweb.org/anthology/2020.acl-main.463.

98. Porter, T. (1995). Trust in Numbers: The Pursuit of Objectivity in Science and Public Life (Princeton University Press).

99. Houser, J., Desrosières, A., and Naish, C. (1999). The politics of large numbers: a history of statistical reasoning. Contemp. Sociol. *28*, 361.

100. Koopman, C. (2019). How We Became Our Data: A Genealogy of the Informational Person (University of Chicago Press).

101. Jacobs, A.Z., and Wallach, H. (2021). Measurement and fairness. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21 (Association for Computing Machinery), pp. 375–385. ISBN 9781450383097. https://doi.org/10.1145/3442188.3445901.

102. Richards, N.M., and King, J.H. (2014). Big data ethics. Wake For. L Rev *49*, 393.

103. Metcalf, J., and Crawford, K. (2016). Where are human subjects in big data research? The emerging ethics divide. Big Data Soc *3*, 2053951716650211.

104. Mohamed, S., Png, M.T., and Isaac, W. (2020). Decolonial AI: decolonial theory as sociotechnical foresight in artificial intelligence. Philos. Technology, 1–26.

105. Birhane, A. (2020). Algorithmic colonization of africa. SCRIPTed *17*, 389.

106. Harvey, A., and LaPlace, J. (2019). MegaPixels: origins and endpoints of biometric datasets "in the wild". https://megapixels.cc.

107. Solove, D.J. (2007). 'I've got nothing to hide' and other misunderstandings of privacy. San Diego L. Rev. *44* (4), 745–772. https://papers.ssrn.com/soL3/papers.cfm?abstract_id=998565.

108. Peng, K. (2020). Facial Recognition Datasets Are Being Widely Used Despite Being Taken Down Due to Ethical Concerns. Here's How (Freedom to Tinker).

109. O'Sullivan, L. (2020). Don't steal data, neurIPS Workshop on Dataset Curation and Security http://securedata.lol/.

110. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., et al. (2020). Extracting training data from large language models. arXiv, arXiv:201207805.

111. Vidgen, B., and Derczynski, L. (2020). Directions in abusive language training data, a systematic review: garbage in, garbage out. Plos one *15*, e0243300.

112. Stodden, V., and Miguez, S. (2014). Best practices for computational science: software infrastructure and environments for reproducible and extensible research. J. Open Res. Softw. *2*, e21. https://doi.org/10.5334/jors.ay.

113. Stodden, V. (2020). The data science life cycle: a disciplined approach to advancing data science as a science. Commun. ACM *63*, 58–66. https://doi.org/10.1145/3360646.

114. Pasquetto, I.V., Randles, B.M., and Borgman, C.L. (2017). On the reuse of scientific data. Data Sci. J. *16*, 8. https://doi.org/10.5334/dsj-2017-008.

115. Belz, A., and Kilgarriff, A. (2006). Shared-task evaluations in HLT: lessons for NLG. In Proceedings of the Fourth International Natural Language Generation Conference (Association for Computational Linguistics), pp. 133–135. https://www.aclweb.org/anthology/W06-1421.

116. Edwards, P.N. (2013). A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. Infrastructures Series, first paperback edition (The MIT Press), ISBN 978-0-262-51863-5 978-0-262-01392-5.

117. Strasser, B.J., and Edwards, P.N. (2017). Big data is the answer … but what is the question? Osiris *32*, 328–345. https://doi.org/10.1086/694223.

118. Radin, J. (2017). "Digital Natives": how medical and indigenous histories matter for big data. Osiris *32*, 43–64. https://doi.org/10.1086/693853.

119. Murgia, M. (2019). Who's Using Your Face? The Ugly Truth about Facial Recognition (Financial Times).

120. Irani, L. (2015a). The Cultural Work of Microwork*17* (New Media & Society), pp. 720–739.

121. Suchman, L. (1995). Making work visible. Commun. ACM *38*, 56–64. https://doi.org/10.1145/223248.223263.

122. Star, S.L., and Strauss, A. (1999). Layers of silence, arenas of voice: the ecology of visible and invisible work. Computer Supported Coop. Work (Cscw) *8*, 9–30.

123. Precarity Lab (2020). Technoprecarious (Goldsmiths Press).

124. Irani, L.C., and Silberman, M.S. (2013). Turkopticon: interrupting worker invisibility in Amazon Mechanical Turk. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '13 (Association for Computing Machinery), pp. 611–620, ISBN 9781450318990.

125. Irani, L. (2015b). Difference and dependence among digital workers: the case of Amazon Mechanical Turk. South Atlantic Q. *114*, 225–234.

