Special Issue: Automating Communication

new media & society
2024, Vol. 26(9) 4955–4972
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14614448241251797
journals.sagepub.com/home/nms
Sage

The social construction of datasets: On the practices, processes, and challenges of dataset creation for machine learning

Will Orr
University of Southern California, USA; Microsoft Research New York City, USA

Kate Crawford
University of Southern California, USA; Microsoft Research New York City, USA

Abstract

Despite the critical role that datasets play in how systems make predictions and interpret the world, the dynamics of their construction are not well understood. Drawing on a corpus of interviews with dataset creators, we uncover the messy and contingent realities of dataset preparation. We identify four key challenges in constructing datasets, including balancing the benefits and costs of increasing dataset *scale*, limited access to *resources*, a reliance on *shortcuts* for compiling datasets and evaluating their quality, and ambivalence regarding *accountability* for a dataset. These themes illustrate the ways in which datasets are not objective or neutral but reflect the personal judgments and trade-offs of their creators within wider institutional dynamics, working within social, technical, and organizational constraints. We underscore the importance of examining the processes of dataset creation to strengthen an understanding of responsible practices for dataset development and care.

Keywords

Accountability, artificial intelligence, datasets, design, machine learning, maintenance

> *Right now, in machine learning, what drives progress is data. Really more than algorithms, is data*. [Interview with Cornebise, WorldStrat]

Corresponding author:
Will Orr, Annenberg School for Communication, University of Southern California, Los Angeles, CA 90007, USA.
Email: orrw@usc.edu

Introduction

How are datasets for machine learning (ML) developed? What are the practices around their construction, and how does that influence the way that ML works today? These are surprisingly difficult questions to answer. While training datasets are the foundation upon which ML models are built, there has been little examination of the practices of dataset production. In our research, we aim to develop a richer understanding of how datasets are made and distributed through in-depth interviews with the creators of some of the most influential datasets in contemporary ML. We contribute to the growing body of literature in critical dataset studies (Thylstrup, 2022) by tracing the technical processes, the social dynamics, and the maintenance and obsolescence of datasets from the perspectives of their creators to uncover the constraints that shape their formation. We hope that this contributes to a nuanced understanding of this essential yet often undervalued aspect of ML research (Sambasivan et al., 2021).

Training datasets form the backbone of ML research and constitute the "epistemic boundaries" of ML models (Crawford and Paglen, 2019). They provide models with a set of "ground truth" examples to which all other predictions are made and compared. Training datasets, therefore, not only shape the possibilities of ML models, they also constitute their bedrock of claims to truth and accuracy. A subset of datasets is also used for evaluating the performance of ML models. Benchmark datasets represent certain tasks or technical challenges and are routinized for comparing, replicating, and reproducing model results (Raji et al., 2021). Benchmark datasets, in this sense, have a normative impact on ML models. They not only set the problems that are deemed worth solving by the community, but they also constitute the very metrics of success. For a model to be considered state-of-the-art, it must correctly predict the outputs of popular benchmarks. So they have enormous power to determine what "good performance" means.

As Cornebise underscores, ML research has been driven by the size and diversity of the datasets used to train and evaluate them. Today, "internet-sized datasets" (Torralba et al., 2008: 11) are the dominant training corpora for large models. For instance, the Common Crawl dataset consists of 200–300 TB of scraped text content released every month (Luccioni and Viviano, 2021). GPT-3 (Brown et al., 2020) incorporates 41 months of content as just one aspect of their training corpus. Increasing the scale of datasets has become closely associated with improving model performance (Hoffmann et al., 2022). But the current emphasis on scale has also brought with it various technical, legal, and ethical pitfalls.

Drawing on the traditions of laboratory studies from Science and Technology Studies (Latour, 2015), we trace the social production of datasets by their creators, focusing on four aspects of their creation: *scale*, *resources*, *shortcuts*, and *accountability*. Rather than analyzing ready-made datasets, we analyze datasets-in-the-making where "context and content fuse together" (Latour, 2015: 6). This is particularly important as technical narratives about ML as an "objective" form of statistical representation fail to account for the personal decisions and institutional forces that shape training datasets (Crawford, 2021). Creators' motivations, practices, and constraints invariably impact the form and substance of datasets which then go on to shape the models used by the ML community. While ML models may be opaque and inscrutable, understanding the datasets that are used to train and evaluate them can provide insights into their outputs and their political and epistemological implications (Crawford and Paglen, 2019). Thus, detailing the cultures through which datasets are created and deployed is an integral step toward better understanding the social practices that underlie ML as well as the challenges and potential paths forward.

To carry out this research, we interviewed 18 dataset creators via teleconference between July and September 2022. Participants were determined by assessing the most cited contemporary datasets, then contacted through personal contacts and snowball sampling. Our sample is diverse but by no means representative. Two-thirds of interviews were conducted with creators in the United States, with the remainder spanning the United Kingdom, Europe, and Australia. Most participants were located at universities and nonprofit organizations. A minority of participants were employed at technology companies. Generally, private sector employees were reluctant to be interviewed. This underscores the secrecy regarding in-house dataset creation and its perception as commercial in confidence. Consequently, detailing the experiences of creators in large technology companies remains an avenue for future research.

We recruited participants via email invitation, and our interviews traced datasets through their origins, usages, maintenance, and eventual obsolescence. Datasets ranged from large, crawled corpora, to natural language processing (NLP) benchmarks, personal recommendation, emotion detection, and action recognition. Data collection methods used by creators were themselves diverse: some creators relied on existing datasets, others hired crowd workers and students to generate data, or conducted scraping on specific sites and the accessible web. While some participants contributed to many datasets reflected in the findings, specifically, the datasets discussed were SQuAD 2.0, GLUE, SuperGLUE, WiC, IEMOCAP, YFCC100M, Common Crawl, C4, LAION-5B, Amazon Reviews, MovieLens, WorldStrat, TweetEval, WinoGrad, WinoGrande, UCF101, Taskonomy, and IKEA Assembly. To allow for proper attribution and anonymity, dataset creators were given the opportunity to be presented within research outputs by their name, by the datasets that they created, or to remain anonymous.

