There has been increased attention toward the datasets that are used to train and build AI technologies from the computer science and social science research communities, but less from legal scholarship. Both Large-Scale Language Datasets (LSLDs) and Large-Scale Computer Vision Datasets (LSCVDs) have been at the forefront of such discussions, due to recent controversies involving the use of facial recognition technologies, and the discussion of the use of publicly-available text for the training of massive models which generate human-like text. Many of these datasets serve as “benchmarks” to develop models that are used both in academic and industry research, while others are used solely for training models. The process of developing LSLDs and LSCVDs is complex and contextual, involving dozens of decisions about what kinds of data to collect, label, and train a model on, as well as how to make the data available to other researchers. However, little attention has been paid to mapping and consolidating the legal issues that arise at different stages of this process: when the data is being collected, after the data is used to build and evaluate models and applications, and how that data is distributed more widely.

* Mehtab Khan, Resident Fellow at the Information Society Project at Yale Law School; Alex Hanna, Director of Research at the Distributed AI Research Institute (DAIR). This work was previously presented at the 2022 Privacy Law Scholars Conference. Special thanks to Emily McReynolds, Jason Schultz, Jen King, Jevan Hutson, Kirsten Martin, Luiza Jarosky, Margaret Mitchell, Negar Rostamzadeh, and Ryan Steed for helpful comments on this paper. Thanks to Jenny Paquette for editorial assistance.
In this article, we offer four main contributions. First, we describe what kinds of objects these datasets are, how many different kinds exist, what types of modalities they encompass, and why they are important. Second, we provide more clarity about the stages of dataset development—a process that has thus far been subsumed within broader discussions about bias and discrimination—and the subjects who may be susceptible to harms at each point of development. Third, we provide a matrix of both the stages of dataset development and the subjects of dataset development, which traces the connections between stages and subjects. Fourth, we use this analysis to identify some basic legal issues that arise at the various stages in order to foster a better understanding of the dilemmas and tensions that arise at every stage. We situate our discussion within wider discussion of current debates and proposals related to algorithmic accountability. This paper fulfills an essential gap when it comes to comprehending the complicated landscape of legal issues connected to datasets and the gigantic AI models trained on them.
INTRODUCTION ............................................................................................................. 174

I. UNDERSTANDING DATASETS ..................................................................................... 182
   A. Reconstruing Dataset Development ..................................................................... 184
   B. ImageNet: A Broad Search Engine Strategy ......................................................... 186
   C. Faces in the Wild: Public Figures in the News ...................................................... 189
   D. Common Crawl: Archiving the Whole Web ............................................................ 190
   E. The Pile: Data Documentation and Safety Considerations ...................................... 193

II. TAXONOMIES OF DATASET DEVELOPMENT STAGES AND SUBJECTS ................. 195
   A. The Stages of Dataset Development ...................................................................... 195
   B. The Subjects of Dataset Development .................................................................... 201

III. LEGAL INFRASTRUCTURE SHAPING DATASET DEVELOPMENT ................................. 204
   A. Copyright and Fair Use .......................................................................................... 206
      1. Fair Learning ....................................................................................................... 208
      2. Fair Use as a Corrective Measure ....................................................................... 215
   B. Privacy Harms ....................................................................................................... 219
      1. Individual Privacy and Dataset Development ..................................................... 219
      2. Systemic Privacy in Dataset Development ......................................................... 224

IV. A MATRIX OF DATASET DEVELOPMENT HARMS .................................................. 229
   A. Applying the Matrix .............................................................................................. 230
      1. Problem Formulation ......................................................................................... 231
      2. Data Collection .................................................................................................... 233
      3. Data Cleaning ....................................................................................................... 236
      4. Data Annotation ................................................................................................... 237
      5. Model Training and Evaluation .......................................................................... 240
      6. Model Implementation and Inference ................................................................... 242
      7. Data and Representation Distribution ................................................................... 244
   B. Normative and Regulatory Approaches to Dataset Accountability ....................... 245
      1. Systemic and Individual Algorithmic Accountability ............................................. 245
      2. Proposed Legislation .......................................................................................... 252
   C. Summary ................................................................................................................. 255

V. CONCLUSION ............................................................................................................. 256
INTRODUCTION

In January 2020, Robert Julian-Borchak Williams was accused of felony larceny for stealing watches from a boutique in Detroit. He had been arrested on his front lawn in front of his wife and two children, held in custody for 30 hours, and released after paying a $1,000 bond. Mr. Williams, a 41-year-old Black man, had been identified via a facial recognition algorithm provided by the company DataWorks Plus and purchased for use by the Detroit Police Department. The face identified in the image, however, was not his.¹

Since his arrest, at least two other Black men in Detroit have been incorrectly identified via facial recognition algorithms.² Facial recognition systems are being used by thousands of police departments in the U.S. Clearview AI is the most discussed of these systems due to its popularity, the size of its image database, and the brazenness of its founders and funders.³ Clearview, according to reporting, has a U.S.-based clientele of over 3,000 customers, mostly from within law enforcement and military entities.⁴ It maintains that it has a database of three billion images scraped from the “public” web but also Facebook, Instagram, Twitter, and Google.⁵ Large social media companies have filed suit against the startup for violations of their terms of service, and

⁵ Hill, The Secretive Company That Might End Privacy as We Know It, supra note 3.
plaintiffs have sought injunctions against Clearview AI for its data scraping practices, but this lawsuit has not provided decisive conclusions about acceptable data scraping practices.\(^6\)

The failures of models that interpret, generate, and translate human language are less visible but no less insidious. In 2017, Israeli police arrested a Palestinian construction worker after capturing a picture of himself leaning against a bulldozer and captioning the text with “good morning” in Arabic on his Facebook page. An automated service used by Facebook translated the caption as “attack them” in Hebrew and “hurt them” in English.\(^7\) Safiya Noble has documented how Google text searches for Black, Asian, or Latina women have disproportionately surfaced pornographic images and sites.\(^8\) Nearly all search engine technology now uses some kind of language model for query expansion and question answering.\(^9\) Language models for language generation also have the ability to surface personally identifiable information (PII); given a prompt, these models have surfaced information like full names, physical addresses, telephone numbers, and email addresses.\(^10\)

More recently, a subset of language models has been used for the generation of computer code. GitHub’s Copilot technology has been framed as a “pair programmer”, a tool that allows software engineers to generate code.\(^11\) This tool works by analyzing a subset of the code in a


\(^{10}\) Nicholas Carlini et al., Extracting Training Data from Large Language Models, ARXIV (June 15, 2021), https://arxiv.org/pdf/2012.07805.pdf [https://perma.cc/C26M-PB2Q].

particular file, along with natural language comments and in-line documents, to suggest new code to be written in the file. Although GitHub’s own analysis suggests that this tool does not simply memorize or regurgitate elements of code, others have noted that there are some unique algorithms that have been copied whole cloth, such as the fast inverse square root algorithm used in the computer game Quake III, with some relatively simple prompts.\textsuperscript{12} From the cybersecurity angle, researchers have found that these models can reintroduce common code vulnerabilities because the training data suggests the most common code instances, rather than coding practices that are built for security and guard against vulnerabilities.\textsuperscript{13} The Free Software community has contended that this kind of code reuse can cause conflicts in licensing terms, calling into question the nature of open licensing practices in machine learning.\textsuperscript{14}

The examples described here have a few features in common. First, the underlying technology is based on large datasets built on millions of data points. These datasets undergird our modern artificial intelligence (AI) infrastructure by being used to train, test, and index modern AI technologies. We engage with these technologies on the web and in public life, often without our explicit knowledge. These datasets can be considered infrastructural, insofar as they do not appear unless the AI models which are built upon them break down; that is, they make egregious technical errors or mistakes in classification.\textsuperscript{15} Others break down when people who are not part of the majority—including racial, religious, ethnic, caste, and gender minorities, disabled people, refugees and migrants, and others in the Global South—interact with them at points of disjuncture, some of them violent and life-altering, as the examples above indicate.

Second, the models developed on large datasets have led to biased, discriminatory, or otherwise harmful effects on ordinary people, often belonging to marginalized social and economic groups. The

\textsuperscript{12} Id.
impact of these technologies can also be seen as the source of new kinds of informational harm.\textsuperscript{16}

Third, these data are often scraped from the web, but they are also often taken from books, legal proceedings, and patents. The practices around data collection and mining are known to take place often with little regard to potential copyright and privacy issues that may arise in the process.\textsuperscript{17} Moreover, large language models can possibly expose personally identifiable information or pose security risks via information leakage.\textsuperscript{18} The development process of large datasets also poses a unique set of questions about the limitations and norms of licensing regimes, and in particular, open licensing options when the data scraped may be subject to various licensing terms.\textsuperscript{19} The controversies mentioned above and the ways in which datasets are shaping public life have raised questions about how training and benchmark datasets are developed and what should be done in response to publicly-available data being used to build harmful technologies.

Lastly, the categorization and annotation of data exacerbate issues of representation and bias. Large visual datasets have well-documented issues related to consent and the types of categorical labels which are applied to humans.\textsuperscript{20} Likewise, large language models have


\textsuperscript{18} Carlini et al., supra, note 10, at 1.


These emerging questions warrant a closer examination of how we as a society should view and regulate dataset development. We aim to lend some clarity on that underlying gap in our understanding of datasets. Aside from identifying the points of intervention in dataset development, we emphasize that datasets form the backbone of AI applications in our public and private lives. We can no longer ignore the central role that datasets play in the deployment of AI applications. We need to be attentive to the ways that owners and curators of these datasets can and should be held accountable for the harms which arise from their development, usage, and deployment.

Large datasets have already been the subject of cursory forms of regulatory intervention. For instance, the FTC ordered the image storage company Everalbum, Inc. to delete its datasets after it found that it used images to train facial recognition technology without consent from its users.\footnote{See Tiffany C. Li, \textit{Algorithmic Destruction}, 75 \textit{SMU L. Rev.} 479 (2022); Kim Lyons, \textit{FTC Settles with Photo Storage App that Pivoted to Facial Recognition}, \textsc{The Verge} (Jan. 31, 2021), \url{https://theverge.com/2021/1/31/21912673/ftc-everalbum-lawsuit-settlement-facial-recognition}.} However, data deletion has limited effectiveness since dataset
development practices mean that there are countless copies of a dataset available on private servers even after it has been deleted by its curators. A recent study on datasets found that retractions of data are ineffective because the data still remains widely available. Therefore, establishing data deletion as a remedy has its drawbacks and underscores the need for interventions much earlier in the dataset cycle.

AI researchers have recently called models that are built on large datasets “foundational”, such that in the future they will undergird much more complicated models and be applied in many different contexts, from healthcare, law, and education. Developing these models has become the new AI gold rush, with large firms like Google and Meta, as well as AI-centric firms like OpenAI and Anthropic, releasing new large language models on a constant basis. The vast majority of the training data from these models are private. Thus, we are at a critical juncture in which the harms arising from the creation of these datasets need to be more carefully enumerated. We need to consider what types of policy interventions can be made to mitigate these harms. While scholars have discussed datasets in the context of specific applications, such as facial recognition, we have not yet seen a general framework to discuss and understand the scraping and processing of image and textual data at the massive scale under which it is being undertaken today.

This article makes four main contributions. First, we describe what kinds of objects these datasets are, how many different kinds exist, what types of modalities they encompass, and why they are important. In Part I, we examine several examples of the datasets used in the development of AI systems. We focus on four datasets: two within the visual domain and two dealing with language. In our exposition of datasets, we outline the processes of their data collection, data cleaning, and data annotation before ever being put to use within a model.


Peng et al., supra note 17, at 1.


See Amanda Levendowski, Resisting Face Surveillance, 100 N.C. L. REV. 1015 (2022) (this paper is a discussion of how building facial recognition by using copyrighted images may not be fair use); Mark Lemley & Bryan Casey, Fair Learning, 99 TEx. L. R. 743, (2020) (this paper proposes a default fair use application for data that is copied for machine learning applications).

Lemley & Casey, supra note 25, at 1045.
Embedded in this analysis is the belief that these steps matter greatly for the downstream consequences and effects of how these datasets are used and their legality.\textsuperscript{27} We hope to clarify some common misconceptions about these datasets and detangle how they work within AI models. In this work, we identify two types of datasets: large-scale computer vision datasets (LSCVDs) and large-scale language datasets (LSLDs).\textsuperscript{28} Discussions within the legal domain thus far have been about downstream representations, such as word2vec,\textsuperscript{29} heterogeneous data sources (both public domain and copyrighted "high quality" works), and large-scale datasets such as ImageNet. Delineating the various stages of dataset development can aid with what many stakeholders seem to be struggling with: how to enact accountability measures for the harms caused by AI.\textsuperscript{30}

Second, we propose two taxonomies, first of dataset development stages and second, the individual stakeholders who are subject to harm at different points of this process. We find this is necessary in order to understand the source of the problems arising from various AI applications and, most importantly, the datasets used in the development of those applications. In Part II, we expand on the taxonomy of the stages and stakeholders involved in the process. We need more clarity about the stages of dataset development because this discussion has thus far been subsumed within broader discussions about bias and discrimination, which focus heavily on issues around data distribution and model deployment. Although these two moments


\textsuperscript{28} Alec Radford et al., \textit{Learning Transferable Visual Models from Natural Language Supervision}, arXiv (Feb. 26, 2021), https://arxiv.org/abs/2103.00020 [https://perma.cc/3KJX-Z397] (finding that these are not the only types of datasets used within AI research, but they are the most relevant and paradigmatic; others include datasets for audio which are used in speech recognition modeling and so-called “multi-modal” datasets which are combinations of more than one modality of data, e.g., images with human language captions; for instance, OpenAI’s CLIP model learns image categories from [image, text] pairs on the web; although we focus here on LSCVDs and LSLDs, our argumentation easily applies to multi-modal datasets and models).

\textsuperscript{29} \textit{Id}.

\textsuperscript{30} See generally Andrew Selbst, \textit{Institutional View of AIAs}, 35 HARV. J. L. & TECH. 117 (this article is an exploration of how a mandate for algorithmic impact assessments may be imposed on private companies); Margot Kaminski & Jennifer Urban, \textit{The Right to Contest AI}, 121 COLUM. L. REV. 7 (2021) (this article builds a case for an individual right to contest decisions made by AI systems).
within a dataset's lifecycle are important and typically the most visible to the public, they are the product of all the previous steps in which potential interventions could be mounted to hold dataset developers accountable.

Third, we expand on a set of overarching legal issues that arise at these stages to foster a better understanding of the tensions and dilemmas involved in the dataset development process. In Part III, we specifically identify intersecting copyright and privacy issues arising at the various stages of a dataset's development. We focus on these aspects to build an overarching framework for dataset accountability. Scholars have discussed the copyright implications in building ML datasets and have also documented how some of these datasets—especially datasets involving facial images—have violated privacy expectations. But on the whole, these interrogations do not make the connections between the development and deployment of the dataset, and the role both copyright and privacy law play throughout. Similarly, privacy and data governance are an underexplored areas when it comes to datasets, and we propose conceptualizing harms arising in the process as deserving their own categories.

The copyright aspects of data collection and development of LSCVDs and LSLDs are not straightforward and intersect existing frameworks surrounding web scraping, data governance, and fair use. There is a complex landscape of rules that govern web scraping in general. Some of these existing narratives view the data collection process as linear and assume that users have the power to consent before their data is collected and used to train AI models. We expand on the privacy and copyright landscape and its implications.

Lastly, in Part IV., we combine the taxonomies of subjects and stages outlined in Part II to construct a matrix of informational harms. This matrix serves as an analytic which helps guide us to consider what harm may exist for each subject at each stage of the dataset development process. We discuss how current regulatory and normative discussions may use this analytic in fostering algorithmic accountability.