126. Berg, J. (2016). Income Security in the On-Demand Economy: Findings and Policy Lessons from a Survey of Crowdworkers. ILO Working Papers (International Labour Organization).

127. Semuels, A. (2018). The Internet Is Enabling a New Kind of Poorly Paid Hell (The Atlantic). https://www.theatlantic.com/business/archive/2018/01/amazon-mechanical-turk/551192/.

128. Gray, M.L., Suri, S., and Ghost, Work (2019). How to Stop Silicon Valley from Building a New Global Underclass (Houghton Mifflin Harcourt).

129. Silberman, M.S., Tomlinson, B., LaPlante, R., Ross, J., Irani, L., and Zaldivar, A. (2018). Responsible research with crowds: pay crowdworkers at least minimum wage. Commun. ACM *61*, 39–41. https://doi.org/10.1145/3180492.

130. Whiting, M.E., Hugh, G., and Bernstein, M.S. (2019). Fair work: crowd work minimum wage with one line of code. Proc. AAAI Conf. Hum. Comput. Crowdsourcing *7*, 197–206.

131. Salehi, N., Irani, L.C., Bernstein, M.S., Alkhatib, A., Ogbe, E., Milland, K., et al. (2015). We are dynamo: overcoming stalling and friction in collective action for crowd workers. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI '15 (Association for Computing Machinery), pp. 1621–1630, ISBN 9781450331456.

132. Callison-Burch, C. (2014). Crowd-workers: Aggregating Information across Turkers to Help Them Find Higher Paying Work (HCOMP).

133. Viljoen S. Democratic data: a relational theory for data governance. Forthcoming, Yale Law Journal. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3727562.

134. Posner, E.A., and Weyl, E.G. (2019). Radical Markets (Princeton University Press).

135. Vincent, N., Hecht, B., and Sen, S. (2019). "Data strikes": evaluating the effectiveness of a new form of collective action against technology companies. In The World Wide Web Conference, pp. 1931–1943.

136. Benjamin, M., Gagnon, P., Rostamzadeh, N., Pal, C., Bengio, Y., and Shee, A. (2019). Towards standardization of data licenses: the Montreal data license. arXiv, arXiv:1903.12262.

137. Khan, M., Hanna, A.. The Legality of Computer Vision Datasets. Under review 2020.

138. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. Int. J. Computer Vis. *115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y.

139. Levendowski, A. (2018). How copyright law can fix artificial intelligence's implicit bias problem. Wash L. Rev. *93*, 579.

140. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., and Zisserman, A. (2010). The Pascal visual object classes (VOC) challenge. Int. J. Computer Vis. *88*, 303–338. https://doi.org/10.1007/s11263-009-0275-4.

141. Merkely, R.. Use and Fair Use: Statement on Shared Images in Facial Recognition AI. 2019.

142. Berg, T., Berg, A., Edwards, J., Maire, M., White, R., Teh, Yee-Whye, et al. (2004). Names and faces in the news. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004, *vol. 2* (IEEE), pp. 848–854. ISBN 978-0-7695-2158-9. https://doi.org/10.1109/CVPR.2004.1315253.

143. Sag, M. (2019). The new legal landscape for text mining and machine learning. Journal of the Copyright Society of the USA *66*, 291–367. https://doi.org/10.2139/SSRN.3331606.

144. Caliskan, A., Bryson, J.J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science *356*, 183–186. https://doi.org/10.1126/science.aal4230.

145. Packer, B., Mitchell, M., Guajardo-Céspedes, M., and Halpern, Y. (2018). Text embeddings contain bias. Here's why that matters. Google Developers https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html.

146. Seaver, N. (2017). Algorithms as culture: some tactics for the ethnography of algorithmic systems. Big Data Soc. *4* (2). 2053951717738104. https://doi.org/10.1177/2053951717738104.

147. Selbst, A.D., boyd, d., Friedler, S.A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 59–68.

148. Bambara, T.C. (1970). On the issue of roles. The Black Woman: An Anthology, 101–110.

149. Raji, I.D. (2020). The discomfort of death counts: mourning through the distorted lens of reported COVID-19 death data. Patterns *1*, 100066.

150. Register, Y., and Ko, A.J. (2020). Learning machine learning with personal data helps stakeholders ground advocacy arguments in model mechanics. In Proceedings of the 2020 ACM Conference on International Computing Education Research, pp. 67–78.

151. Vanschoren, J., and Yeung, S. (2021). Announcing the NeurIPS 2021 datasets and benchmarks track. https://neuripsconf.medium.com/announcing-the-neurips-2021-datasets-and-benchmarks-track-644e27c1e66c.