We use the term "dataset creator" to denote individuals who had substantive first-hand experience producing datasets. Participants were all directly involved in the creation of their respective datasets in both junior and senior capacities. Many of these datasets were produced in teams, with participants contributing to certain aspects, or overseeing the whole process. Most of our participants worked on multiple datasets, and their experiences from other projects often informed their responses. Interviews averaged approximately 1 hour long, ranging from 40 minutes to 1.5 hours. Interviews were transcribed and thematically coded iteratively allowing new themes to emerge as interviews progressed. This process produced four major themes that were present across all interviews and contexts. Checks on coding were done by both authors to hone the conceptualization of these themes and to ensure reliability. Interview data were supplemented with dataset documentation, such as description files, licenses, and publications. These approaches allowed for the social, technical, and organizational dynamics of dataset creation to emerge.

In this article, we address four central concerns that our respondents raised about the development of datasets: navigating the promises and costs of *scale*, access to *resources, shortcuts* and workarounds in creating datasets and assessing their quality, and attributions of *accountability* for a dataset and its impacts. As we will see, creators grapple with the social, structural, and capital constraints these challenges impose upon the datasets and their original objectives, frequently making complex trade-offs that shape datasets and their social impacts in the process.

Dataset creation as a practice: the existing literature

Dataset construction depends on substantial amounts of human conceptualization, judgment, and values, particularly in the collection, curation, annotation, and preparation of data (Crawford, 2021; Jaton, 2021; Plantin, 2019). While ML is often represented as an objective form of statistical processes, the process of compiling training data is based on subjective determinations, and is riddled with contingencies, indeterminacies, and assumptions (Kang, 2023). Yet as Orr and Davis (2020: 7) show, AI practitioners are "extrinsically bound" and must negotiate conflicting stakeholder goals and values. For Jaton (2021), creators face uncertain decisions that materially shape the impacts of a dataset. There are no standard practices for resolving these tensions. Practitioners rely on ad hoc and pragmatic solutions to negotiate competing priorities and constraints (Jaton, 2021; Orr and Davis, 2020).

The lack of standard practices for creating and circulating datasets has significant impacts. For instance, without standardized mechanisms for maintaining and retiring datasets and informing the community, datasets remain in circulation and use well after they are perceived to be out of date (Luccioni et al., 2022). All datasets are eventually viewed by their creators—and the wider ML field—as obsolete. They may be perceived as "solved" as models achieve high prediction scores (e.g. Baldominos et al., 2019). They may also contain information that is perceived to be outdated. For example, models trained on datasets created before COVID-19 cannot provide information regarding the pandemic, such as detecting faces in masks. Sometimes, datasets may also be retracted by dataset creators (e.g. Torralba et al., 2008). Despite the prevalence (and inevitability) of dataset obsolescence, datasets often remain freely available online through official and third-party sources and continue to be used for years.[1]

Despite the importance of this labor in shaping the perspectives embedded in datasets, data work is often devalued, erased, and hidden from view (Plantin, 2019; Sambasivan et al., 2021). As Plantin (2019) argues, the standardization of datasets into accepted formats and the centralization of these artifacts in repositories for distribution erase the labor and uncertainties of creating datasets. Ready-made datasets, as the infrastructural scaffolding for ML models, are often presented as objective representations of reality and neutral technical substrates. But this relies on a "form of collective forgetting, or naturalization, of the contingent, messy work" of their making (Bowker and Star, 1999: 299). Moreover, these messy processes of constructing datasets are rarely communicated within research articles (Geiger et al., 2020).

This erasure has consequences: it reinforces the flawed notion that ML is separate from personal judgments and value-based decisions, and it reinscribes the misperception that dataset construction is immaterial for the effectiveness of an ML model. We can understand this as part of a much longer tradition in the sciences where scientists are trained to hide subjective decisions in their work, which merely masks them, "making them unexaminable by others” (Douglas, 2004: 459). This is why examining the social, technical, and institutional processes that underlie dataset creation is essential if we are to understand the values that drive their production. In the following sections, we highlight four significant challenges in the construction of their datasets: the pressures of *scale*, the limitations of *resources*, the reliance on *shortcuts*, and the allocation of *accountability*. As we will see, these themes are interrelated and often inform each other.

Scale

> *As machine learning people, we wanted to have everything.* [Interview, anon.,YFCC]

Scale is a primary objective of dataset construction. In the early 2010s, datasets grew to a scale that was previously never attempted. Containing millions of examples, large-scale datasets “completely changed what was possible with deep learning,” leading to state-of-the-art models and better techniques (Goodfellow et al., 2016: 21). The abundance of data is seen to reduce the necessary skill and model complexity for achieving high performance (Hoffmann et al., 2022). Large-scaled datasets are seen as more objective as they are assumed to minimize human choices: rather than choosing a particular set of cats, a dataset includes as many cats as possible from online sources. But even in this case, humans make choices—how many cats are enough? How much care is given to assessing that they are correctly labeled or in which contexts they appear?

The promises of scale

Recent literature in natural language processing (e.g., Brown et al., 2020) and image-text multimodal models (e.g., Ramesh et al., 2021) argues that increasing the scale of training datasets has intrinsic benefits for model performance. Our interviews with creators supported these claims, with one explaining, “just by increasing the scale, you force the system to develop capabilities to solve new tasks” [Interview with Jitsev, LAION-5B]. Increasing the scale of a dataset thus became the default starting point for many, almost as an instinctive presumption, as one creator explained: “I just kind of thought . . . what’s the biggest possible dataset of that category that we could possibly collect?” [Interview with McAuley, Amazon Reviews].