The overall aim of this article is to aid recent efforts to respond to individual and societal interests connected to the use of large datasets.

One of our subsequent aims is to help support the development of datasets such that individual rights are at the center, while not depressing novel computer vision and natural language processing research. We provide this overview to help understand those limitations and situate the discussion within larger debates about the accountability and governance of AI systems. It is essential to inform policymakers, researchers, and industry leaders about the complicated landscape of legal issues connected to large datasets, and a deeper understanding of the dataset development process is the first step.

I. UNDERSTANDING DATASETS

A great deal of focus on machine learning and other sociotechnical systems has aimed to understand them as systems that, like human actors, can result in decision-making that has disparate impacts towards groups with protected classes, such as race, gender, and class.32 Critical literature have congealed into a body of work that has fallen under a set of umbrella terms: fairness, accountability, and transparency (which has given rise to an academic conference of the same name: FAccT.)33 Also often included in this list are interpretability and explainability. Most of this work, which has focused on AI models, has characterized disparate results across protected classes as a form of bias, and accordingly, amending systems have focused on algorithmic interventions to alter the results of a system towards unbiased outcomes.34 Many of the paradigmatic cases within fairness research

32 See Barocas & Selbst, supra note 16.
have focused on the use of any kind of AI modeling system in high-risk domains, such as criminal justice, welfare allocation, and hiring.

A major line of work has turned attention away from focusing purely on the outcomes of sociotechnical systems towards focusing on data. Data, after all, is the foundation of the decisions that models make, and is often neglected in favor of doing the work of model development. Focusing on data has turned an eye toward training datasets for facial recognition which do not represent darker-skinned women and do not represent individuals, social contexts, or objects in the Global South. This has led to a number of interventions toward increasing documentation for datasets used in machine learning. Developments within AI research have attempted to remedy the lack of


attention to purposeful dataset development, with major research avenues instituting specific tracks for datasets.\textsuperscript{42}

These interventions are welcome and present a way forward for prospective dataset development. For our purposes, however, hundreds, if not thousands, of AI datasets exist in the wild. All of these datasets have their own contingent histories, conditions of creation, and, for the purposes of this article, legal and policy entanglements. Towards this end, we echo others in that there is a need to trace the histories and development of datasets used within machine learning.\textsuperscript{43}

A. Reconstruing Dataset Development

In this section, we describe several histories of the development of large datasets used in computer vision and natural language processing. LSCVDs and LSLDs have contextual and contingent histories. In this work, we broaden the focus from the technical dimensions of these datasets and into a broader genealogy of these datasets. A genealogy focuses on identifying the strategic emergence and transformation of subject positions and data roles and their deployment in a network of power relations.\textsuperscript{44} In this regard, this not only relies on paying attention to the technical dimensions of a dataset which may be reported in data documentation, but also being attentive to the organizations and institutions which have collected the data, the individuals which have labored on the data, and where and how these data were hosted and distributed. This intersects with the legal analysis of these datasets because legal restrictions and ambiguities on data strategies shape the practices and conditions in which large-scale AI datasets are developed. When we discuss the genealogies of these datasets and the assumptions and practices that dataset curators use to create their classifications and categories, we must also discuss the legal structures that allow or restrict these practices, the rights implicated in the process, the related harms that have yet to be defined.


\textsuperscript{43} Plasek, supra note 27, at 6 ("[W]e need [to write] histories of the datasets themselves.").

In this part, we describe both datasets which are used for the training of AI systems, but also datasets used for benchmarking AI systems. The distinction here is subtle but important. Training data is the information used to teach AI models what they know. Everything a model knows about the world comes from the data it is trained on. A benchmark dataset is used to judge how well a computer has learned to perform a task. Benchmarks are special sets of data that allow developers to compare their machine-learning methods against each other. They are measurement devices that provide an estimate of how well AI software will perform in a real-world setting. This distinction matters because the applicable scope of legal obligations would change based on these characteristics of a dataset. LSCVDs and LSLDs can be both training and benchmark datasets. While many of the LSCVDs and LSLDs which are used to train commercial AI products are proprietary and restricted to in-house development, there are a large set of benchmark datasets that are freely available online to machine learning researchers, academics, and industry practitioners. Some of these are housed in centralized repositories,^45 are packaged on secondary analysis sites for machine learning or data science development, such as Kaggle^46 and MLCommons,^47 or are integrated into major machine learning development environments such as TensorFlow^48 or PyTorch.^49

In the analysis below, we discuss two LSCVDs (ImageNet and Labeled Faces in the Wild) and two LSLDs (Common Crawl and The Pile). We focus on these datasets because they are some of the most well-cited in the literature, or they are some of the largest and the data on which much of the most modern machine learning models are trained on or evaluated against. Moreover, we chose to focus on these based on the variability of their sourcing, copyright status, and the privacy implications involved. LSCVDs and LSLDs are commonly used in both

^45 UCI MACHINE LEARNING REPOSITORY, https://archive.ics.uci.edu/ml/index.php [https://perma.cc/58K9-SYN8] (last visited July 31, 2022) (one of these being the UCI Machine Learning Repository, although these are used less as benchmarks and more as toy datasets for pedagogical reasons).
academia and industry as a means to report state-of-the-art results for computer vision algorithms, and the metrics associated with benchmarks are equated with scientific progress in computer vision subfields.

Computer vision (CV) is a field of research that, in a word, aims to allow computers to “see” elements of the real world.\textsuperscript{50} Tasks within computer vision include, but are not limited to, facial analysis and its subtasks of facial recognition, facial classification, and facial verification;\textsuperscript{51} object detection and localization; scene detection and segmentation; and gait detection and recognition. There are anywhere from 500 to 1,600 LSCVDs that are available and traceable through the computer vision literature, each of which has its own architectures of collection and sourcing.\textsuperscript{52}

Natural language processing (NLP) is a discipline at the nexus of computer science and linguistics.\textsuperscript{53} As the name suggests, the field is oriented towards interpreting and handling human languages (as distinct from machine-oriented programming languages). Tasks that rely on LSLDs include sentiment analysis, question answering, word sense disambiguation, text classification, and query expansion. There are somewhere in the range of 1,500 LSLDs,\textsuperscript{54} and, like LSCVDs, they all have their own means of collection and sourcing.

Most of these datasets are downloadable by anybody on the web – because they need to be available for others to evaluate their algorithms – and are often accompanied by a short Terms of Service (ToS) agreement. However, some do not have any ToS agreements that a user has to agree to before downloading the dataset.

B. ImageNet: A Broad Search Engine Strategy

\textsuperscript{50} Fei-Fei Li, \textit{How We’re Teaching Computers to Understand Pictures}, TED (March 2015), https://www.ted.com/talks/fei_fei_li_how_we_re_teaching_computers_to_understand_pictures [https://perma.cc/XSA6-7NC5].

\textsuperscript{51} See Scheuerman et al., \textit{supra} note 20 (helpful diagram on these subtasks).


\textsuperscript{53} DAN JURAFSKY & JAMES H. MARTIN, \textit{SPEECH AND LANGUAGE PROCESSING} (3rd ed. 2021) (there are many textbook level treatments of natural language processing).

\textsuperscript{54} \textit{PAPERS WITH CODE}, \textit{supra} note 52 (based on the number of datasets with the “text” modality).
The ImageNet dataset was developed by a team of researchers at Princeton and Stanford to support research and development into generalized visual object recognition. Consisting of over 14 million images organized into about 20 thousand categories, at the time of its creation it was one of the largest human-annotated image datasets ever developed and is orders of magnitude larger than its predecessors. Images include many human beings; these images may explicitly label a photograph as containing a human (i.e., if the image is labeled with a category that is derived from the top-level “person” category), but humans also appear throughout as incidental in other images in the dataset. Humans are thus located throughout. The categorical structure for the dataset is derived from WordNet, a large database of English words organized into a hierarchy based on semantic relationships developed by cognitive psychologist George A. Miller in the 1980s. The keywords for each concept were used to scrape web search engines for images associated with the concept. Led by Fei-Fei Li, Jia Deng, and Olga Russakovsky, the ImageNet team gathered images from the internet using multiple search engines (including the newly released Google Image Search) and entered queries into them based on the nouns in WordNet.

Several key affordances, some facilitated by developments in copyright law, came together to make ImageNet possible. First, the rise of digital image sharing during the decade prior to ImageNet’s creation,
coupled with search engines capable of indexing images on the web provided a mechanism for the ImageNet creators to construct a massive dataset of high-resolution images based on simple text-based queries. Because the ImageNet dataset scrapes data from the web, ImageNet does not claim to own the copyright to any of the images in the dataset and does not host any of the images on its own servers and instead chooses to host only the image URLs. For researchers who want to use the dataset and download images manually, the website requires registration and for users to sign a Terms of Service agreement. However, as of writing, links to past copies of the dataset are not available for download. ImageNet authors have not noted any privacy mitigation efforts, either in the original papers or within the follow-ups. Although a blog post that previews the 2020 update notes that they will “blur faces for privacy preservation.” However, given that currently, researchers can only download the original image URLs, this is not possible to implement.

Recent projects have been critical of ImageNet and large-scale computer vision datasets from an ethical perspective. The excavating.ai project and a short introduction to image datasets have focused on the “weird metaphysics” of ImageNet and the images which are featured within each category, including highlighting images that are bizarre, offensive, racist, sexist, homophobic, and transphobic. This has prompted its authors to sanitize the dataset by manually removing images that may contain offensive images and balancing the images across gender, age, and skin tone. It has also led to a broader interrogation of LSCVDs by highlighting how many of these datasets have not discussed the lack of consent.

---

60 Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015); Perfect 10, Inc. v. Amazon, Inc., 508 F.3d 1146 (9th Cir. 2007) (allowing copying of books and images for what courts deemed to be transformative purposes).
63 Crawford & Paglen, supra note 31.
64 Malevé, supra note 20.
65 Yang et al., supra, note 62.
66 See Prabhu & Birhane, supra note 31; Scheuerman et al., Computer Vision Values Dataset, supra note 52 (in a recent survey of many LSCVDs, Scheuerman and co-authors find that consent is rarely accounted for, unless the images are collected under strict laboratory conditions).
C. Faces in the Wild: Public Figures in the News

Faces in the Wild (FW) is a dataset developed for the purpose of labeling faces and is explicitly motivated by facial detection (identifying if an image contains a face), verification (given a facial image, verifying whether another image matches that facial image), and recognition (identifying if an image matches a face in a database).\(^67\) FW has been expanded into a more significant effort by the name of Labeled Faces in the Wild (LFW). LFW is a subset of the larger Faces in the Wild dataset, the major difference being that the faces in the LFW contain manually annotated names matched to the correct faces, whereas the Faces in the Wild dataset draws probably names from the news captions and assigns a name to each face. Faces in the Wild was developed by computer scientists Tamara L. Berg and David Forsyth while Berg was a graduate student at UC Berkeley. LFW was developed by a team at the University of Massachusetts-Amherst, including Gary B. Huang, Manu Ramesh, Erik Learned-Miller, and Berg. At the time of writing, it appears that FW was quietly taken off the web sometime in late 2020, according to periodic snapshots from the Internet Archive.

FW is drawn from “half a million news pictures and captions from Yahoo News,”\(^68\) while the final dataset contains 30,281 images of faces. The FW website has a small disclaimer on usage (“This dataset is for academic research purposes only”), while LFW contains none. Although not written in the documentation or metadata, the images are downloaded into daily folders from the years 2002 and 2003.\(^69\) Although Yahoo News has produced original content (at one point, journalist Katie Couric hosted a video news program on the website), it has primarily been an aggregator of other news sites, including the


\(^68\) See Berg et al., supra note 67, at 7.

\(^69\) Id.
Associated Press, the BBC, and others.\textsuperscript{70} To that end, FW or LFW do not log information on copyright in any readily accessible metadata. However, we were able to retrieve copyright information by parsing the captions in the FW dataset with a short Python script, given that captions contain copyright and authorship information.\textsuperscript{71} Most of the copyrights are held by Reuters and the Associated Press (over 27 thousand of the 30 thousand), with a small number owned by Agence-France Presse (about one thousand) and even fewer by named individuals with no institutional affiliation.\textsuperscript{72} Moreover, given that these images were taken from news media, no privacy mitigations were taken for those individuals in the photos.

D. Common Crawl: Archiving the Whole Web

The Common Crawl (CC) dataset is one of the most popular datasets used in the training of what have typically been called large language models. The CC dataset has been used for the seed data for OpenAI’s GPT-3 model,\textsuperscript{73} Google’s BERT,\textsuperscript{74} Facebook’s FastText,\textsuperscript{75} CLUECorpus2020,\textsuperscript{76} and XLM-R.\textsuperscript{77} Common Crawl is a San Francisco-based non-profit organization led by technologists who do not have formal academic affiliations.\textsuperscript{78} The Common Crawl organization has constructed the dataset by taking periodic snapshots of the web through

\textsuperscript{70}Id.
\textsuperscript{71}Thank you to Erik Learned-Miller for this information.
\textsuperscript{72}University of Massachusetts Amherst, Labeled Faces in the Wild Dataset (last visited May 18, 2023), http://vis-www.cs.umass.edu/lfw/#explore [https://perma.cc/LG52-2AV7].
web crawling, which they have done since 2008.\textsuperscript{79} Since March 2014, they have released monthly public snapshots of the web.\textsuperscript{80} The January 2022 CC snapshot is composed of 320 terabytes (TB) of data uncompressed, or 73.5 TB compressed.\textsuperscript{81} As a comparison, all of the text in English Wikipedia, the Wikipedia which contains the most articles, is about 20 GB compressed.\textsuperscript{82}

Common Crawl has done very little documenting of the dataset itself, so much of what we know about what’s in the data has come from third-party audits.\textsuperscript{83} report the most common domain name in the dataset from the Google patent database (patents.google.com), followed by English Wikipedia, several publications (New York Times, Los Angeles Times, The Guardian), and PLoS One (an open-source academic journal). They problematize the text which comes from the Google patent database because much of that text has been machine translated and scanned in with optical character recognition, thereby going through multiple stages of machine reading.

Because much of the data is undocumented, it is difficult to identify the licensing regimes of the different types of content in Common Crawl. The Common Crawl organization claims that they obey certain directives written in the robots.txt files of individual sites,\textsuperscript{84} but that says little about the copyright status of the materials collected. Given the heterogeneity of domains in CC, it’s likely that there are a range of licensing regimes attached to each of the archives, from the most restrictive (All Rights Reserved) to the least restrictive (Creative Commons and other open licensing). The analysis in Dodge et al. suggests that there is a large volume of news text in CC, thereby meaning many of the copyright holders are news organizations, similar to the copyright situation in FW. Similarly, given the large number of patents in the dataset, each patent is subject to different countries’ intellectual property regimes.

\footnotesize
\textsuperscript{80} Id.
\textsuperscript{82} enwiki dump progress on 20230201, https://dumps.wikimedia.org/enwiki/20230201/ [https://perma.cc/F8RC-ZCV3].
\textsuperscript{83} Dodge et al. supra note 21.
Moreover, Common Crawl’s lack of documentation means there is no discussion of potential privacy issues. Privacy implications of a text-based LSLD could include details like an individual’s address, email address, phone number, or social insurance or social security number. The LinkRun project\(^{85}\) analyzes the top domains on the web linked through Common Crawl. The top domains are social media sites, including Facebook, Twitter, YouTube, Instagram, LinkedIn, WordPress, and Pinterest. Given that these sites are subject to self-disclosure, but also revenge posting, doxing, and abuse, it is very likely that many acute instances of privacy violations are present through the dataset.\(^{86}\) Moreover, individuals have no meaningful way to consent to be part of this collection.