Creators also presented scale as a technological alternative to data cleaning, filtering, and curation. As Jaton (2021) signals, while the phenomena in the world they represent continue to change and develop, datasets provide *a view into the online past*: a snapshot of reality that is situated in the geographies and temporalities of their creation. Creators often identified erroneous, low-quality, or harmful content after the dataset had already been released. In this way, datasets are always out of date, in need of revisions, or “broken” (Pink et al., 2018). As Chun (2021) underscores, relying on temporally situated data can only ever reproduce the injustices and inequities of the past into future systems. Creators are well aware of these limitations. Despite the immense scale of some datasets, several participants recognized that they were by no means complete nor representative. One participant explained that “it is not ground truth data. It’s just this big noisy mess” [Interview with McAuley, Amazon Reviews].

Increasing the scale of a dataset also heightens the probability of including "garbage data" [Interview with Schuhmann, LAION-5B], such as blurry images, duplicate pages, and harmful content. However, echoing the "scale beats noise" discourse of large-scale datasets (Birhane et al., 2021), our participants frequently suggested that low-quality data could be drowned out by scale. Some creators also perceived that increasing the scale of their dataset would mitigate the socially harmful content present within their dataset. Scale was thus presented as not only a technological solution but also of intrinsic good in its own right. Yet, as Hanna and Park (2020) have argued, this kind of focus on scale will deepen existing social inequities.

The limitations of scale

The perceived benefits of scale have led to increased pressures from the research community to collect more data and release larger datasets. Even creators of the largest publicly available datasets recounted criticisms that their creations are "not big enough, or not complete enough" [Interview with Nagel, Common Crawl]. For some, these pressures manifested in dataset users' desire to have increasingly intimate information about data subjects: "People are always asking for more and more stuff about the [data subjects] . . . Who are they? We don't know. And boy as someone who's developing personalization technology, it'd be really great to know more about them" [Interview with Harper, MovieLens]. The appetite for larger, more complete, or comprehensive datasets was coming at the cost of the data subjects' personal information as well as consistent data quality.

Furthermore, making large-scale datasets comes with significant challenges. Some creators expressed the importance of caring for their creations, but due to the sheer scale of current datasets, they didn't feel able to validate that the data represents what is intended, or discern underlying patterns, relationships, or trends within it. One creator explained with concern, "At some point, the dataset becomes too large to audit, especially via manual means" [Interview, anon., C4]. Creators could not view every picture, read every article or entirely know what is contained within their own dataset. They rarely looked at the data they pulled into training sets. Here, the practical limits of transparency resulted in creators feeling at 'arm's length' from a sense of accountability. Moreover, the scale of a dataset led creators to instead focus on concerns within future iterations and projects, rather than assessing and maintaining released datasets. As one participant explained:

> Running a crawler, you are frustrated often. Because the data quality is less than expected, or less good than you would like to have it. But it's difficult to improve it, especially if you want to do it for future crawls and not just by filtering and cleaning previous crawls. . . We are more in a sense looking forward, we do not care about all the data, and try to improve the upcoming datasets. [Interview with Nagel, Common Crawl]

While datasets-of-the-future are imagined as more carefully constructed, flawed datasets-of-the-past remain circulated and used, to be reproduced in perpetuity (Chun, 2021). Creators of large datasets expressed their frustration about the inability to adequately manage or understand the contents of their dataset. Although recent literature advocates for practices of care regarding the development, maintenance, distribution, and deprecation of datasets (Luccioni et al., 2022), traditional care practices may be frustrated by the increasing imperative for scale. As highlighted by Seaver (2021), while care and scale continue to be considered to be contradictory goals, they will be treated as such. This points toward a need to reimagine practices of both care and scale, not as opposed but decorrelated and simultaneously possible.

Resources

> *It's all a matter of cost-benefit . . . We were just kind of like, "Okay, let's just like figure out the fastest way to satisfy these constraints."* [Interview, anon., C4]

A strong theme that emerged from participants' accounts of making their datasets was the paramount importance of resources: specifically, money, compute, data, and human labor. These resources were needed to make a large-scale dataset, as well as to functionally train a model. Given the resource costs of constructing datasets, creators were often motivated to ensure that their datasets were publicly accessible to assist the broader research community. However, just making a dataset public does not ensure that it is usable. Training large models requires immense energy and labor costs that limit who is able to participate (Crawford, 2021; Koch et al., 2021). Here, we will highlight how creators negotiated these tensions.

The costs of dataset creation

"You're under time and budget pressure, and we didn't want to throw out data that we'd already collected. So, we collected all of the data rather than go for gender diversity" [Interview with Gould, IKEA]. This participant's response is paradigmatic of the kinds of trade-offs and concessions that creators must make in order to satisfy the resource constraints of their work. While gender parity was desired by the dataset creator, resource restrictions entailed settling for a dataset that overrepresented male participants. Our participants all mentioned the pressure on resources such as time, money, and compute in the construction of datasets. Crawlers required immense computational demands while employing crowd workers or students necessitated a non-trivial financial investment.

Furthermore, creators mentioned multiple kinds of strict time constraints such as conference submission deadlines, funding pressures, corporate schedules, and fears that rival academics may be working on a similar dataset. Thus, financial, computational, and organizational constraints ultimately shaped the characteristics of the datasets they created and released. It also underscores why datasets increasingly originate from elite, well-funded institutions that traditionally offer more support (Koch et al., 2021). While our participants did represent a diversity of institutions, they all reiterated the significant resource requirements to produce and maintain datasets.

Labor demands

Datasets were also constrained by the demands on creators' time. In most cases, participants largely acted in very small teams, which reflects findings that data work is not seen as high-status or desirable (Sambasivan et al., 2021). Limited human resources also shaped the datasets as creators had to decide the most important tasks that needed to be undertaken. The lack of personnel intersected with the time pressures, forcing dataset creators to 'make do,' as one creator explained: "I have to be very lazy, just as an engineer. Do only tasks which, let's say, just could be done in a short period of time, because I cannot just focus on a single task" [Interview with Nagel, Common Crawl]. In this case, the creator was the only person technically capable of collecting and curating the dataset for release. Larger problems that required structural solutions, further strategizing or intensive time investment were forsaken for easier solutions to smaller problems. These tensions were further evident in the cleaning and curation of datasets. The limited availability of time, money, compute, and labor meant that datasets exist as a compromise between the original intentions of creators and their situated resource constraints. Datasets were never perfect but only ever "good enough" [Interview with Jitsev, LAION-5B]. And even this was understood as very time-limited, until the next bigger dataset was released.

Accessibility

Given the immense resource requirements to create datasets, making datasets available to the public was often a great motivator for those who could afford it. Creators were very sensitive to the problem that large corporations have held a monopoly on data for state-of-the-art ML applications. This is evident through popular large models such as DALL-E 2 (Ramesh et al., 2021), and GPT-3 (Brown et al., 2020) being trained on proprietary datasets that were not released to the public. Making datasets available to the broader research community was therefore viewed as an important motivation for dataset development, as one creator explained:

> [Large corporations] were able to drive a lot of innovation based on this, on the availability of the data. And the idea was just to give more people, I mean researchers, students, startups with less resources also the opportunity to work with large amounts of web data, without the need to run a web crawler by themselves. [Interview with Nagel, Common Crawl]

The accessibility of datasets also gains symbolic meaning: creators perceive that by making datasets accessible, they can "begin to democratize this study on large-scale models across the broad community" [Interview with Jitsev, LAION-5B]. Here, the accessible dataset is seen as a resource for the community, to empower community researchers to contribute to the trajectory of ML research that has previously been controlled by corporate interests and proprietary technology. However, access alone is insufficient for truly "democratizing" AI use, development, profits and governance (Seger et al., 2023).

Indeed, just making a dataset public does not make it usable: creators noted the significant computing demands to train models with their datasets. These demands of computational intensity are particularly extensive for creators of image, video, and multimodal datasets. For some participants, this meant relying on expensive and energy-intensive supercomputer infrastructure to use their own datasets. Even the work of downloading some datasets is not trivial. Some creators faced computation and site restrictions when attempting to access their dataset online. One creator outlined the challenges of downloading the images from their dataset onto their computer:

> We crashed some file systems with the data, because usually large-scale, high-performance computer systems are built in a way that they were working very well on few files that are large. We had millions of tiny files, not every system is able to do all of that. [Interview, anon., YFCC]

This case highlights the resource and infrastructural barriers faced by users in downloading and making use of ML datasets. These challenges are only exacerbated by scale, as larger datasets require greater resources to process them, and are prone to infrastructural barriers such as human verification steps. Thus, the benefits and dividends of accessible datasets flow not to the "broad community," in the words of one participant [Interview with Jitsev, LAION-5B], but to those with resource requirements to adequately make use of them—namely, resource-rich corporate entities. As Srnicek (2022) argues, tech companies' dominance of compute resources has significantly contributed to their competitive advantage and increasing monopolization of AI technologies.

Corporate adoption

Public datasets hold a key, often unacknowledged role in corporate ML research and development. Despite the perception that large technology companies such as Google, Amazon, and Microsoft have more than enough internal data to develop ML systems, creators noted that many publicly accessible datasets were most commonly adopted in corporate settings. "The industrial research community," as one participant explained, "is huge and data-starved, in terms of open data" [Interview with Harper, MovieLens]. For corporate ML researchers, gaining access to internal company data is "so slow, so bureaucratic, it's so political as well" [Interview with McAuley, Amazon Reviews]. They often face regulatory barriers and internal policy processes, hindering their ability to easily access and use company data (often for good reasons, such as user privacy). These resource restrictions turn corporate researchers toward publicly available datasets, which are often adopted as a proxy for proprietary data, as one creator explains: "you can drive research much faster by using low sensitivity, proxy datasets." He continued, "I get all kinds of queries from people in Amazon about my dataset. And they would use it, probably while they were going through the bureaucracy of getting official permission" [Interview with McAuley, Amazon Reviews]. However, the extent of the use of datasets in corporate contexts is uncertain beyond citations within articles. Baio (2022) identifies that this public-to-private pipeline of datasets facilitates "data laundering" where private corporations devour and exploit public datasets, even relicensing content for commercial use and benefit without data subjects' knowledge or consent. While public artifacts may not perfectly model their internal data, corporate researchers may adopt these datasets as readily available makeshift solutions to bureaucratic barriers. As such, some creators perceive that there is a "mismatch" between the two sectors: dataset creation is "probably a lot more valuable for industry than really academia" [Interview with Wang, GLUE; SuperGLUE]. This highlights the porous boundaries of the laboratory whereby proxies intended for research and knowledge production leave "indelible marks" upon industry technologies (Mulvin, 2021: 143).

Shortcuts

> *These benchmarks are proxies and come with the collateral damage of sometimes being unrepresentative of the bigger goal that we want to solve.* [Interview with Zamir, Taskonomy; UCF101]

Creators described relying on several shortcuts to construct and circulate datasets. The significant resource commitment to construct and utilize datasets encourages shortcuts that simplify the knowledge production process. As Star (1983) argues, the production of scientific knowledge requires simplifications at every stage of research work. Datasets themselves are shortcuts—proxies for something in the world they wish to model. Thus, as Mulvin (2021) suggests, we should be attentive to why and how they are given the power to stand in for the world. As these shortcuts become institutionalized within the production and circulation of datasets, creators themselves can overlook their "reconciliation work" (Star, 1983: 206) of aligning the proxies with the phenomena they represent. In other words, shortcuts are reified, naturalized, and their contingencies forgotten. In this section, we detail the shortcuts that creators rely upon in the creation and use of datasets, and what this produces. These shortcuts can give way to systemic failures that not only affect the contents of datasets, and the models trained on them, but also the integrity of the institutions that deploy them.