Most of the secondary analysis on CC has focused on the biases which emerge from the dataset. Dodge et al. note that there is a stark bias against certain ethnicities based on sentiment analysis. “Jews” are most positively viewed in the text while “Arab” is most negatively expressed. Moreover, blocklists that exclude “offensive” language also exclude documents that make mention of sexual orientation, or are African American English or Hispanic-aligned English. Moreover, most of the text comes from English-speaking nations in the West, with little representation from nations with a large English-speaking population in Africa or Asia. Luccioni and Viviano\(^{87}\) analyze the presence of hate speech and sexually explicit content in CC and find that nearly 5% of its content contains hate speech, and anywhere from 1-3% contains sexually explicit content. Analyzing OpenAI WebText, a dataset developed as an alternative to CC “which emphasizes document quality” and is used to train OpenAI’s GPT-2 model, Gehman et al.\(^{88}\) find that much of the data is sourced from the social media site Reddit, which is generally known for its content which is abusive and biased against racial, ethnic, and gender minorities. Moreover, they find that explicitly banned and quarantined subreddits feed 63 thousand messages of OpenAI WebText, 40 thousand from r/The_Donald, and four thousand from r/WhiteRights.


\(^{87}\) Luccioni & Viviano, supra note 21.

E. The Pile: Data Documentation and Safety Considerations

In contrast to CC, researchers at the NLP collective EleutherAI released The Pile, an 800 GB dataset gathered from a set of diverse sources on the internet. EleutherAI is a self-described “decentralized, grassroots collective” of NLP researchers who are explicitly oriented towards open sourcing the datasets related to the building of LLMs. The Pile is advertised as both a benchmark and a possible source for training data. To date, it has been used in training several LLMs, including the open source GPT-Neo and GPT-NeoX, trained by EleutherAI itself.

The Pile stands out uniquely amongst large-scale datasets because the dataset itself is well-documented in both the original paper which discusses its release and its accompanying datasheet. Of all the datasets discussed here, it is the only one which has released explicit data documentation as suggested by most recent data documentation guidelines. The authors of The Pile themselves note that “no dataset intended to train massive language models has been seriously documented by its creators”. This is possibly a function of when these data were released (2020), but we have not seen an effort to produce post hoc datasheets by any of the other projects, which are still actively maintained.

The documentation included in the paper and the datasheet provide information on the component parts of the dataset, which itself is drawn from a cleaned subsample of the Common Crawl, as well as 22 different datasets, ranging from collections of books, academic papers, and abstracts, and computer code. The paper documents how much size each of the component datasets includes, how much they’ve weighted each dataset (e.g., which balances quality of content), language and dialects which are detected within the text, and how much each component diverges from Common Crawl.

The Pile authors explicitly discuss some possible problems arising from copyright, terms of service conflicts, and authorial consent. They claim that the use of the dataset for training and benchmarking machine learning systems falls under the fair use exemption of US

---

92 Gebru et al., supra note 41.
93 Gao et al., supra note 90 at 11.
copyright law and that their use is transformative. Although the state that all the datasets are “public”, insofar as they are openly available on the internet, they also note that 17 of the 22 datasets are in compliance with the Terms of Service of the dataset, and only 12 of the 22 explicitly have authorial consent, typically according to the license (e.g. all of English Wikipedia is licensed as Creative Commons Attribution ShareAlike).\(^\text{94}\) They also acknowledge that certain Terms of Service are terms set by the data host, but not the data subjects or in some cases, not even the authors of the data. For instance, the authors state about the Enron Emails dataset, was “not collected with the permission of the authors, but was collected by the U.S. government as part of a criminal investigation.”\(^\text{95}\) Implications for privacy are similar to those with other LSLDs, including Common Crawl. In the accompanying datasheet, The Pile authors state that they are “not distributing any previously inaccessible confidential data, as all data contained in the Pile was already widely and publicly available on the internet” and that some data may be considered psychologically harmful to readers or some data subjects.\(^\text{96}\)

The Pile authors also document potential bias-based harms which may be present in the dataset. They note that the vast majority of the text is in English (97.4%), which is similar to other LSLDs. Interrogating binary gender pronouns, they find no difference in sentiment bias between men and women, although acknowledge that these biases may be present. In analyzing religion, they find a high co-occurrence of the term “radical” with “Muslim”. “Jew” is reported with the lowest co-occurring sentiment score, which is a marked difference from Dodge et al.’s analysis of CC above. Lastly, focusing on race, they note that the terms “black” + some personal identifier (such as “man”, “woman”, or “person”) have word co-occurrences that seem to imply news reports on police brutality (“unarmed”), protest or rights-based activity (“civil”), or racist connotations (“criminal” and “scary”). “Black” also has the lowest sentiment of all racial groups tested.

In sum, although harms stemming from toxicity, online abuse, and bias have been better formalized in the literature, harms emerging from copyright, privacy violations, and lack of participation in the creation of language technologies are less understood in the context of LSLDs.

The above discussion is admittedly deeply in the weeds of dataset development. We have corresponded (or attempted to

\(^{94}\) However, note, that while licensing may allow for consent for the author, it does not account for the rights of the data subject.

\(^{95}\) Gao et al., supra note 90 at 14.

\(^{96}\) Biderman et al., supra note 91 at 15.
correspond) with all of the authors of these datasets. Some of them have reported that we are the first people to reach out to them to request documentation of the dataset. Much of this information is not written down or formalized as knowledge in any particular way. But we contend as one of the major points of this article, that the devil is in the details. The harms that are consequences of the use of LSCVDs and LSLDs can stem from different points in the dataset creation process, from data collection, data annotation, data distribution, model training, model evaluation, and model inference. However, the majority of the discussion of harms that stem from AI tends to only focus on harms that emerge from inference, e.g., the point at which predictions are made, images are automatically labeled, or new text is generated. In Part II, we focus on developing a taxonomy of the subjects and stages of dataset development to aid a discussion of the potential informational harms that emerge from each step.

II. TAXONOMIES OF DATASET DEVELOPMENT STAGES AND SUBJECTS

This section explores and organizes how we may understand the development of large AI datasets. Much of the focus when it comes to understanding datasets has been on the impact upon deployment, hence the emphasis on biased data and input. However, the other stages in the cycle are also worth consideration. Legal scholars, legislators, and regulators need to understand what happens in what Lehr and Ohm refer to as the “middle stages” of the machine learning model building lifecycle, and ways to look at the process beyond just data collection and model deployment. In order to understand these various other steps, and the nuances of the dataset development process, we need a common language for conversations about datasets. We fill a key gap in this understanding by offering a taxonomy for dataset development.

A. The Stages of Dataset Development

Throughout this paper, we refer to two kinds of large-scale AI datasets: training datasets and benchmark datasets. Unlike in traditional software development, machine learning engineers do not write explicit rules that tell a computer exactly what to do. Rather, they enable a computer to “learn” what to do by discovering patterns in data. The

98 See Bommasani et al., supra note 24.
information used for teaching computers is known as training data. Everything a machine learning model knows about the world comes from the data it is trained on. Consider an engineer who wants to build a system that predicts whether an image contains a cat or a dog. Their cat-detector model is trained only on cat images taken inside homes, the model will have a hard time recognizing cats in other contexts, such as in a yard. Machine learning engineers must constantly evaluate how well a computer has learned to perform a task, which will in turn help them tweak the code in order to make the computer learn better.

To evaluate a model, engineers expose it to another type of data known as testing data. For the cat-detector model, the testing data might consist of both cats and other animals. The model would then be evaluated based on how many of the cats it correctly identified in the dataset. Testing data is critical to understanding how a machine learning system will operate once deployed in the world. However, the evaluation is always limited by the content and structure of the testing data. For example, if there are no images of outdoor cats within the testing data, a cat-detector model might do a good job of recognizing all the cats in the testing data, but still do poorly if deployed in the real world, where cats might be found in all sorts of contexts.

Finally, a benchmark dataset is used to judge how well a computer has learned to perform a task. Benchmarks are special sets of training and testing data that allow engineers to compare their machine learning methods against each other. They are measurement devices that provide an estimate of how well AI software will perform in a real-world setting. Most are circulated publicly, while others are proprietary. Most importantly, benchmarks guide the course of AI development. They are used to establish the dominance of one approach over another, and ultimately influence which methods get utilized in industry settings.

For the purposes of the next few sections, we refer to the following stages of development and usage in a large-scale AI dataset. These stages are drawn from prior typologies of model construction, and are helpful to explain for the purposes of the potential harm which may arise at each stage. We divide these into the following stages: problem formulation; data collection; data cleaning; data annotation; model training; model evaluation; model deployment and inference; and data distribution. These are not strict stages of dataset development, and many of the stages blend into each other. However, this is stylized into

---

a way that is helpful to discuss who can be potentially affected by harms arising at a specific stage, and who should be accountable for the decisions which emerge from these stages.

Problem formulation: Prior to collecting data for a large-scale dataset, dataset curators need to take care to define what the problem they are attempting to define is. Defining the task may emerge from commercial, researcher, or funder-based interests. The definition of the task will then determine what the quality and the nature of the data ought to resemble. While problem definition may seem straightforward, ethnographic, participatory, and other interpretative social science research on machine learning has shown how problem formulation itself is complicated and shifting as the understanding of the problem and the available data and features change in the machine learning development lifecycle.\textsuperscript{100}

Data collection: Data collection is the iterative process of collecting data instances (e.g., individual images, video stills, and/or accompanying metadata or captions, textual documents, code repositories) for the purpose of constructing a large-scale AI dataset. Data collection is typically facilitated through automation via web scraping or using an API provided by a platform or website. This process invariably involves making multiple copies of the collected data. There are more rare cases in which data collection occurs in a more controlled setting, such as taking photos or videos from individuals in a lab environment. For instance, Facebook recently released their Casual Conversations dataset,\textsuperscript{101} in which short videos of individuals were taken in a controlled environment. In that case, many of the issues of consent and rights of the data subject which arise in LSCVD development have been circumvented. However, the large-scale nature of modern models (especially large vision and language models) has necessitated data which operates on the scale of the entire web.\textsuperscript{102}

Data cleaning: Data which is collected from a “raw” data dump is messy. For visual data, many images will not meet the criteria which has come out of the problem formulation stage. Perhaps the dataset


\textsuperscript{101}CASUAL CONVERSATIONS DATASET, \url{https://ai.facebook.com/datasets/casual-conversations-dataset/} (last visited July 31, 2022).

\textsuperscript{102}Alon Halevy, Peter Norvig & Fernando Pereira, \textit{The Unreasonable Effectiveness of Data}, \textit{IEEE INTELLIGENT SYSTEMS} 24:2, 8-12 (2009).
authors were looking for images which only featured clear, unobstructed views of faces. Or they were looking for images which have a primary subject matter and do not have other objects, natural elements, or people in the background. For textual data, much of the text may be surrounded by extra HTML code which needs to be removed. Data cleaning also entails preparing the data for the machine learning process. This can include deduplication of data instances which occur more than once. This is an important step because of what is called “data leakage”, when instances used in the training data appear verbatim in the testing or benchmark data, which can artificially boost the evaluation metrics in the model evaluation stage.

Data annotation: Data annotation is the optional stage of attaching new information to an individual data instance, often called a “label” or an “annotation.” For instance, a curator or a third-party data annotator will label an image of a cat with the label of “cat” for an LSCVD used in object classification or draw a “bounding box” around the cat for an LSCVD used in the task of object localization. Data annotation can also be automated, with “naturally-occurring” labels which can be attached to data instances. For example, for all the Wikipedia entries which are collected from the domain “en.wikipedia.org,” we can assume that all those entries come from English Wikipedia, as opposed to “de.wikipedia.org,” which would come from German Wikipedia. Of the datasets presented above, only ImageNet had humans label the individual images with annotations and draw bounding boxes around the object of interest. LFW used an automated method to find the faces within the image and used the news text to label the individuals within the image. CC and The Pile had no annotations attached to individual data instances, besides metadata.

Model training and valuation: Model training is the act of modifying a machine learning model by being exposed to data instances within the dataset. Models update a group of numerical weights (also called parameters) based on the data which has been seen. Models may be trained on all or part of an LSCVD or LSLD. They are typically only trained on a subset of the data, known as the “training split”, which may be defined by the curator or decided upon by the people creating the new machine learning model. Model evaluation is the act of assessing the accuracy or performance of a machine learning model. Typically, for a benchmark dataset, the machine learning model will be trained on a subset of a dataset, and then will be evaluated on a “held-out” part of the data which the model has not been exposed to before.

Model implementation and inference: After models are developed, trained, and evaluated, they are then used in a production
setting. In a purely research setting, models are often not used in a production environment. However, this is typically the goal in data science and commercial settings. This is the point at which most legal evaluations typically analyze informational harms, for instance, whether inference in a criminal justice setting can cause adverse outcomes which may be in violation of due process.\textsuperscript{103}

Data and representation distribution: This involves the act of posting or hosting a dataset or a downstream pre-trained model/representation more publicly, either on a lab or personal website or institutional repository. Data can be distributed along with a Terms of Service, Release Agreement, or Software License, although these types of agreements are often not sufficient to prevent unauthorized distribution or usage. Copies are also accessible via popular machine learning software packages, such as TensorFlow and PyTorch,\textsuperscript{104} making it trivial for AI researchers to copy and access these datasets.

It is worth defining more clearly what a pre-trained model or representation is. A pre-trained model is a machine learning model, which has been developed with a set of data instances, often from an LSCVD or LSLD, without the original data. Benjamin et al. define a pre-trained representation as "[a] different form, format or model that mimics the effects of given data, but that does not contain any individual data points or allow third parties to infer individual data points with currently existing technology."\textsuperscript{105} Pre-trained models and representations are released because the act of doing model training from scratch can be computationally and environmentally expensive.\textsuperscript{106}

Often, researchers will release the first few layers of a neural network model (known as the convolutional layers) to help with downstream tasks of a larger computer vision model. For instance, the convolutional layers trained on ImageNet feature in a process called


\textsuperscript{105} Benjamin, supra note 99.

\textsuperscript{106} See, e.g., the environmental cost of training large models: Emma Strubell et al., Energy and policy considerations for deep learning in NLP, preprint arXiv:1906.02243 (2019); Carole-Jean Wu et al., Sustainable AI: Environmental Implications, Challenges and Opportunities, 4 PROCEEDINGS OF MACHINE LEARNING AND SYSTEMS, 795–813 (2022).
"transfer learning"\textsuperscript{107} and have been applied in many different domains, in fields as varied as refugee resettlement to the identification of military targets.\textsuperscript{108} For text, pre-trained embeddings have been used to reduce the dimensionality of a large corpus of textual documents to a much smaller one. One prominent example is word2vec, which was trained on a massive internal Google News corpus and released by Google researchers.\textsuperscript{109} This is an important practice to note, as it shows the norms around dataset exchange amongst various commercial and non-commercial entities, as well as the general public.

The release and distribution of datasets and pre-trained models/representations is typically not seen as part of the machine learning pipeline, given that most models in production are not trained on large-scale AI datasets.\textsuperscript{110} Large-scale datasets are unique insofar as they are often used and distributed. Unfortunately, this also means that data which has been shown to be harmful may have afterlives which are beyond the control of the original data curator. When we identify points at which dataset development can become more intentional, we also have to factor in pre-trained models which are often out of reach in terms of transparency and accountability.