Filtering

As datasets are compiled, creators rely on automated shortcuts to reduce their own labor burden. One such shortcut is using filtering mechanisms to remove unwanted data from their dataset. Filtering algorithms allow creators to automate some elements of curation, thus alleviating the resource burden of having to manually identify and delete undesirable content. Deciding the content deemed 'undesirable' is itself a subjective and messy process (Jaton, 2021; Thylstrup and Waseem, 2020), ranging from duplicate and low-quality data, to content deemed inappropriate or socially harmful. Filtering mechanisms, however, introduce further contingencies and potential errors.

For example, filtering algorithms allowed creators and institutions to maintain an appearance of neutrality when trying to decide what constituted pornographic content. In the case of the C4 team, they chose to apply a publicly available list called "Dirty, Naughty, Obscene or Otherwise Bad Words" to filter their dataset:

> We didn't want to try to come up with a definition of what is and is not porn, or is and is not obscene . . . We just took that list [of obscene words] as is. Because again, if we took the list and we changed it, then we were putting our own beliefs about what is and isn't bad. [Interview, anon., C4]

As shortcuts, filtering techniques encourage a disavowal of accountability by creators for the artifacts they construct. In this case, the politics embedded within filtering mechanisms were unquestioned and allowed to shape the filtered dataset. A creator of the C4 dataset reflected that the filtering mechanism was "very poor and boneheaded for the end goal of trying to just remove some porn from the dataset." In attempting to 'clean' the dataset of harmful content, these filtering mechanisms disproportionately removed content produced by and for minoritized groups, such as health content for LGBTQI+ communities and marginalized English dialects (Dodge et al., 2021), thus amplifying the harmful cis-white-heteronormative paradigm that is already prevalent within scraped datasets (Luccioni and Viviano, 2021).

There are many subjective issues when it comes to identifying "dirty, naughty, or obscene" content (For whom is it dirty? What constitutes naughty?). But automating these decisions reifies them as objective and authoritative while also allowing creators to avoid taking responsibility for their limitations (Thylstrup and Waseem, 2020). In the case of C4, this flawed filtering process was considered acceptable because the primary objective of the dataset was to train a language model to perform well on conventional benchmark datasets, and according to the creator, "because most benchmark datasets don't talk about sex, it probably doesn't hurt the model" [Interview, anon., C4]. These assumptions and shortcuts end up shaping data infrastructures and model evaluation practices. As a shortcut, filtering techniques can introduce new problems rather than addressing the structural limitations of a dataset.

Academic validation

Academic institutional mechanisms were also employed to verify the quality of a dataset. Porter (1999 [1996]) highlights how the academic peer review system is seen as the primary mark of impersonal, objective assessments of academic rigor and success. In the case of datasets, the publication process was seen as a suitable proxy for evaluating the quality of a dataset. One creator explained: "Because most datasets were already published, we didn't even check for these kinds of biases and stuff like this" [Interview, anon., TweetEval].

Creators also perceived the citation count of a dataset to be an objective measure of dataset quality. For instance, benchmark datasets with high citation counts were considered to be "implicitly community vetted" [Interview with Wang, GLUE; SuperGLUE] and reliable representatives of certain tasks. However, using citation count as an indication of dataset quality can create a "closed ecosystem with positive feedback" [Interview with Zamir, Taskonomy; UCF101] in which datasets continue to receive citations well after they are perceived to be outdated. Creators explained how the peer review system expects them to evaluate their models using the same benchmarks as previous state-of-the-art models, leading to using outdated, unnecessary, or flawed datasets. As one creator explained: "When you put a goal post in front of people, they're going to fixate on that goal post and run towards it" [Interview with Zamir, Taskonomy; UCF101]. In practice, the citation count of a dataset is a signal of the general awareness of a dataset, not the quality of that dataset.

The perceived objectivity of these shortcuts for assessing dataset quality is encouraged by the peer review system itself. One creator explained the general perception that peer review would penalize work that is open and earnest about its limitations: "You are going to get rejected more often because the reviewers would just look at limitations and copy-paste into a review" [Interview with Zamir, UCF101; Taskonomy]. Thus, not only do institutional shortcuts encourage the perception that datasets are fit-for-use, objective, and of high quality, it discourages alternative interpretations by encouraging dataset creators to hide limitations and contingencies. These 'peer-washing' practices maintain datasets as authoritative proxies even when they are shown to be harmful or problematic.

Accuracy scores

The performance of ML models on benchmark datasets is calculated as a percentage of correctly identified outputs, which is called an accuracy score. Accuracy scores are used as shortcuts to evaluate a model's performance. As noted, benchmarks can only ever provide a partial evaluation of a model's performance. And yet the research community depends on these metrics as reliable evaluation tools (Raji et al., 2021). Publication in highly ranked conference proceedings and journals often depends on achieving state-of-the-art performance on benchmarks.

However, creators also noted the limits of benchmarks as shortcuts for evaluating a model's performance. For instance, some large models have exploited statistical patterns in datasets, producing state-of-the-art performance while being unable to solve a simple problem that is not found within the dataset. Known as 'overfitting,' models can produce inflated accuracy scores without making substantive technical progress. But without formal standardization for evaluation, model creators cherry-pick benchmark datasets that are more favorable to increase their likelihood of publication: "People don't necessarily like a more realistic but harder problem. People like an easier problem that is more publishable" [Interview with Zamir, Taskonomy; UCF101].

Overfitting obscures a lack of progress within the field as it is difficult to distinguish between models that have solved a problem 'correctly' versus those that have sought statistical shortcuts. Although accuracy scores are relied upon for model evaluation, they are imperfect shortcuts for verifying the performance of a model or the technical progress of the field.

Accountability

> *I came in, so T5 was the nine of us. And none of us was the dataset person . . . none of us were saying, "Hey, we're really, really, really, going to own the dataset and make it as good as possible."* [Interview, anon., C4]

Creators often grappled with the distributed and diffuse nature of accountability for the contents and impacts of datasets. Without standardized practices of dataset development, or clear and enforceable documentation practices, accountability for datasets remains limited (Gebru et al., 2021; Luccioni et al., 2022). Studies with artificial intelligence (AI) practitioners reveal how accountability for the systems they create is deferred to stakeholders throughout the pipeline of AI development (Orr and Davis, 2020). As modular technologies with multiple stakeholders, datasets are another case of "dislocated accountabilities" (Widder and Nafus, 2023). Creators illustrated a deferred chain of responsibility for datasets with the only exceptions being (a) where it may present legal liability to them, and (b) where it can be entirely placed with the user of the dataset. While this echoes arguments that platforms are neutral infrastructures and thus not responsible for the content that their users post (Gillespie, 2010), it is a particularly ill-suited metaphor for datasets as they are profoundly important in shaping models. There is, in other words, no way for users to produce a model that is not shaped by the dataset. Yet, the dataset creators we interviewed often deferred accountability to their legal advisors and to the users of their creations. Although some participants recognized the inadequacy of this approach and located some accountability with themselves, they were also ambivalent about their own ability to exert any real control over their creations. As one participant explains: “You cannot take care of all the things that are happening along the value chain” [Interview, anon., YFCC].

Legal departments and constraints

Creators primarily located accountability for their creations with the law and the legal departments of their institutions. For those operating in the private sector, corporate lawyers would dictate the final form of the dataset based on liability concerns. While corporate legal counsel rarely assessed the specifics of datasets directly, they would impose strong constraints upon what dataset creators could do. For instance, some creators said that they were prohibited from publishing images, and must instead publish *links* to images, to allow users the ability to remove their data if they desired. This has become a common technique used by dataset creators to avoid perceived legal accountability.

Often, these constraints were motivated by protecting their corporate image and proprietary information. As a former Google employee explained: “Google is not going to release its own scrape, they will never do that” [Interview, anon., C4]. And yet, they were also discouraged from cleaning and rereleasing public data as “Google really doesn’t want to be in the position of ever saying this web content is better than this web content” [Interview, anon., C4]. As such, their legal department denied them the ability to publish a dataset at all, but instead allowed them to publish the instructions of how to make their dataset from publicly available sources. This is a common practice of *decentralizing risk* in dataset construction. Due to the centralized control that legal departments had over the creation of datasets, creators felt that their own role was limited: “We had constraints. We were navigating a narrow path. There was not much degree of freedom” [Interview, anon., YFCC]. Creators commonly deferred accountability for the dataset to the legal departments while framing themselves as passive operators of someone else’s commands.

Creators noted the numerous legal constraints that they were working within and were cautious that their datasets were compliant. These negotiations were most prevalent as creators navigated the various licenses of the data included within their dataset. Licenses detail the copyright restrictions placed on data, informing how data are able to be used and repackaged. These copyright restrictions flow on and shape the legal use restrictions of creators’ datasets. Violating these terms opens creators up to legal actions (e.g. MegaFace, see Hill and Krolik, 2019). Consequently, legal teams were “super, super sensitive” [Interview, anon., YFCC] to the restrictions imposed upon a dataset.

However, as licenses were perceived as the primary arbiter of what creators were *forbidden* from doing, they were also embraced as a guide of what creators were *permitted* to do. For instance, creators explained that “Because people annotated creative common, we could just publish that” [Interview, anon., YFCC]. Here, the declaration of a Creative Commons license became conflated with the consent of uploaders for the use of their data within ML datasets. This is despite the recognition that content creators rarely knew or consented to their data being used for ML datasets. Misusing Creative Commons licenses beyond their stated restrictions (such as the requirement of attribution or non-commercial use) raises copyright infringement concerns as well as concerns about enabling the surveillance of data subjects and undermining rights to individual privacy and consent (Harvey, 2022). Notably, there is no current Creative Commons license that explicitly forbids data from being taken for use in an ML training dataset.

Users

Creators also located accountability with the users of their datasets. As datasets are used in diverse ways, often beyond creators' intentions and without their knowledge, creators often felt it was not their responsibility to know or track how their datasets were used once they were released. Dataset users were perceived as the primary responsible actors: "You create something, and there is a moment where . . . you cannot make decisions because it is owned by the community" [Interview, anon., YFCC]. Creators recognized that users' legal restrictions may differ heavily depending on their domestic context, and they generally felt unfit to advise users about their specific legal constraints. As one participant explained: "It's always on the users to decide whether they can do what they want in their legal context" [Interview with Nagel, Common Crawl]. Moreover, while some datasets place extensive use restrictions on their datasets, these were always left for the users to interpret and implement. At the same time, creators have no way of knowing if users read, comprehend, or abide by these rules. As such, creators said that there is "no practical way to enforce our terms of use" [Interview with Nagel, Common Crawl]. This deferral of accountability echoes Langdon Winner's (2001 [1978]) description of a technological Frankenstein phenomenon, where "a man who creates something new in the world . . . then pours all of his energy into an effort to forget" (p. 313). Despite the global impact of datasets, they are generally released "with no real concern for how best to include it in the human community" (Winner, 2001 [1978]: 313). Although creators are tasked with making complex trade-offs in their work, they ultimately placed the locus of accountability with users.

Personal responsibility

Some creators did acknowledge their own personal responsibility in the development of their datasets. Despite deferring responsibility for the dataset to legal departments and users, in practice, creators recognized that they were the ones who had the power to make decisions. It was ultimately the creators' role to balance competing priorities and negotiate conflicting values. For instance, some creators justified the presence of harmful and inappropriate content in their dataset as concern for social harm was beyond the scope of the dataset's aims. As Jitsev [Interview, LAION-5B] explained: "We were less worried about the bias amplification because we were more in the realm of reproducing the results." In other cases, these trade-offs were made between the perceived utility of a dataset and the privacy of data subjects represented within it. The deferral of accountability cabined creators from these considerations. For instance, while one creator admitted that consent and privacy are "real concerns" for scraped datasets, they explained that "we haven't thought much about them, or maybe we never would have done this in the first place" [Interview with McAuley, Amazon Reviews].