Peng et al. tracked three major image datasets to analyze how much of an afterlife these datasets had after they were released.\textsuperscript{111} The datasets – LFW, MS-Celeb-1M, and DukeMTMC – are all datasets which involve images of people. LFW is well-documented above; MS-Celeb-1M is an LSCVD developed by Microsoft for facial analysis tasks. It is constructed of nearly one million images of celebrities drawn from “public” images, whereas DukeMTMC is a dataset constructed of


\textsuperscript{111} Peng et al., \textit{supra} note 17.
non-consensual images taken of individuals walking around the Duke campus. The latter two datasets were retracted by their curators in April 2019 after artists and activists Adam Harvey and Jules LaPlace publicized their existence dedicated to highlighting privacy violations of image datasets.  

Peng et al. found, unfortunately, that these datasets outlived their prior existence, finding derivatives of those datasets (20 of LFW, 8 of MS-Celeb-1M, and 7 of DukeMTMC). They also found several thousand citations of the original datasets and their derivatives in subsequent research papers, evidence of potential production use of both the retracted datasets and their derivatives, and the distribution of several dozen pre-trained models and representations on the web.  

In summary, we need to do closer work on the legal and policy issues that emerge at the different stages of the machine learning development pipeline identified here. Moreover, these issues will implicate different subjects of the dataset development process, to which we now turn our attention.

B. The Subjects of Dataset Development

To clarify the scope of potential harms of large-scale AI datasets, it is necessary to identify different stakeholders impacted in the dataset development process. In the economy of the dataset, we have identified five different stakeholders: the curator, the data annotator, the data subject, the copyright holder, and the model subject. Each of these stakeholders should have a specific set of rights and responsibilities, and a key role to play in the process.

A curator is a person, group of people, or entity involved in collecting a large-scale AI dataset. A data annotator is the person or group of people involved in annotating data. They are often the same as the curators, but as datasets become larger, this work often gets outsourced to remote workers who are paid for annotations in a

---

112 Adam Harvey & Jules LaPlace, EXPOSING.AI, (last visited Aug. 1, 2022) exposing.ai [https://perma.cc/9GPF-KSXW].
113 Peng et al., supra note 17.
114 Scheuerman et al., supra note 20.
115 This is not a fixed category, and we mean it to include a wide range of people involved in the annotation process. See, e.g., Aaron Mok, Kenyan data labelers were paid $2 an hour to label child sexual abuse, bestiality, and other horrific content for ChatGPT creator OpenAI, report says, BUSINESS INSIDER (January 18, 2023), https://www.businessinsider.com/openai-kenyan-contract-workers-label-toxic-content-chatgpt-training-report-2023-1; Khari Johnson, MIT researchers find 'systematic' shortcomings in ImageNet data set, VENTURE BEAT (July 15, 2020), https://venturebeat.com/ai/mit-researchers-find-systematic-shortcomings-in-imagenet-data-set/ (systemic issues with annotating in ImageNet).
piecework fashion. A copyright holder is the person or entity which holds a number of exclusive rights over a work. A data subject is the person whose biometric information is collected/visible in an image. Sometimes a data subject's rights are implicated when a decision is made about them based on an application trained on a large-scale AI dataset. Lastly, the model subject is the person subject to the decisions of a downstream model. We propose this division of roles to distinguish the ways these different actors are impacted in the data pipeline.

Consider two examples of how this may play out in practice, and this can be helpful to articulate all of these subject positions. Say Croydon (the copyright holder) takes a picture of their daughter Sophie (the data subject) at a carnival and uploads the picture to a photo sharing site such as Flickr. Dr. Kurry (the curator), in collaboration with her grad students, scrapes Flickr for images which have any faces in them for assembling a database meant to improve the accuracy of facial analysis tasks, which can include facial recognition and facial detection. To ensure that the images are sufficiently diverse for their purposes, they want to annotate the images with labels like: is the image taken under sunny conditions? Does the individual in the image present as


117 This is Based on what is covered by 17 U.S.C. § 106 – “Exclusive rights in copyrighted works.”

118 This language is meant to mirror the language of the European Union General Data Protection Regulation. See Article 4, Definitions, GDPR. https://gdpr-info.eu/art-4-gdpr/ (Last accessed Aug. 15, 2022).

119 There is no precise language for this individual, some scholars have defined this broadly to also include the data subject. See Alicia Solow-Niederman, Information Privacy and the Inference Economy, 117 N.W. UNIV. L. REV., 357 (2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3921003. Others have highlighted the non-consensual nature of this individual and have called them the “affected individual” or the “affected non-user” (thanks to Deb Raji for this suggestion). In this case, we are using the term “model subject” to differentiate them from the data subject who is represented in the large-scale dataset, and to note that this person is subject to the decisions or output of a machine learning model. There is also the distinction between a knowing user of a model and an unknowing subject of a model who is not a user. However, that distinction is not meaningful for our purposes.

120 Flickr is a photo sharing platform; some users make their works openly licensed and available for download and use. See generally Flickr, https://www.flickr.com/about (last visited May 5, 2023).
more masculine or feminine? What is the skin tone of the individual?\textsuperscript{121} Dr. Kurry’s lab has collected nearly 50,000 images, which would be prohibitively expensive to annotate with the graduate students in the lab, so they outsource it to Amazon Mechanical Turk,\textsuperscript{122} where many Turkers (the data annotators) label the images for a few cents per item. They post the dataset on the web, alongside a model which shows that their facial detection algorithm exceeds the state-of-the-art on facial detection. A year later, a private company SecurFace, uses the dataset to evaluate the fairness of their facial recognition model, and uses this dataset to assure that their technology works well across multiple different genders and skin tones. They sell this technology to the San Junipero Police Department. SJPD, however, uses the technology to wrongfully apprehend Malik (the model subject).\textsuperscript{123} In this case, very real and lasting harm has been affected on Malik. There are a number of pressing concerns here: What kind of a legally permissive environment facilitated all of the actions that led to the creation of the final application? What are the harms (and accountability mechanisms) that may be considered for others in this network of subject positions?

The second example involves a text-based example. Sunipa (the data subject) is a pro-choice activist who has been subject to online harassment and targeting. She gets doxed, and her name, address, and phone number are non-consensually posted to an anti-choice Reddit forum by Courtney.\textsuperscript{124} Sunipa files a complaint to take down the post, which is eventually actioned upon by a Reddit moderator. However, in the interim, the GatherEverything Collective, an informal research group (the data curator), crawls Reddit, amongst many other webpages, to create a broad English-language LSLD. They do this without specific attention to the terms of service agreements on these social media websites. They post this snapshot on the web in perpetuity but do not create a model with it. The data are not annotated in any way, but GatherEverything has performed several rounds of data cleaning. As part of this data cleaning, they use automated methods to attempt to remove private-looking information such as

\begin{itemize}
  \item Annotating for race and gender is problematic, and most facial analysis datasets do not explicitly label for these attributes. See Scheuerman et al., \textit{supra} note 20.
  \item Amazon Mechanical Turk is an outsourcing platform for a variety of tasks. See \textit{generally} Amazon Mechanical Turk, \url{https://www.mturk.com/} (last visited May 5, 2023).
  \item Johnson, \textit{supra} note 2.
  \item \texttt{u/RunDNA}, \url{https://www.reddit.com/r/TheoryOfReddit/comments/3uzsl1/who_owns_the_copyright_on_reddit_comments_and/} (last visited Aug. 1, 2022) (consists of a user analysis of Reddit’s user content, which says that users own content but grants Reddit an expansive license to do what they want with that content.)
\end{itemize}
phone numbers, social security numbers, etc. However, this process is not full-proof and Sunipa’s information remains in the dataset. Later, the for-profit company OpenLLM uses the data to train a massive, generative language model. They do not release the model but allow individuals to access the API endpoint. When a user types Sunipa’s full name into the model, it spits out a phone number and an address.\(^{125}\) In this case, Sunipa is both the data subject and model subject. The harm of Sunipa’s doxxing goes beyond her status as a data subject, but also as a model subject. Again, what facilitated this process, and what types of harm (and accountability) may be considered here?

While these are both stylized examples, they do not differ too greatly from how LSCVDs and LSLDs are collected, distributed, and used by model developers in practice. These hypothetical examples illustrate the ways in which the subjects we identify are interconnected and serve as key stakeholders in the various stages of dataset development that we have identified through our taxonomy. These examples should force us to consider how we view the legal structures around data collection, curation, and annotation. Furthermore, we attempt to illustrate here the challenge of defining what harms occur at each stage, who is impacted by whom, and what the course of action for each subject should be. In the sections below, we elaborate the legal environment in which the dataset development process is taking place. We focus on the overarching frameworks because it situates each stakeholder within broader confines of what society has already deemed permissible/impermissible in information law and policy.

### III. Legal Infrastructure Shaping Dataset Development

Thus far we have discussed what large datasets are most prominent in the current AI landscape. We have proposed a taxonomy to break down the steps involved in the development stage. A framework for dataset accountability also matters for the subjects we identify above who have been impacted in various ways by large-scale AI datasets. This gives us a starting point to discuss current and potential avenues for dataset regulation.

Dataset regulation is important due to the harms caused by the use of automated decision-making systems that could be addressed at

the developmental stage.Datasets are also a useful focal point because there are harms which are emergent from the development process itself that currently lack conceptualization in legal literature.

Many existing and recently introduced laws deal with datasets indirectly. Some laws focus on the model inference stage, by attempting to regulate applications like facial recognition technology and hiring software. Oftentimes, such rules have been devised in response to outcry over abusive use. For example, many laws and ordinances around the country related to facial recognition prohibit specific use by law enforcement agencies, and in general prevent the use from intruding upon residents’ everyday lives. These regulatory responses also place necessary emphasis on individual rights, such as the right to get access, notice-and-consent, and requesting deletion.

There are obligations on companies as well, such as general auditing requirements for the AI systems use, but whether that includes documentation around dataset development is unclear. These regulatory measures are mainly responsive to downstream impacts, and do not specifically address the developmental stage of datasets that led to the creation of problematic technologies.

Therefore, accountability means, in a broad sense, that we should be attentive to the legal and policy environment that leads to the creation of the datasets in the first

---


127 See generally Joel R. Reidenberg, Lex Informatica: the formulation of information policy rules through technology, 76 TEXAS L. REV. 553 (1998) (overview of how code makes important decisions about rights and is embedded in a system’s architecture).


130 Wiggers, supra note 126.
place. This part explores the legal environment which enables the creation of datasets, and the blind spots and novel harms that may be emerging in the process. We specifically examine current understandings of copyright and privacy doctrine that are shaping how datasets are seen in the regulatory sphere.

There are two aspects of legal and policy environment that enable dataset creation that we endeavor to examine here. The first is an exploration of how copyright law facilitates the developmental stage where data is being collected, curated, annotated, and processed. Here, we talk about relevant copyright law because the current legal regime has the potential to influence how the development stage takes place. The second aspect looks at both the individual and systemic informational harms that may arise in the process. We discuss these aspects together as a way to holistically understand the dataset development process. Our taxonomy of the dataset development process can help identify the stages at which policymakers can intervene. Our aim is to make the dataset development process more legible, such that we can conceptualize the individual and societal harms associated with large datasets.

A. Copyright and Fair Use

Data collection practices have been a key contention in discussions about how copyright law applies to dataset development. Data is being collected in large volumes from a variety of public and private sites and usually scraped from a variety of sources including free knowledge projects like Wikipedia, open-source libraries, and public domain works. At the data collection stage, the most obvious stakeholder is the copyright holder. And much of the discussion on data collection practices has focused on the impact on the copyright holder, and more recently the data subject. We bring attention to the impact on the model subject as well, since they are usually at the receiving end of the downstream AI application. In this section, we explore the connection between the copyright holder, data subject, and the model subject. We create a picture of the ways they are impacting one another,

---

131 Barocas & Selbst, supra note 16; Lehr & Ohm, supra note 97.
132 Clearview AI has raised questions about privacy rights of data subjects, see Hill supra note 3.
133 We coin the term "model subject" to refer to a category of persons who are impacted by automated decision-making systems. See also Kaminski & Urban, supra note 30, at 1957 (conceptually we rely on scholars discussing individual rights of action and we propose for these individuals to be called "model subject" when they are specifically contesting datasets).
and how data collection practices buoyed by current understandings of copyright law implicate these actors.

At a fundamental level, copyright law prevents copying and dissemination of works without permission from the copyright holder, and the copyright holder may sue for infringement when their copyrights are infringed upon.\textsuperscript{134} In the United States, a copyright holder has a number of exclusive rights such as the right to reproduction and the right to create derivative works of a protected work.\textsuperscript{135} However, not all copying is automatically wrong and the development of new technologies has always challenged the scope of copyright protection. This means the copyright holder implicated in dataset development does not necessarily have a cause of action worth pursuing.

The law around the extent of copying allowed for machine learning is in flux. Courts have dealt with various technological developments over the last few decades, such as peer-to-peer file sharing,\textsuperscript{136} time-shifting technologies,\textsuperscript{137} and the functions of search engines in displaying and disseminating snippets of copyrighted images, sharing excerpts of books,\textsuperscript{138} but have yet to deal with the use of a large amount of copyrighted works to build datasets.\textsuperscript{139} These cases have all addressed questions such as whether making copies at some stage of the technology’s function was legitimate. For example, peer-to-peer sharing platforms allowed users to download and store copies of copyrighted works on their personal hard drives.\textsuperscript{140} With time-shifting technologies, courts have dealt with the question of whether making a temporary copy to watch later was legitimate. In the case of search engines, the courts considered whether thumbnail copies of an image on a third-party site constituted infringement.\textsuperscript{141}

There are features unique to the data collection process that are testing the boundaries of existing copyright law. The process of building and training large-scale AI datasets involves decisions made by curators to use or not use copyrighted works without permission. These decisions

\textsuperscript{134} The law defines a “copy” as “material objects . . . in which a work is fixed by any method now known or later developed, and from which the work can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device.” See 17 U.S.C. § 101 (2012).

\textsuperscript{135} 17 U.S.C. § 106.

\textsuperscript{136} A&M Records, Inc. v. Napster, Inc., 239 F.3d 1004 (9th Cir. 2001).


\textsuperscript{138} Authors Guild, supra note 60.

\textsuperscript{139} Perfect 10, supra note 60.

\textsuperscript{140} See, e.g., MGM Studios, Inc. v. Grokster, Ltd., 545 U.S. 913 (2005); A&M Records, Inc., 239 F.3d 1004 (9th Cir. 2001)

\textsuperscript{141} Authors Guild, supra note 60.
are shaped by a combination of awareness about copyright rules, website and platform terms of service agreements, and a set of institutional and research practices established by curators. The large volume of copying is also unprecedented, as well as the purpose for which the copying is taking place.

Copyright law allows unauthorized copying through the fair use exception, but the determination is done on a case-by-case basis. There is a doctrinal uncertainty in copyright law when it comes to determining the legal basis of copying for machine learning from millions of works without permission from copyright owners. Furthermore, the data constituting large-scale datasets includes not just one-time unauthorized copying but the replication, transformation, and appropriation of those copies. This is different from past instances of copying considered by courts, because the use for machine learning is much more multifaceted in terms of character and purpose. In training for instance, the copy is stored, used multiple times, and distributed. The character and purpose of the use of this data and the lack of clear determinations by courts makes a decisive fair use assessment difficult. Lack of certainty around fair use also means that it is not clear how data collection may take place, and how copyright holders, data subjects, and model subjects may seek potential recourse.

This is why scholars have weighed in on whether copying copyrighted works to build machine learning applications is and should be considered fair use. We consider their perspectives on fair use in machine learning and delineate two emerging themes. These thematic perspectives on fair use can be roughly considered from least restrictive to most restrictive with regard to what falls under the fair use exception.