Another creator recognized how the constraints of their work conflicted with their own values: "As a consequence of the constraints we were working in and the goals of the project, we ended up implementing a pipeline that did something that feels kind of bad and wrong in certain ways" [Interview, anon., C4]. These cases underscore the disconnection between the legal requirements and the ethical implications of a dataset in practice. The goals of the ML projects were prioritized over the potential harm the dataset may cause. Even if datasets do adhere to legal constraints, that alone is no guarantee that they will not cause harm. While creators recognized these issues, it was generally seen as beyond their control, as the ultimate decisions were up to dataset users.

As we have shown, dataset creators face a multitude of challenges related to *scale*, *resources*, *shortcuts*, and *accountability* that shape datasets and their social impacts. Together, these challenges have meant that many dataset creators feel unable to adequately care for their creations, and they can experience a kind of 'accountability distancing' where they do not see a way to control or maintain responsibility for either the contents or the ultimate uses of their datasets. Yet the datasets they produce have significant impacts on data subjects, dataset users, corporations, academic institutions, and those subjected to ML systems around the world. This tension speaks of a wider uncertainty in the field: where should responsibility be located when the systems trained on these datasets have harmful, dehumanizing, or discriminatory impacts?

Conclusion

> *Creating datasets well is actually just very difficult for all these reasons. There's so many things to keep in mind at the same time . . . I think it's very valuable work and very difficult to get right.* [Interview with Jia, SQuAD 2.0]

Dataset development profoundly shapes ML systems and the impact they have on the world. Yet dataset work remains an undervalued and under-researched juncture in the path toward understanding the culture of ML and limiting the harms of automated systems. Dataset development is far from cohesive, and the lack of standardized practices reveals a field that is struggling with a shared set of issues: the constant pressures of *scale*, the struggle for *resources*, the adoption of *shortcuts*, and confusion about *accountability*.

In sum, our participants expressed the ways in which creating high-quality datasets is a practice that is in flux, and requires considerable individual judgment, hard work, and an ongoing struggle for resources and ever-increasing scale. Clearer structures and lines of accountability are needed in order to transform the current ad hoc design trade-offs into thoroughly considered and transparent decisions. As one participant articulated, there was a desire to share stories as a way to build "best practices for the field" [Interview with Gould, IKEA], which is also the subject of ongoing work by the authors. Our research points to the need for the recognition of the social process of dataset-making, and its significant impact on ML systems, as well as the accountability vacuum that is growing in the ML field. Before this can be addressed, the field must recognize the limits, challenges, and shortcuts that dataset creators confront, even as their work is undervalued, hidden, or assumed to be primarily the result of automated processes.

By centering the voices of those charged with creating these influential artifacts, we hope that our research contributes to a much-needed cultural shift toward understanding and valuing dataset creators and their work, strengthening forms of accountability, and reorienting the understanding of ML as objective and automated toward the ways in which it is socially constructed and based on human judgments and values from the outset.

Acknowledgements

This work began as an internship project at the Fairness, Accountability, Transparency and Ethics (FATE) group at Microsoft Research and continued at the Knowing Machines Project. The authors express gratitude to their colleagues across MSR, NYU, and USC, with specific thanks to the following members of the Knowing Machines project who gave helpful feedback and assistance of many kinds: Mike Ananny, Edward Kang, Sasha Luccioni, Jason Schultz, and Hamsini Sridharan. Above all, we would also like to thank the interview participants for their generous contributions to this research. The Knowing Machines Project is supported by the Alfred P. Sloan Foundation.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Microsoft Research and the Alfred P. Sloan Foundation supported the research for this article.

ORCID iD

Will Orr https://orcid.org/0000-0002-6720-5794

Note

[1] Sites like Academic Torrents continue to circulate datasets that have been removed by creators. Recent literature advocates for practices of care regarding the development, maintenance, and distribution of datasets and includes mechanisms like Digital Object Identifiers to make dataset removal easier to track (Luccioni et al., 2022).

References

Baio A (2022) AI data laundering: how academic and nonprofit researchers shield tech companies from accountability. Available at: https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield-tech-companies-from-accountability/ (accessed 12 October 2022).

Baldominos A, Saez Y and Isasi P (2019) A survey of handwritten character recognition with MNIST and EMNIST. *Applied Sciences* 9(15): 3169.

Birhane A, Prabhu VU and Kahembwe E (2021) Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv [preprint]*. DOI: 10.48550/arXiv.2110.01963.

Bowker GC and Star SL (1999) *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Brown T, Mann B, Ryder N, et al. (2020) Language models are few-shot learners. In: Solla SA, Leen TK and Müller KR (eds) *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, pp. 1877–1901. Available at: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (accessed 19 September 2022).

Chun WHK (2021) *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition*. Cambridge, MA: MIT Press.

Crawford K (2021) *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press.

Crawford K and Paglen T (2019) Excavating AI: the politics of training sets for machine learning. Available at: https://excavating.ai (accessed 12 June 2022).

Dodge J, Sap M, Marasović A, et al. (2021) Documenting large Webtext corpora: a case study on the colossal clean crawled corpus. In: *Proceedings of the 2021 conference on empirical methods in natural language processing*, Punta Cana, Dominican Republic, November, pp. 1286–1305. Kerrville, TX: Association for Computational Linguistics.

Douglas H (2004) The irreducible complexity of objectivity. *Synthese* 138(3): 453–473.