1. Fair Learning

Mark Lemley and Bryan Casey introduce the concept of “fair learning,” which holds that copying in data collection should be considered fair use in machine learning contexts. Fair learning posits that as long as the purpose of copying works is to facilitate the training of machine learning models, there should be a doctrinal presumption for

---

142 17 U.S.C § 107 (courts apply a four-factor test to assess whether a use is fair, including looking at the 1) purpose and character of the use, 2) nature of the copyrighted work, 3) the amount or portion used, and 4) the effect on the potential market.)
143 Lemley & Casey, supra note 25 at 4.
144 See Perfect 10, supra note 58 Thumbnail copying is not the same as machine learning copying.
145 General here does not refer to a legal term but a normative position.
that use to be fair. According to the authors, unauthorized copying for the purposes of learning from the data and not using it for the same expressive purpose as the original author may be allowed under current copyright doctrine. This is based on what is known as the transformativeness standard within fair use. Transformative use excuses unauthorized copying when it is done for a different purpose from the original work. For example, a parody of a book or film is potentially a transformative use. This idea goes to the heart of what copyright exceptions such as fair use allow: to let individuals use the work for its underlying idea or facts. Researchers working on the technical aspects of this issue think similarly. The Stanford Foundational Models report notes how issues with training models may be considered transformative enough to fall under fair use exceptions; it acknowledges that this is a gray area. Most aggressively, policy researchers in the AI startup space have argued that outputs from generative models should be considered transformative.

The basic argument is that accessing the "facts" of an image of a cat in order to learn more about its features is not the same purpose as the photographer copying a photo of a cat for creative or personal purposes. A machine learning model may learn certain features about the image of the cat, or many images of cats, such as the pointy-ness of their ears or the shape of their irises, for the purpose of classifying images of cats. Lemley and Casey also note that the emphasis on purpose can also be used to find that a use is not fair. For example, something like copying an Ariana Grande song for the purpose of creating a machine learning system which generates songs in a similar style may make fair use exceptions more difficult to justify, but in general, such potential uses should not prevent copying audio files.

In addition to emphasizing purpose, the authors argue that there are policy reasons for a finding of fair use when it comes to collecting data for machine learning. First, fair use allows for broader access and

146 See Lemley & Casey, supra note 25.
147 Id. at 26.
150 See Lemley & Casey, supra note 25 at 6.
151 See Bommasani et al., supra note 24 at 145.
153 Lemley & Casey supra note 25 at 746.
154 Id. at 8.
scrutiny of the data that constitutes machine learning applications, for better research, and for continuous improvement of machine learning results. Second, it is impractical to license millions of individual works; a potential market for licensing for copyright holders to benefit from should not be reason enough to discount a fair use finding. Third, including more copyrighted works in the datasets would help reduce bias that stems from solely using works in the public domain or which have been licensed with the express purpose of creating datasets. Overall, applying the concept of fair learning to cases questioning whether machine learning is fair use would lend much needed clarity to dataset curators to create datasets without worrying about copyright infringement.

These are compelling policy reasons to endorse a “fair learning” presumption for machine learning. However, we highlight a few assumptions made about the dataset development process that should lead us to pause before broadly applying a presumption of fair learning.

First, a key caveat is that the purpose for copying in order to build large datasets is not easily discernible and may not be as easily categorized. As we have noted above, there are subtle differences in the ways benchmark and training datasets are constituted. It is critical to note that machine learning cannot be described as a single step process. Initially, at the problem formulation stage, and even before collecting data, dataset curators need to clearly define the problem they are attempting to solve through the model. The task at hand will determine the quality, nature, and source of the material copies — and therefore, “purpose”. Furthermore, different kinds of “learning” occur at multiple points in the process, such as experimenting with the data, evaluating the model at multiple different points, and lastly, training the model to be used in production and testing for issues like accuracy and fairness.

Therefore, the overarching purpose of copying may be “machine learning”, but there are substantive decisions made in the process that should be critical to determining whether a use is fair. Returning to the example of a model built on Ariana Grande’s music, there would be points in the process, such as during data collection and model deployment, which would raise important questions about the purpose.

155 Id. at 753.
156 Id. at 759.
157 Id. at 771.
of copying. For instance, is the copier a commercial or non-commercial entity?\textsuperscript{159} Are the music files being used for a classificatory or generative purpose, meaning would they be used to identify Ariana Grande's music, or also enable the creation and implementation of a generative model that would allow the curator to develop songs like hers without credit?

Second, there is an increasingly blurred line between commercial and non-commercial use when it comes to machine learning. Benchmark datasets may start as an academic endeavor to advance a narrow task, but then later may be used in an unknown number of applications, across different national jurisdictions with different copyright regimes. For instance, Peng et al. found that three datasets – Labeled Faces in the Wild (LFW, discussed above), MS-Celeb-1M, and Duke MTMC – were developed with academic purposes in mind, but had afterlives which could not be accounted for.\textsuperscript{160} These afterlives become relevant when benchmarks are used to create harmful technologies, such as tools for mass surveillance.\textsuperscript{161} In response, the retraction of a model, especially by a commercial entity, adds further confusion to the overall purpose of the use. Microsoft retracted MS-Celeb-1M, offering vague reasons for doing so. Even worse, Duke MTMC was taken down because the data had been collected non-consensually by taking still images from people on the Duke University campus.\textsuperscript{162}

This is important to note because even if creators try to limit the purpose of use to something that may be fair under copyright law, it is difficult to implement these terms in practice. Commercial entities may scramble to retract or delete models after a controversy, even though they used datasets that were created for non-commercial purposes. A further complication in this ecosystem comes from the introduction of various terms of service and licenses that are applied to datasets in both commercial and non-commercial settings. These details about the commercial or non-commercial nature of copying are important when deciding whether we have achieved the balance for public interest, which lies at the core of fair use doctrine.

\textsuperscript{159} There is no negative presumption against commercial copying in fair use doctrine, but a purely commercial endeavor may weigh the fourth factor, market effect, in favor of the copyright holder. Furthermore, Matthew Sag’s empirical research has found that commercial use makes finding fair use less likely. See Matthew Sag, \textit{Predicting Fair Use}, 73 Ohio St. L. J. 47, 58 (2012); See, \textit{e.g.}, American Geophysical Union v. Texaco Corp., 60 F.3d 913 (2nd Cir. 1994).

\textsuperscript{160} Peng et al., \textit{supra} note 17.

\textsuperscript{161} Hill, \textit{supra} note 3.

\textsuperscript{162} Harvey & LaPlace, \textit{supra} note 31.
Third, having a broad presumption of fair use for machine learning does not consider in enough detail the kinds of stakeholders involved in the process, such as data collectors and curators. These stakeholders make critical decisions such as what kind of data to collect, what the sources of that data may be, what license terms to respect and subsequently apply to the datasets created by them. These decisions are built into the process of dataset development, and there needs to be a closer examination of the practices and motivations of these key actors. Interestingly, early instances of dataset curators working on face recognition datasets paid attention to relevant copyright and privacy issues, and sought permission from the data subjects in photographs, and who were also sometimes compensated for their participation. Whether this is possible at the scale on which current datasets are created is questionable, and as researchers have noted, this level of carefulness has faded with increased data scraping. An examination of the stakeholders thus would enable better understanding and scrutiny of how and why a dataset is developed, deployed, and used down the line.

Thinking about the subjects would also help policymakers devise the responsibilities that curators may have towards data and model subjects. Presently, the norm is to scrape data from the web, but simply because information about a data subject is publicly available online does not necessarily mean that it should be used to train large datasets. Oftentimes, personal information about data subjects ends up in these datasets.

Julie Cohen calls personally identifiable information as being part of the “biopolitical public domain: a repository of materials that are there for the taking and that are framed as inputs to particular types of productive activity.” Information about a data subject is treated as “raw” data collected to train an AI system. And we argue that treating biometric information being a part of the “biopolitical public domain” may be entrenched by an expansion of fair use that does not include an undertaking of a dataset curator’s role. If it is fair use for curators to not require permission from copyright holders, they can continue scraping thousands of images without necessarily doing the work of identifying the impact on data and model subjects. The scope of using publicly

---

165 Id. at 1 (Cohen refers to the repository of public domain information as “raw material”).
available information was tested in the lawsuit against Clearview AI, a company which has come under fire for training facial recognition on millions of images without consent and then selling surveillance technologies to law enforcement agencies.\textsuperscript{166} Giving dataset development the cover of fair use is necessary in some ways, but not enough to address the ethical and legal questions that arise from the applications of such datasets. A general fair use presumption also needs to incorporate the distinctions between different kinds of data, such as whether those data are code, faces, objects, or natural language text—since that would influence the first factor, the character and purpose of copying.

Finally, there is an assumption that copying for machine learning can generally be considered non-expressive use. A part of the fair use analysis is considering whether that use is expressive or non-expressive.\textsuperscript{167} Non-expressive use refers to copying that is not "intended to enable human enjoyment, appreciation, or comprehension of the copied expression as expression."\textsuperscript{168} A use is non-expressive in the machine learning context when the copies are being made to extract meta-level information, such as patterns in speech or language from a collection of copies of novels. Matthew Sag argues that text and data mining is non-expressive use\textsuperscript{169} and therefore making copies for text and data mining should not conflict a copyright owner's exclusive rights.\textsuperscript{170}

Conversely, Benjamin Sobel sheds light on the blurred line between expressive and non-expressive use in machine learning contexts. He argues that it is not a straightforward determination to deem all machine learning as non-expressive use.\textsuperscript{171} Non-expressive use has historically been applied to assess the fairness of previous instances

\textsuperscript{166} Clearview AI ultimately settled the lawsuit with the ACLU. See Hill, supra note 3. The settlement included terms that would protect data subjects in Illinois by giving them the option to "opt-out" but did not address deeper questions of access and public scraping. See ACLU of Illinois, In Big Win, Settlement Ensures Clearview AI Complies With Groundbreaking Illinois Biometric Privacy Law, ACLU (May 9, 2022). https://www.aclu.org/press-releases/big-win-settlement-ensures-clearview-ai-complies-with-groundbreaking-illinois.

\textsuperscript{167} Benjamin L. W. Sobel, Artificial Intelligence's Fair Use Crisis, 41 COLUM. J. L. & ARTS 45 (2017).


\textsuperscript{169} Id.

\textsuperscript{170} Id. at 302. We use the term "raw", even though no data is ever really "raw": it is the product of many different decisions and choices. See Lisa Gitelman et al., Raw Data Is An Oxymoron, (2013).

\textsuperscript{171} Sobel, supra note 167 at 51.
of unauthorized copying by computers. Sobel states that this is premised on two facts: that "the mechanical ingestion of works is a non-expressive purpose, provided it is not to facilitate human engagement with the works' expression" and second that "these uses do not affect works' potential markets [of the work] ... because copyright owners' entitlements does not encompass the non-expressive components of their works – the very components with which computerized analysis engages, and from which it can derive value." In other words, the act of ingestion/data collection itself does not entail an expressive purpose, and the post-ingestion analysis is where the value-add occurs and we may consider factors such as transformativeness and market substitution.

The line between expressive and non-expressive use is now increasingly blurred primarily because of the generative potential of large datasets. Fair use may not protect expressive machine learning applications that have a similar purpose to the original work, such as the creation of new artworks or texts, which mimic the style of the data which the model ingests or model outputs that create works similar to a copyrighted work and help "reconstruct idiosyncrasies of input data" – just as we alluded to in the example of a generative model that would create music like Ariana Grande's songs. On Sobel’s second point, this very potential to generate expressive works means that the fourth fair use factor, the effect on the market for the original work, may not weigh in favor of finding fair use. Whenever there is a threat of market substitution, copying for machine learning becomes more difficult to justify. This applies to dataset development as a whole since the resulting dataset could very well have expressive value that could threaten market substitution for the original work. Indeed, there is already a recognition of the market for user data, as exemplified by the terms of service agreements of platforms like Facebook that users agree to license their data to. Due to widespread use of user data, the large size of the dataset, and the infinite number of models that could be trained on a dataset, having a presumption that the creation of such a dataset is non-expressive does not seem accurate.

172 Id. at 51-57 (discussion on non-expressive use).
173 Id. at 12.
174 Id. at 18.
175 Id. at 20.
176 Id. at 12.
177 Id. at 29; see, e.g., Terms of Service, FACEBOOK, https://www.facebook.com/legal/terms (last visited Aug. 1, 2022.)
In sum, we have considered the most expansive interpretation of fair use in machine learning contexts, and offered nuances to it based on the dataset development process. Our goal here is to shed light on the tensions and questions that remain unanswered when we think about how such a presumption would apply on the ground. These tensions should enable us to examine more closely the environment which allows dataset creation in the first place.

2. Fair Use as a Corrective Measure

A second, and more restrictive perspective on fair use applicability in machine learning is the idea that certain machine learning applications should not be considered fair use at all because they lead to harmful impacts such as facial recognition, emotion detection, and other invasive types of surveillance, which run counter to the public interest. This perspective also looks at the purpose of copying but reaches a different conclusion, finding that fair use may instead be seen as a corrective tool for inappropriate copying.

Amanda Levendowski notes that a reason that companies like Clearview AI that are building face surveillance technologies have not been overwhelmed with copyright infringement claims is fair use. And as we have noted above, data mining practices have largely been seen as fair use. She argues instead that it may not be fair use to make copies to build the dataset used in this application because the final product weighs against public interest, which is what fair use intends to uphold. She engages in a fair use analysis of the copying carried out by Clearview AI and finds that it may not be fair. On the first factor looking at the purpose of copying, she characterizes the use by Clearview AI as the creation of a massive database of faces collected without consent, and sold to law enforcement without oversight, as having no clear public benefit. She distinguishes this use from past instances of mass copying where the public interest was clear, and the courts found fair use, such as the creation of a searchable database of books for the public’s benefit, and the ability to search for images on the internet. Furthermore, she notes that fair use exists so that it can

179 Id.
180 Id.
181 Id. at 37.
promote public good when a new technology enables the public to make reasonable use of the works, and that face surveillance technology does no such thing.\textsuperscript{183} In part because of this emphasis on the public interest motivations behind determining the purpose of copying, Levendowski argues that fair use can be used as a mechanism to curb copying that allows dataset curators to build harmful technologies.\textsuperscript{184}

This is a novel and creative application of fair use analysis. However, it is narrow in its application because it depends on the obviously egregious and unaccountable use of non-consented images by a corporation. Fair use has the potential to be a corrective tool, but dataset development is more than facial recognition datasets, and may present use cases that are not always as controversial. It is useful to consider fair use's inherent flexibility in a situation where regulation is slow to respond otherwise. In such a case, fair use helps us recognize that the dataset development process needs more attention in the early stages than is currently being given.

Another way fair use can be used as a corrective measure is when it can be used to address issues of bias in datasets. Levendowski has previously argued that training AI models on copyrighted work should be considered fair use and can be a method of mitigating bias.\textsuperscript{185} She contends that many works that are under open license and in the public domain can be the source of embedded biases and misrepresentations in a dataset.\textsuperscript{186} Instead, if engineers are able to access more diverse images and text sources, even if they are copyrighted, the assumption is that the models they build will also be able to make inferences which do not impute undue bias.\textsuperscript{187} Levendowski distinguishes between different kinds of data, and argues that public domain datasets like Enron and representations like word2vec are "low-friction" because they allow AI researchers without access to large troves of data to perform machine learning inference easily, or to train their models on existing, public domain datasets. Public domain data

\textsuperscript{183} Levendowski, supra note 178.
\textsuperscript{184} Interestingly, Sobel notes that fair use is being used to shift benefit away from the public and towards powerful companies instead. Sobel, supra note 167, at 37.
\textsuperscript{186} Id.
may contain demonstrable biases, especially given that many works enter into the public domain after their copyrights have expired after 95 years. Established biases have also been shown to exist in downstream visual and linguistic representations.

However, it is not clear that there is much separation between what Levendowski calls “biased, low-friction data” and the data constituting some of the large datasets we describe here. Ostensibly, Google does not own the copyright on Google News stories. Like Yahoo News, they are an aggregator and produce no new net content. The same can be said of MS-COCO and ImageNet, whose photos are drawn from Google Image Search and Flickr. These LSCVDs are built on an unknown combination of copyrighted, public domain, and Creative Commons works. Existing LSCVDs already contain an unknown mix of what may be “biased low-friction data”, and LSCVDs may not necessarily be improved even if copyrighted works are added to the dataset. Furthermore, representations have been created from all three of these datasets, but none of the organizations which created those representations are the owners of the copyrighted works used in that development. This complicates the issue of identifying representations that may result in bias, and the copyright status of the underlying works.

Accessing copyright work may also be necessary for the purpose of auditing, testing, and mitigating bias in datasets. Copyright limits access to training data and confines curators to using either public domain data -- such as the Enron email dataset, which was released by the Federal Energy Regulatory Commission after the high-profile investigation of the company after its mass fraud and corruption case -- or representations obtained from large-scale, internal datasets, such as word2vec, which was trained on a massive internal Google News corpus.
and released by Google researchers.\textsuperscript{194} Here, it may be useful to rely on the flexibility of fair use, and support access for researchers and auditors. For dataset development however, the connection between copyrighted works and reduced bias is tenuous and needs further investigation of the sources of data that dataset curators are using.

Finally, thinking of data in terms of biased/unbiased overlooks the underlying narratives around the sources from which it is collected. For example, GPT-3 has been criticized for generating hateful and Islamophobic text.\textsuperscript{195} If the language model has been trained on media or literary sources consistently espousing a certain view about a religion or ethnicity, this may still be "good" data because it is representative of larger discourses. For instance, mainstream news media and entertainment in the West have had well-documented patterns of hostility towards Arabs and Muslims.\textsuperscript{196} Therefore, relying on this "good" data still leads to biased representations about a marginalized community. Using copyrighted works does not guarantee that the narratives embedded in a large dataset will be fair and balanced.\textsuperscript{197}

In sum, the inherent flexibility of fair use can both be a reason for pause in dataset development, and a reason to hope for the mitigation of some of the impacts of the process. In either case, there needs to be a deeper incorporation of the nuances of the stages of dataset development, and a consideration of the data and model subjects. Fair use as a corrective measure does not have the ability to capture the question of the rights of model subjects, which should be a key goal in any regulatory intervention. However, these two perspectives on the application of fair use constitute the broader legal environment in which dataset creation is taking place and should be a starting point for any accountability frameworks and debates.

\textsuperscript{194} See Mikolov et al., supra note 109.
B. Privacy Harms

Many of the controversies surrounding datasets involve appropriation or misuse of personal information, offending existing legal conceptions of privacy. In this section, we discuss how these privacy frameworks can help us identify what dataset development practices implicate privacy. We examine the ways harms arise in both the individual, but also the aggregate and collective sense. Because the focus on dataset development is novel, existing notions of privacy do not address datasets directly. Thus, we expand into what interests matter in dataset development, whose interests they are, and why they are relevant for privacy doctrine as a whole.

1. Individual Privacy and Dataset Development

Individual-focused conceptions of privacy harm feature primarily in current legal regimes in the US. An individual harm occurs when information is used or produced to negatively impact a data subject. In this section, we identify how some individual privacy interests apply to dataset development, and where these conceptions fall short of capturing the full extent of privacy violations in the process.

We start by considering the defining characteristics of what a privacy harm is. Ryan Calo expands on privacy harms being a unique type of injury. He attempts to define the boundaries and characteristics of such actions so that it is easier to fit newer harms based on the inner and outer bounds of a basic offense. Calo argues that there are two main categories of privacy harms: subjective and objective. Subjective harms are those where there is a perception of unwanted observation and that causes unwelcome emotional states. Objective

---

198 See Salome Viljoen, Relational Theory of Data Governance, 131 YALE L. J. 573, 592-596 (2021). Viljoen argues that privacy law is focused on the individual, with many laws providing individual-centered rights like notice, consent, and deletion. Bolstering her argument is other scholarship that shows corporation direct their efforts towards the individual and not groups impacted by their privacy practices; See, e.g., Neil Richards & Woodrow Hartzog, A Duty of Loyalty for Privacy Law, 99 WASH. U. L. REV. 10 (businesses rely on “notice” and “choice” to exploit data); Daniel J. Solove, Privacy Self-Management and the Consent Dilemma, 126 HARV. L. REV. 1880, 1883-85 (2013); Woodrow Hartzog & Neil Richards, Privacy’s Trust Gap: A Review, 126 YALE L.J. 1180, 1197-98 (2017).


201 Id at 1133.

202 Id.
harms are those that include the "unanticipated or coerced use of information concerning a person against that person." For example, using a person's personal information to send them spam is an objective harm. Ultimately, both kinds of harms relate to some form of loss of control over one's information.

An interesting feature of subjective harms is that they can occur even in a systematic manner that is part of a plan or pattern, for instance, persistent surveillance of individuals in a city with expansive surveillance infrastructure is a subjective harm. This conception is important when we consider dataset development as a systematic strategy of ingesting personal information, and the ways in which dataset construction practices can be seen as a pervasive source of subjective harm. If we normalize mass data collection to the extent that individuals expect their personal information to feature in large datasets, this would significantly alter notions of subjective privacy harms, and the ways to address them. Furthermore, both computer vision and language datasets often contain sensitive personal information. During the dataset development process, this personally identifiable information can be sought out from within the millions of data points and that can lead to objective harm such as identity theft, misidentification, or biased representations.

Calo's approach to privacy harms attempts to capture the absence of a human actor involved in a process for there to be a privacy harm, which he argues is not necessary. He notes that the machines involved in data processing are capable of acting upon private information in harmful ways without there ever being a human being that sees it. This may be the case when it comes to data processing and model building based on the data annotation steps that we have identified in this paper. During data annotation, the arbitrary decisions of annotators can lead to categories and representations that may cause downstream impact such as misidentification in facial recognition, or denial or jobs and social services. These harms to an individual can occur without there being a human involved in the decision process, since many of these algorithmic systems are deemed to be black boxes. As automated systems become ubiquitous, subjective notions

---

203 Id. at 1133.
204 Id. at 1145.
205 Id. at 1154.
206 Id. at 1159.
207 See generally Scheuerman et al., supra note 20. (discussing facial recognition mislabeling leads to technologies unable to recognize nonbinary persons).
of privacy harm may have to be expanded to include the stress, anxiety,
and the anticipation of loss that may come from critical decisions being
made based on a large dataset. Similarly, these systems may cause
objective harm because of underlying issues in the model and the social
and economic contexts in which they are deployed.

Since automation is ever-increasing, it is likely that many
objective and subjective harms might be occurring without there
necessarily being a human involved. 209 This conception of privacy harm
is useful for dataset development. This is because conceptions of
privacy harm based on the two categories do not have to be confined to
harms that can only be anticipated beforehand and then included in any
given privacy notice or regulatory intervention. We can instead focus
on whether the overarching practices of collecting data and processing
it, and the resulting models and applications cause subjective or
objective harm to people.

Expanding on how individual harms arise in dataset
development, we look towards some basic categories. A basic
individual privacy harm is when information is collected without the
consent of an individual. 210 Many privacy laws have attempted to
mitigate this by instituting notice and consent requirements, attempting
to give data subjects more control over what information they hand over
to platforms. 211 But a simple notice-and-consent regime falls short in the
context of dataset development for a few reasons. Typically, the notice-
and-consent system is based on “opt-in” by a data subject. 212 But this
system is passive in that it does not give meaningful choice or
information about how the data is going to be used to either data subjects
whose images are being collected and used, and the model subjects of
the applications trained on the large datasets. Opting-out is difficult,
especially when it comes to dataset development, where data is scraped
in large troves and often without much regard to the source. Furthermore,
a notice-and-consent regime would be challenging to
implement for dataset development, because it must be decided when
data subjects or model subjects may have the choice to opt out – should
there be an opt-out at every stage of data collection and processing, or

209 Calo, supra note 200, at 1154.
210 Viljoen, supra note 199, at 595.
211 See, e.g., CCPA, CAL. CODE 1.81.5. § 1798.100 - 1798.199.100 (2018). California
Privacy Rights Act requires notice and consent; GDPR requires notice for six purposes
detailed under Article 6.
212 Solon Barocas & Helen Nissenbaum, On Notice: The Trouble with Notice and
MGMT. OF PERS. ELEC. INFO. (2009).
only at the beginning? And how can we develop meaningful notice-and-consent for model subjects when the stages of dataset development before deployment are largely opaque? Hanely et al. note that in their overview of the dataset creation process and its ethical issues, and state that a subject cannot consent to every possible use of a dataset and should be informed about the scope of the use of the dataset, which would likely have practical challenges.213

There are many datasets which contain images taken by dataset curators without consent of data subjects, as was the case with the UnConstrained College Students Dataset, in which photos were taken by University of Colorado-Colorado Spring researchers using a “long-range high-resolution surveillance camera without their knowledge,”214 which creator Professor Terry Boult claims to be legal according to Colorado state law.215 The MegaPixel project216 documents many of the most egregious of these, including the Brainwash dataset217 (taken from CCTV camera footage in a San Francisco café), Duke MTMC218 (students photographed on the Duke University campus), and Oxford Town Centre219 (CCTV in a downtown area of Oxford, UK). The Microsoft’s MS-Celeb-1M dataset was gathered in much the same manner of Faces in the Wild, with additional images from Flickr.220 Users who upload their images to Flickr have the option to use Creative Commons licenses but it is not clear how many of the collected images were under a CC license.221 Both of these latter projects have been decommissioned after intense scrutiny by Harvey and Jules, and subsequent news reporting from the Financial Times.222 The scrutiny

213 See Margot Hanley et al., An Ethical Highlighter for People-Centric Dataset Creation, NEURIPS (2020).
214 Harvey & LaPlace, supra note 112.
217 Id.
218 Id.
219 Id.
221 Due to the dataset being comprised of millions of images, there is no record available for what number of images were copyrighted, open licensed, or in the public domain.
222 Murgia, supra note 215.
that comes after such revelations shows that there is an incongruence between the efficacy of notice-and-consent requirements, and the actual mitigation of harms to individual data and model subjects. These examples illustrate the difficulty of designing a consent regime and identifying the scope for which the dataset is being created.

Another basic conception of an individual’s harm is the reidentification of an individual’s private data. This is a pressing concern acknowledged by existing privacy laws and will also be an issue in dataset development. Reidentification occurs when personal data is released due to a leak, or in case of datasets, through reverse engineering processes that reveal the identifying information of data subjects. There are laws that protect against the harms of reidentification and related consequences such as breach of medical or financial confidentiality. However, there is a unique reidentification concern in dataset development. Reidentification can occur when automated tools, such as language or code generating models are released for public use. This concern arose when Github’s Copilot was released to the public. The criticism was that a tool built on publicly available code may introduce unknown vulnerabilities into systems built based on it, and also allow an innumerable people access to personal data through such bugs.

Finally, inaccurate representation and discrimination can be an individual harm. This occurs when data-driven tools make decisions about individuals that may result in them losing opportunities or access to critical social services. Large datasets can be the driving force behind such an individual harm. This is because the datasets used to build applications may contain biased representations against individuals belonging to a certain race, class or ethnic background.

225 A study showed that language model may leak private information like addresses, identities, etc. Nicholas Carlini, Privacy Considerations in Large Language Models, GOOGLE RESEARCH (Dec. 15, 2020), https://ai.googleblog.com/2020/12/privacy-considerations-in-large.html [https://perma.cc/4BD2-D4UJ].
Thus, there is a need to specifically recognize the role that data collection, annotation and model deployment processes can play in exacerbating this particular individual harm. Here, the concept of "fusion harms" can be a useful construct. Solove and Citron discuss "fusion harms" that occur when two kinds of harms to an individual are occurring due to the same action. For example, there can be both a physical and psychological harm when doxing takes place.

Further, there are other examples of fusion harms that affect both individuals and marginalized populations: We can characterize the arrest of Robert Julian-Borchak William mentioned in the first part of this article as a fusion harm, in which both discrimination and autonomy harms occurred due to his wrongful arrest. Misrepresentation of gender can be a discrimination harm when an individual who has been misgendered by an AI application is denied entry into a public or private space based on gender identity. They may also suffer an autonomy harm since the AI application denies them the dignity of living true to their identity. Furthermore, if a dataset creates unfair associations between a class of job candidates and their eligibility for a job, these individuals may suffer both economic and discrimination harms since they would be denied economic opportunities based on them belonging to a marginalized group.

Lastly, we contend that when it comes to conceptualizing privacy harms, data subjects and model subjects should be seen as having distinct interests. An individual becomes a model subject when they are subject to the decisions made by an AI system. We contend that this individual is distinct from the data subject because the data subject is a person about whom data is collected, often to build the large dataset. Therefore, conceptualizing individual harms in the dataset development process by identifying the distinct individuals involved can help us expand our understanding of when and how a harm occurs.

2. Systemic Privacy in Dataset Development

---

229 Citron & Solove, supra note 16, at 834.
230 See id.
Thus far we have discussed how existing conceptions of individual privacy harm map onto the dataset development process. Datasets are constructed from millions of individual data points, and these data points form the backbone of a broader infrastructure in which AI technologies are developed and deployed. A framework for dataset accountability would be incomplete without a consideration of the systemic effects of large datasets.

A notion of group harm already exists in privacy scholarship. Group harms are useful as a separate category when an injury may appear too small to be actionable when viewed in isolation, but detrimental when viewed in aggregate. For example, an unwanted email or ad is inconvenient on an individual level, but en masse spam emails can be seen as an erosion of people's expectations about how their email addresses are used by businesses.

A useful lens to examine systemic privacy harms is investigating the social impacts of data use. Salome Viljoen writes about focusing on social inequalities, and not just individual autonomy, as being a core part of what should be considered an informational harm. She contends that the patchwork of privacy legislation in the US gives too much weight to the individual, that it focuses on data as an individual medium. She argues that instead the social effects of privacy erosion should be given more relevance instead of just individual interests. Ari Ezra Waldman also notes that the current focus in privacy regulation is on personal control over data, and not population-level impact and systemic inequities.

Central to Viljoen’s conception of privacy harms is the ways in which data may introduce informational harms at the population-level, rather than merely at the individual level. This leads to highlighting novel kinds of information harms, such as race and sex discrimination, and other ways of domination and oppression, that are facilitated by algorithmic systems. Data should not only be seen as a piece of

---

233 Citron & Solove, supra note 16 (articulating main types of privacy harms: physical harms, economic harms, reputational harms, discrimination harms, relationship harms, psychological harms, and autonomy harms).
234 Id. at 3.
235 Id. at 582.
236 Id.
237 Id. at 579.
239 Viljoen, supra note 199, at 573, 579.
240 Id. at 581.
information in isolation, but that data in aggregate has the capacity to transmit social and relational meanings, which means the impact of data use goes beyond the individual data subject about whom the data is collected.\textsuperscript{241}

This observation fits well with our taxonomy of model subjects, which identifies the model subject, or specific groups of model subjects, as the main stakeholders impacted by these population-level insights which emerge from datasets. Whereas a data subject may have to maneuver how their data is collected and used, a model subject would be concerned with how they factor in with the population-level insights derived from data in aggregate. Viljoen uses the example of biometric data to illustrate what we note as essentially the relationship between the data subject and model subject.\textsuperscript{242} Biometric data collected about one person flows across several parties, such as the person it was collected from, to the company making models, to a government agency relying on the application, to a person identified through that application.\textsuperscript{243} The crux is that data collected about one person may be used against another person.\textsuperscript{244}

As we have discussed, during data collection, for each data point collected there are several stakeholders involved, such as the data annotator and the model subject. They are both impacted by how the data is annotated, and how models are developed and deployed. Model subjects are particularly noteworthy because the impact on them is more clearly seen in real-world applications. Furthermore, the decisions made by data annotators in the data collection and annotation processes, impacts how model subjects are affected by critical population-level insights and patterns embedded in the final model.