Gebru T, Morgenstern J, Vecchione B, et al. (2021) Datasheets for datasets. *arXiv [preprint]*. DOI: 10.48550/arXiv.1803.09010.

Geiger RS, Yu K, Yang Y, et al. (2020) Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In: *FAT* '20: Proceedings of the 2020 conference on fairness, accountability, and transparency*, New York, 27 January, pp. 325–336. New York: ACM.

Gillespie T (2010) The politics of "platforms." *New Media & Society* 12(3): 347–364.

Goodfellow I, Bengio Y and Courville A (2016) *Deep Learning* (Adaptive Computation and Machine Learning). Cambridge, MA: MIT Press.

Hanna A and Park TM (2020) Against scale: provocations and resistances to scale thinking. *arXiv [preprint]*. DOI: 10.48550/arXiv.2010.08850.

Harvey A (2022) On creative commons—the exploitation of photography: how creative commons licenses enable surveillance. Available at: https://ahprojects.com/creative-commons/ (accessed 9 October 2022).

Hill K and Krolik A (2019) How photos of your kids are powering surveillance technology. *The New York Times*. Available at: https://www.nytimes.com/interactive/2019/10/11/technology/flickr-facial-recognition.html

Hoffmann J, Borgeaud S, Mensch A, et al. (2022) Training compute-optimal large language models. *arXiv [preprint]*. DOI: 10.48550/arXiv.2203.15556.

Jaton F (2021) *The Constitution of Algorithms: Ground-Truthing, Programming, Formulating*. Cambridge, MA: MIT Press.

Kang EB (2023) Ground truth tracings (GTT): on the epistemic limits of machine learning. *Big Data & Society* 10(1): 20539517221146122.

Koch B, Denton E, Hanna A, et al. (2021) Reduced, reused and recycled: the life of a dataset in machine learning research. In: *Thirty-fifth conference on neural information processing systems datasets and benchmarks track*, 20 August. Available at: https://openreview.net/forum?id=zNQBIBKJRkd (accessed 12 June 2022).

Latour B (2015) *Science in Action: How to Follow Scientists and Engineers Through Society*. Cambridge, MA: Harvard University Press.

Luccioni A and Viviano J (2021) What's in the box? An analysis of undesirable content in the common crawl corpus. In: *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 2: short papers)*, Online, August, pp. 182–189. Kerrville, TX: Association for Computational Linguistics.

Luccioni AS, Corry F, Sridharan H, et al. (2022) A framework for deprecating datasets: standardizing documentation, identification, and communication. In: *FAccT '22: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, New York, 21 June, pp. 199–212. New York: ACM.

Mulvin D (2021) *Proxies: The Cultural Work of Standing In*. Cambridge, MA: MIT Press.

Orr W and Davis JL (2020) Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society* 23(5): 719–735.

Pink S, Ruckenstein M, Willim R, et al. (2018) Broken data: conceptualising data in an emerging world. *Big Data & Society* 5(1): 2053951717753228.

Plantin J-C (2019) Data cleaners for pristine datasets: visibility and invisibility of data processors in social science. *Science, Technology, & Human Values* 44(1): 52–73.

Porter TM (1999 [1996]) *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (History and Philosophy of Science). Princeton, NJ: Princeton University Press.

Raji ID, Denton E, Bender EM, et al. (2021) AI and the everything in the whole wide world benchmark. In: *Thirty-fifth conference on neural information processing systems datasets and benchmarks track*, 20 August 2021. Available at: https://openreview.net/forum?id=j6NxpQbREA1 (accessed 12 June 2022).

Ramesh A, Pavlov M, Goh G, et al. (2021) Zero-shot text-to-image generation. *arXiv [preprint]*. DOI: 10.48550/arXiv.2102.12092.

Sambasivan N, Kapania S, Highfill H, et al. (2021) 'Everyone wants to do the model work, not the data work': data cascades in high-stakes AI. In: *CHI '21: Proceedings of the 2021 CHI conference on human factors in computing systems*, New York, 7 May, pp. 1–15. New York: ACM.

Seaver N (2021) Care and scale: decorrelative ethics in algorithmic recommendation. *Cultural Anthropology* 36(3): 3.

Seger E, Ovadya A, Garfinkel B, et al. (2023) Democratising AI: multiple meanings, goals, and methods. *arXiv [preprint]*. DOI: 10.48550/arXiv.2303.12642.

Srnicek N (2022) Data, compute, labor. In: Graham M and Ferrari F (eds) *Digital Work in the Planetary Market*. Cambridge, MA: MIT Press, pp. 241–261.

Star SL (1983) Simplification in scientific work: an example from neuroscience research. *Social Studies of Science* 13(2): 205–228.

Thylstrup N and Waseem Z (2020) Detecting "dirt" and 'toxicity': rethinking content moderation as pollution behaviour. SSRN Scholarly Paper no. 3709719. Available at: https://doi.org/10.2139/ssrn.3709719

Thylstrup NB (2022) The ethics and politics of data sets in the age of machine learning: deleting traces and encountering remains. *Media, Culture & Society* 44(4): 655–671.

Torralba A, Fergus R and Freeman WT (2008) 80 million tiny images: a large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(11): 1958–1970.

Widder DG and Nafus D (2023) Dislocated accountabilities in the "AI supply chain": modularity and developers' notions of responsibility. *Big Data & Society* 10(1): 20539517231177620.

Winner L (2001 [1978]) *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought*. Cambridge, MA: MIT Press.

Author biographies

Will Orr is a doctoral candidate in the Annenberg School for Communication and Journalism at the University of Southern California. His research explores the culture and politics of data, focusing on the sociotechnical challenges faced by creators throughout the machine learning pipeline.

Kate Crawford is a research professor in Communication and STS at the Annenberg School, and a faculty affiliate of Science, Technology, and Society at the University of Southern California. She is also a senior principal researcher at Microsoft Research New York, and the Lead PI of the Knowing Machines Project.