In addition to identifying who or what groups of people are impacted by data at a systemic level, we can use our taxonomy of the stages of dataset development to discuss aggregate impacts in a more consistent manner. Viljoen’s conception of the uneven impact on some marginalized groups is useful when considering the impact of dataset development as a whole.\textsuperscript{245} She states that in some data collection scenarios, some groups may have a much lower risk of being identified through a system.\textsuperscript{246} The harmful social consequences of data collection about members of some socially significant group may be used at

\begin{itemize}
\item \textsuperscript{241} Id. at 583.
\item \textsuperscript{242} Id. at 604.
\item \textsuperscript{243} Id. at 605.
\item \textsuperscript{244} Id. at 586.
\item \textsuperscript{245} Viljoen, supra note 199, at 653.
\item \textsuperscript{246} Id. at 632.
\end{itemize}
inference time to oppress a different group of subjects who also are associated to belong to that group. In effect, information about one set of subjects is used to oppress another. For instance, a hiring software built on a candidate pool that consists of one specific gender and ethnic group will invariably be used to deny opportunities to populations about whom decisions are being made to exclude them from the job opportunity for not matching the “ideal” profile.

Building on the idea of systemic impact on marginalized groups, it is critical to note that large datasets have been widely criticized for being the source of discriminatory and arbitrary categories about specific groups in society. The presence of ethically questionable concepts or annotations within datasets is pervasive. Examples range from quantifying beauty to predicting sexual orientation and classifying gender. A well-known instance involving LSCVD was when it was revealed that ImageNet included offensive categories for women and underrepresented racial groups. However, there are no clear policy remedies about how to address these examples of systemic harm, nor any clear and organized way of identifying the stakeholders involved.

Centering the dataset development process shows us that data collection, annotation practices as well as the model training and implementation stages of the process can be the source of the exacerbation of systemic harmful impacts on marginalized communities. Many of these annotation practices are carried out by

---

247 Id. at 632.
248 Sánchez-Monedero et al., supra note 37.
249 Yang et al., supra note 62, at 2.
252 See Scheuerman et al., supra note 20.
254 Recently, a group of machine learning practitioners and legal scholars have suggested a system for dataset deprecation which, at the least, outlines how to safely remove datasets, provide a rationale for the deprecation, and allow future researchers to analyze the data post hoc. See Alexandra Sasha Luccioni et al., A Framework for Deprecating Datasets: Standardizing Documentation, Identification, and Communication, FACCT ’22.
companies behind closed doors, often protected by proprietary regimes, giving little opportunity to researchers and impacted individuals to critique the practices. For example, Facebook has historically been criticized for hiring contract workers to label status updates and photos uploaded by users with a set of keywords.256

At the same time, the data curators and data annotators have little to no clearly delineated responsibilities towards data and model subjects for these effects. When controversy arose with regard to representations in ImageNet, the only action taken by ImageNet was to remove the dataset and provide outbound links to images from their website, rather than hosting the dataset themselves. This ultimately lacked much effectiveness since downloaded copies of datasets can exist in an uncountable number of private storage drives. The lack of accountability at different stages means we can end up with model implementations that have widespread systemic consequences.

We must expand our understanding of the privacy harms that result from the private development and use of large datasets. This need to develop a clearer accountability framework for datasets is made more urgent when we see companies releasing large datasets to the public voluntarily, but also holding a lot of control over the terms for the release. For example, Meta released a large language model named OPT-175B (Open Pretrained Transformer) trained on completely public data sources (including The Pile) in an attempt to foster transparency and scientific discourse.257 It is important to note that even though this release is unprecedented for a large tech firm as far as algorithmic transparency is concerned, this does not necessarily support accountability over the impact of large datasets. This model, according to Meta, is not the one being used to moderate social media content, nor

---


does it contain any Meta user data. Incidentally, the use of large language models to moderate content, is a source of a systemic harm and a means for mass disinflation campaigns due to the ease of generation of text, directly impacting the autonomy and self-expression interests of millions of social media users. But because there is currently no way of conceptualizing dataset accountability in a way that would recognize this potential group/systemic harm, such large datasets continue to be released to the public without a clear articulation from the public and legislators of what the company should be disclosing. Scholars have also noted the need for better access to private datasets on terms determined by law or government agencies, but there is currently no safe harbor or sets of obligations for data access.

The lack of clarity over the normative demands we should be making from the companies means that companies can decide the terms of meaning of the release of datasets. There can be more clarity in our expectations from companies by first identifying the stakeholders and then the interests involved in dataset development as a whole. In this section, we have connected the ways in which current conceptions of group harm in privacy and data governance can be applicable to dataset development, but questions remain about specific rights and responsibilities of the individuals and entities creating those datasets.

IV. A MATRIX OF DATASET DEVELOPMENT HARMs

In Part II, we elucidated the stages and subjects of the dataset development, illustrating how different subjects can possibly be
implicated in the process of the AI dataset development process. In Part III, we provided an overview of the informational harms that may emerge in this process, and the underlying copyright and privacy norms shaping dataset development. In this section, we combine insights of both of these sections by describing how the distinct taxonomies of stages and subjects intersect to form a matrix of dataset development impacts, and how individual and collective harms illustrated above slot into each section of this matrix. In Part IV.A, we describe this matrix in detail, focusing on each stage of dataset development and addressing how particular legal and ethical issues may connect to particular subjects. We discuss how policymakers can apply this matrix in the assessment of not only whether they should consider the implementation of a machine learning model, but how particular stakeholders are impacted in the process, and what areas need to be tackled by legislative and regulatory interventions. In Part IV.B we situate this discussion within broader approaches to algorithmic accountability. Lastly, in Part IV.C we briefly illustrate how legislative proposals may incorporate the dataset development framework.

A. Applying the Matrix

In Table 1, we offer a matrix created when considering both the stages and subjects of dataset development, and which stakeholders are implicated at what stage of development. Of the subjects noted above, we omit the curators of datasets, since they are typically not subject to any of the particular informational harms noted in Part III. The rest of the subjects -- data subjects, data annotators, copyright holders, and model subjects -- however, are included. Our intention by creating such a matrix is to create an analytic through which policymakers and regulators can more easily understand at what point and which stakeholders are subject to a set of informational harms. This matrix is, of course, a simplification of the iterative, messy, and often poorly documented process of creating a dataset. However, we hope that it can help highlight the areas of most concern, and for whom, in the dataset development process.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Who is implicated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data subjects</td>
</tr>
<tr>
<td>Problem formulation</td>
<td></td>
</tr>
</tbody>
</table>
Table 1: A breakdown of which group is susceptible to harm at every stage of the AI dataset and model development stage.

1. Problem Formulation

As noted above, problem formulation refers to how a machine learning task is conceptualized and constructed. As Passi and Barocas note, "[problems] must be made into questions that data science can answer." Questions about social life which are vague and not sufficiently operationalized into discrete variables must be made to be so. Moreover, how to formulate a question into one which is tractable with AI can have downstream legal and ethical ramifications. In a recent example illustrated by Obermeyer et al., a risk assessment algorithm used in healthcare was demonstrated to be biased against Black Americans. The primary problem lay in the stage of problem formulation; the target variable for healthcare need was assumed to be a proxy for healthcare expenditure. However, due to systematic disparities in healthcare access, Black Americans spend less on healthcare.

261 Passi & Barocas, supra note 100.
262 See Ziad Obermeyer et al., Dissecting racial bias in an algorithm used to manage the health of populations, 366 Science 447–453 (2019); Martin Jr. et al., supra note 100.
In the dataset development matrix, the problem formulation stage is denoted by a grayed out set of cells. This is because problem formulation may implicate every subject before data collection begins. In the Obermeyer et al. example above, Black model subjects felt the direct impact of the harm of the dataset development process, given that healthcare spending was the variable collected in the data, and served as a poor proxy for their actual healthcare needs. However, one could also envision a formulation of healthcare allocation in which other subjects are affected. A healthcare prediction model could use data from individuals without their consent, or from an app which collects biometric or health-related data without informing individuals of their privacy rights or makes decisions or decides insurance premiums from biometric data.\footnote{Insurance providers, for instance, have been offering discounts if policy holders wear biometric trackers for several years now. See Kai Ryssdal, Trading Insurance Discounts for Health Data, MARKETPLACE (Apr. 8, 2015), https://www.marketplace.org/2015/04/08/trading-insurance-discounts-health-data/ [https://perma.cc/BU8M-M3ML].}

Moreover, data annotators could also be implicated in a health model which used photos of individuals or their medical imaging, such as AI models developed by Google which the company claims can detect diabetic retinopathy, a condition which may cause blindness if not caught early.\footnote{Will Douglas Haven, Google’s Medical AI was Super Accurate in a Lab. Real Life was a Different Story, MIT TECHNOLOGY REVIEW (Apr. 27, 2020), https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/ [https://perma.cc/Z8V9-FPUL].}

Lastly, copyright holders could be implicated in the healthcare setting, such as if individuals track their fitness or recovery progress in online image collections or social media platforms with expansive terms of service agreements that assign copyright privileges to the platforms, and curators then use these data to make healthcare allocation decisions. For the curators of large-scale AI datasets, problem formulation is typically not defined solely by the curator, but may be agreed upon by a larger research community. Some tasks, like facial recognition and analysis, have been problems within computer science for decades.\footnote{Shaun Ravivi, The Secret History of Facial Recognition, WIRED (Jan. 21, 2020), https://www.wired.com/story/secret-history-facial-recognition/ [https://perma.cc/ZUSB-5K7U]; Morgan Klaus Scheuerman et al., Auto-Essentialization: Gender in Automated Facial Analysis as Extended Colonial Project, 8 BIG DATA & SOCIETY (ISSUE 2) (2021).} Others, such as multi-modal image generation, as demonstrated by tools like OpenAI’s DALL-E and its successor DALL-E 2, are emergent tasks which have only been possible with the development of LSCVDs,
LSLDs, and large-scale datasets which contain captioned images. However, the curator still has some individual agency in how they choose to collect data and what labels they assign to each of them. While a facial recognition system has a clearly defined task (match an image with a face or some internal representation within a larger database), for a dataset used for object classification, like ImageNet, a curator can arbitrarily define the set of category labels for the images within the dataset. This process is largely undocumented and not subject to scrutiny or review under any current regulatory regimes.

Moreover, choices over which data to use and how annotator labor is incorporated can skew the problem definition further. These task formulation choices have implications on a more micro scale as well, such as when technology vendors create customized automated hiring systems for companies to use. In the diabetic retinopathy example noted earlier in this section, although the AI model was trained on high-quality images of retinal scans and performed well under laboratory testing conditions, it performed poorly in the field, as patients’ images were taken under low-quality or low-light conditions.

In short, all subjects identified above in the data development pipeline are implicated in the problem formulation stage. It sets the legal and ethical stakes for what is to follow in the form of potential informational harms. The goals of the data collection process are defined at the problem formulation stage, and are often decided under conditions in which the subjects are not consulted or only considered after the fact.

2. Data Collection

Data collection, as a process, has been discussed broadly in legal and social science scholarship related to data mining. Scholars have

---


268 Haven, supra note 264.

269 See generally Barocas & Selbst, supra note 16; Zeynep Tufekci, Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls,
examined the relationship between sources and practices of data collection and the creation of datasets, and the subsequent biases that may emerge in the data mining process. Data collection, then, has implications for data subjects, copyright holders, and model subjects.

Data subjects, in the most obvious case, risk data privacy violations both on the individual and collective level. An individual’s privacy may be violated by having their identity or sensitive information exposed, or an aggregation of their images or text messages collected in one location. These are, as Calo suggests, objective privacy violations. There may also be a subjective privacy harms, due to the feeling of being surveilled online. For instance, New York Times reporting describes how many Flickr users were shocked that their images had been used for training facial recognition, a subjective harm. It makes less sense to discuss population-level privacy harms at this stage, because judgements about members of a particular population-level group are not subject to decision-making at the data collection stage.

Moreover, data collection practices can exacerbate existing social stratification and subsequent discrimination for model subjects. What happens when an underrepresented group’s data is not collected and included at a systematic level? The “quality of individual records of members of these groups be poorer as a consequence, but these groups as a whole will also be less well-represented in datasets, skewing conclusions that may be drawn from

---

Barocas & Selbst, supra note 16; Tufekci, supra note 269; Lerman, supra note 169, at 57; Boyd & Crawford, supra note 269, at 666-68; see also Gitelman, supra note 170.

270 See generally Barocas & Selbst, supra note 16; Tufekci, supra note 269; Lerman, supra note 169, at 57; Boyd & Crawford, supra note 269, at 666-68; see also Gitelman, supra note 170.


272 Lerman refers to this as “data antisubordination” building on equality doctrine of antisubordination which holds that we cannot have equality in the face of pervasive social stratification. Lerman, supra note 269, at 60.
an analysis of the data." 273 Those from marginalized groups may be less experienced in video interviewing and lack of access to technologies (like laptops, webcams, and proper light). This further has the effect that these candidates may not be routinely included or considered in the pool of data that is used to train the automated hiring system. At the point of data collection, then, many subjects may be left out of the equation as a whole.

Copyright holders may be the least impacted to informational harms at the point of data collection compared to the other subjects, but it is at the point of collection in which the original copying is taking place and most likely that copyright violations may occur. It is difficult to isolate if a violation has occurred at this stage without understanding the downstream usage of the data instance. However, once the data instance has been collected into a larger dataset, the copyright holder then has an interest in those downstream uses, and is therefore further implicated.

In the first stylized example above, the copyright holder (Croydon) takes a picture of their daughter (Sophie) and uploads it to a photo sharing site. Dr. Kurry, the curator, grabs the image to store in a dataset for evaluating facial recognition models. Prior to the photo being aggregated into that dataset, under the US privacy regime, Sophie may experience a subjective privacy harm due to having her picture aggregated into a dataset without her knowledge, without any actionable way to obtain consent from her or, if she is younger than 13 years of age, her parent. Croydon, as the copyright holder, would have a different set of concerns. First, the image uploaded to a photo sharing site may be subject to the terms of that website, which may include a license granted by Croydon to allow the website to use the image (which is what usually occurs when users agree to Terms of Service agreements). As discussed above in the context of whether scraping this image is fair use, an individual’s licensing market may not be significant enough on its own, unless Croydon is a professional photographer and needs to be compensated for the use of their images. Therefore, as far as current data collection practices are concerned, an individual, whether the copyright holder or the data subject, has little recourse in case they wish to object to their images being scraped from the internet.

In the second hypothetical example, the data subject (Sunipa) has her private information doxxed and posted to a pro-life section of

Reddit. Her private information is collected by a curator (the GatherEverything Collective). At the point of posting, Sunipa is subject to an objective privacy harm, which can put her at risk of unwanted interactions and harassment, especially with pro-life activists, identity thieves, or other bad actors who act on the privacy violation. The original posting of her information to Reddit and the resultant harms she faces are not a function of the dataset collection process. However, because the GatherEverything Collective has collected those data, they are giving new life to that potential privacy harm. Sunipa has little recourse, as that information remains in a dataset, perhaps unbeknownst to her.

We can see the connections here between the various subjects at this stage, and this is useful to begin mapping how data collection starts to impact individuals from the very beginning.

3. Data Cleaning

Data cleaning is the process of taking information from a less structured source (such as the web) and turning into it something that’s useful for analysis. While Lehr and Ohm have referred to most of the steps in constructing the machine model prior to deployment as the process of “playing with the data”, the process of data cleaning is possibly the one in which machine learning practitioners and dataset curators are doing most of the experimenting and playing with a new set of data instances. Data instances may be incomplete or not meet some predefined criteria, say, for quality of images, or legibility or clarity of text excerpts. Data cleaning is the process of removing those elements, or perhaps imputing missing data so the curator does not have to throw away a data instance. This process implicates two groups of subjects—data subjects and model subjects—but in distinct ways.

This stage is likely to impact model subjects and reflected in what types of images AI models are evaluated and trained upon. For instance, absence of particular quality or legible data instances may not be randomly distributed across salient social categories. Elements of someone’s attire, such people who wear hijab or other religious headwear, may be considered obscuring of an individual’s head and not suitable for downstream applications. In the example noted above, individuals who have lower quality cameras may be lower income or from communities of color, which exclude them from data samples.

274 See Lehr & Ohm, supra note 97.
which set a bar for image quality.\textsuperscript{275} This, then, has implications for what types of predictions the model makes for individuals not represented in that dataset. This model of exclusion has been better explored in prior literature, especially work which examines violations of disparate impact as an emergent property of biased data.\textsuperscript{276} This step is closely related to the data collection process described above in that inclusion and exclusion choices have consequences for how the dataset is deployed.

Data subjects may also be impacted in the data cleaning process, notably in terms of particular types of privacy violations which curators cannot account for. Although privacy is rarely considered as a top-line priority for dataset curators,\textsuperscript{277} more recent dataset creation efforts, especially in the space of LSLD and large-language model creation, have been attentive to removing information which may be considered private, such as social insurance numbers, addresses, and telephone numbers. However, this process is not perfect, nor is there any legal, professionalizing, or normative requirement to compel curators to perform this type of obfuscation or privacy-preserving work in the data cleaning step if the data instance does not contain any biometric data (e.g., an image of the data subject).\textsuperscript{278}

This need is most evident in the second stylized example above, in which Sunipa’s private information was non-consensually released by anti-choice activists and was scraped into a repository by the GatherEverything Collective. The Collective attempted to “clean” out her private information, but their attempts were not sufficient. The individual privacy harm to Sunipa should be clear. We highlight this as this aspect of data cleaning as a step in safeguarding privacy is underappreciated both in the legal and computer science literature.

4. Data Annotation

\textsuperscript{275} EEOC, \textit{supra} note 273.

\textsuperscript{276} See e.g., Barocas & Selbst, \textit{supra} note 16; Rashida Richardson et al., \textit{Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice}, 94 N.Y. UNIV. L. REV. 16-50 (2019).

\textsuperscript{277} See Scheuerman et al., \textit{supra} note 20, at 24-25.

\textsuperscript{278} Jason Schultz has, however, recently argued that the right of publicity may provide recourse for the distribution of the likenesses of data subjects. Critically, the right of publicity does not require a subject’s actual biometrics, but any aspect of their identity, including but not limited to personal stories or uniquely identifying characteristics and roles. See Jason Schultz, \textit{Can the Right of Personality Rein in Facial Recognition?}, \textit{Privacy Law Scholars Conference} (2022).
Data annotation has received possibly the least attention of any of these stages in our matrix in the copyright and privacy harms literature. Data annotation is the process of a human attaching a label or annotation to a piece of data, such as an image or a text. Data subjects are most likely to be impacted at this stage of the dataset development process. Wang et al. discuss the fairness of image captioning practices and outline a few harms that may result in the process. These authors break down bias into specific categories of representational harm, which they define as “harms that occur when some social groups are cast in a less favorable light than others, affecting the understandings, beliefs, and attitudes that people hold about these social groups – caused by image captioning systems.”

Most relevant for data subjects are the issues that occur at the stages of “human-generated” labels and captions in the dataset formation process, when data annotators try to capture everything in an image and create descriptions for it. At this stage, harms such as stereotyping or demeaning may occur. Stereotyping can impact both individuals and groups of data subjects in various ways. For example, if the word “gun” is associated with the images of a Black person, then a racial stereotype is further entrenched in the dataset.

It is important to note that these representational harms are not fully captured in existing notions of privacy harms, especially when privacy harms are juxtaposed onto the dataset development process. It is by dissecting the process of data annotation that we can recognize the ways data subjects are impacted.

However, annotation is not a straightforward task; there are many interpretations which could be attached to an individual data instance. Although how individuals in images are tagged has shown to be problematic, as shown in ImageNet and other LSCVDs, the work of labeling instances of texts is inherently difficult and contextual. For instance, in the areas of content moderation and monitoring for hate speech, this work depends heavily on the local understanding of annotators who may be supplying the training data for the algorithms that power AI screening tools. This work is continually outsourced to parts of the Philippines and Kenya, and policies around what content is moderated may change swiftly depending on the directives of policy.


280 Id. (citing Solon Barocas et al., The Problem With Bias: Allocative Versus Representational Harms in Machine Learning, SIGCIS CONF. (2017)).

281 Wang et al., supra note 281, at 327.

282 Id. at 329.

283 Crawford & Paglen, supra note 31; Prabhu & Birhane, supra note 31.
managers. In this contextual process, it is difficult to pinpoint who may be responsible for the downstream effects of data labeling. It is also difficult to configure data subjects’ rights when the process of dataset development is spread out in many different jurisdictions.

Moreover, field studies of what data annotators do in their day jobs has revealed the contingent and difficult nature of this work. When data annotation work has been outsourced to crowd computing platforms, such as Amazon Mechanical Turk, workers are pressured to meet a high bar of quality in their annotations. There is not much, if any room, to deviate from the instruction sets which are laid out by requestors, and workers have little room to push back against employers in these situations. Therefore, annotators must hew closely to instructions which may be incomplete or unclear. Moreover, because much annotation work is fully distributed and online, large companies which manage this work can encourage a “race to the bottom” in global competition for data annotators who speak proficient English. Many of these annotators are people in precarious employment positions in the Global North and South, including refugees and skilled migrant workers who have little other option for employment.

Although data annotators feature in our matrix, they are not subject to the copyright or privacy-based harms in which other individuals are implicated. Much of the harms that annotators face are allocative in nature and based in the vagaries of a growing sector of the global labor market and desire for “high-quality” data to feed to AI algorithms. If anything, data annotators may feature a secondary harm when it comes to data and model subjects seeking accountability from curators who rely on data annotators to produce training data. The blame for failures of the AI may be passed on to the data annotators, and contracts with the outsourcing companies which employ the annotators may not be renewed, or, in distributed services like Mechanical Turk, those annotators are not relied upon anymore for a particular curators


285 Id. [https://perma.cc/4B4H-C3FQ].

tasks. In this case, annotators may become what Madeleine Clare Elish has called “Moral Crumple Zones” of AI, in which culpability for errors gets redirected. This reality underscores the need to create greater accountability for the platforms and companies that rely on data annotators instead of focusing on the annotators themselves.

In our first example, Malik’s wrongful apprehension may be attributed to a lack of data quality of the dataset obtained from Dr. Kurry’s lab for fairness testing, and the data annotators considered to be the ones to blame. San Junípero Police Department may point their figures at SecurFace, who may double down on their methods publicly. Privately they may be reevaluating their strategies for fairness evaluation. In any case, finger pointing will likely filter down the machine learning supply chain to deflect criticism of SJPD.

5. Model Training and Evaluation

After data cleaning and annotation, we move into model training and evaluation. These steps may be performed by someone other than the curator, although the curator most likely is the first to train and evaluate models on this dataset. As a matter of practice, model training and evaluation go together, as they are two sides of the same coin. A model is trained and evaluated at multiple steps as it is still training, since it can take days if not weeks and months to fully train a large-scale AI model. Or many different models are trained and evaluated, and the one with the best performance metrics is the model which may be chosen to be put into production.

Metrics of evaluation and success of a model are typically associated with datasets themselves. These metrics measure some type of accuracy, whether that is precision, recall, F-score, or area under the RoC curve. For instance, the image classification task associated with ImageNet relies on the top 5 guesses of what’s in an image. Large language models are evaluated on a suite of tasks which are associated

288 Training Checkpoints, TENSORFLOW, https://www.tensorflow.org/guide/checkpoint [https://perma.cc/UB6T-4QLF]. Most modern deep learning frameworks allow model developers to see evaluation metrics as model training is still progressing. E.g., Google’s TensorFlow framework periodically saves checkpoints which document all the current parameters of a model and allow a point estimate of how accurate the model is performing.
289 “For the image classification task we allowed an algorithm to identify multiple (up to 5) objects in an image and not be penalized as long as one of the objects indeed corresponded to the ground truth label.” Russakovsky et al., supra note 55, at 15.
with natural language understanding, such as how well the model can “answer” a question or predicting the sentiment of a movie review.\textsuperscript{290}

Because model training and evaluation attempts to minimize error according to some mathematical function, these algorithms, at least out of the box, are not attentive to potential harms. We have seen that informational harms can emerge from prioritizing accuracy over how well the model performs with particular subclasses of data instances. The desire for accuracy may overshadow other concerns such as representation or performance for other salient social classes of individuals. Accuracy here means that a model may work in accordance with its designed purpose, or solve the problem formulated at the first step of its development. But accuracy can come at the expense of including representative data, or not correcting the gaps that may have been left at the data cleaning and annotation stages. In this case, the model subject is the subject most implicated here: there is typically no way that fairness can be guaranteed against socially significant categories. Harms here, again, are not related to that of copyright and privacy, but potentially of disparate impact, and overlooking procedural safeguards such as explainability and transparency.\textsuperscript{291}

This has been the most significant focus of the technical literature on algorithmic fairness. Definitions of fairness have promulgated by computer scientists and statisticians, which roughly hew to a particular interpretation of US anti-discrimination law.\textsuperscript{292} These technical definitions typically treat fairness as a problem of training and evaluation; instead of relying on a single mathematical function and aiming to minimize error, algorithms must also meet other evaluation criteria to ensure that the outcome is “fair,” depending on the definition.\textsuperscript{293}

\textsuperscript{290} TENSORFLOW, supra note 288.


\textsuperscript{292} For a review of this work, see Shira Mitchell et al., Algorithmic Fairness: Choices, Assumptions, and Definitions, 8 ANN. REV. OF STAT. AND ITS APPLICATION 141 (2021); For a critical review of these definitions and the internal criticisms of US anti-discrimination law writ large, see Anna Lauren Hoffmann, Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse, 22 INFO. COMM. & SOC’Y 7 (2019).

\textsuperscript{293} In a widely circulated talk at the Fairness, Accountability, and Transparency conference, computer scientist Arvind Narayanan has defined up to 21 definitions of fairness in this space. Arvind Narayanan, Tutorial: 21 Fairness Definitions and Their Politics, YOUTUBE (Mar. 1, 2018), https://www.youtube.com/watch?v=jLXJuYdnyyK [https://perma.cc/99M9-5MQ7].
This is also a case in which systemic privacy harms emerge, as outlined by Viljoen, in which data about one set of subjects is used to impinge on the population-level rights of another set of subjects. While data subjects are not directly implicated at this stage of the dataset development pipeline, they have a relational connection with the model subjects who are implicated. This is significant to note because at this stage of dataset development, data subjects may not have much recourse to extricate themselves from such a relational connection.

The first stylized example illustrates this best. Although SecurFace makes assurances to its customers for both accuracy and fairness using Dr. Kurry’s dataset, their guarantees of fairness are difficult to adjudicate. Third-party audits of facial recognition algorithms are few and far between, and there are few regulatory bodies with the authorization of conducting them. Malik, as a wrongfully arrested model subject, is the one most affected and whose liberty and right to privacy is reduced. This has ramifications more broadly on a population-level, as Black individuals are more susceptible to this harm.\(^\text{294}\) Recognizing the consequences on model subjects at this point in the process should allow us to recalibrate what it means for the AI model to be both accurate and fair.

6. Model Implementation and Inference

Once a model is in a production environment, it has reached the implementation and inference stage. These are typically models used in corporate and state applications, and less commonly are deployed by academic researchers themselves. The model implementation and inference stage of dataset development is particularly impactful for both data and model subjects. For example, models may be used to surveil communities of color, or decide grades for final exams for students in secondary schools.\(^\text{295}\) In these scenarios, data subjects whose privacy interests are involved may be impacted because private details may be

\(^{294}\) It is worth mentioning here that third-party auditing is no panacea for the fairness of particular types of carceral technologies, such as facial recognition. Given that these tools are deployed disproportionately in communities of color and fit within a system of criminal justice which is already systematically racist in its orientation, “fairness” is a limiting frame in this example. See also Sarah T. Hamid, Community Defense: Sarah T. Hamid on Abolishing Carceral Technologies, LOGiC (Aug. 31, 2020), https://logicmag.io/care/community-defense-sarah-t-hamid-on-abolishing-carceral-technologies/ [https://perma.cc/2ZMT-LACK].

exposed via model inference. This was a concern when GitHub released Copilot, when researchers found that Copilot could generate code that could be susceptible to cyber security attacks. Model subjects are distinctly affected by model implementation, but are at the receiving end of the inferences and decisions made based on a model and therefore subject to similar types of information harms noted in the model training and evaluation section above.

The model implementation stage may also involve copyright holders where there is a commercial interest involved. This could possibly happen in two ways. First, if the model implementer is using the dataset for commercial gain, then this may weigh against the fourth factor, effect on the market, of the fair use doctrine. This is because a copyright holder could potentially have an interest in their work not being used for commercial purposes. Moreover, it may also be in violation of open licenses, such as Creative Commons, and the license terms that may prohibit commercial use of datasets. Secondly, the copyright holder may be implicated if the model reproduces, in its entirety, a piece of copyrighted work. Github’s Copilot model reproducing the fast inverse square root algorithm used in the computer game Quake III is a prime example of this in practice. This is also a concern when open models like DALL-E are used to generate artworks that look similar to existing ones. The concern becomes both ethical and legal when the art belongs to small-scale artists who place their works online in hopes of generating interest and visibility, and do not have the means to seek compensation for the use of their works.

In the second stylized example above, Sunipa – as both data subject and model subject – faces a further privacy harm beyond her initial doxxing, as her private data has a secondary afterlife in the OpenLLM generative text model. Typing her name as input to the OpenLLM model republishes her private phone number and address. Sunipa has little recourse here – she can ask OpenLLM to exclude those types of results, but they may not have a straightforward path under existing privacy law regimes in the US.

296 Pearce et al., supra note 13.
297 Barber, supra note 11.
7. Data and Representation Distribution

Lastly, recall that data and representation distribution is the act of reposting the data or some downstream representation (such as word embeddings or pre-trained model architectures) to a lab website or as part of a machine learning software package. The act of hosting these data and representations presents special issues, because, even post-retraction, these datasets can have afterlives and remain in circulation.299 Because of this, distribution has an impact on data subjects, copyright holders, and model subjects.

Data subjects are implicated in data distribution by way of individual privacy harms that may result from these models being used largely without any oversight or knowledge of the data subjects about being part of this dataset, now available for (semi-)public consumption. One of the conventional approaches to privacy regulation has been the notice and consent model, whereby users agree to certain terms and uses before accessing a web service or platform. In practice, notice-and-consent models have shown to be ineffective in changing exploitative data processing practices, or even in adequately informing users of how their data is being processed.300 Notice-and-consent models have been critiqued heavily and are not seen as the optimal solution to privacy concerns in recent years.301 Current practices related to how datasets are distributed make it even more difficult to obtain informed consent from data subjects. Even if data subjects could consent to their data being processed for specific uses, there is no way to trace these uses once these datasets are used to build a variety of applications with multiple downstream uses.

Similarly, copyright holders may be unaware of their data being used to train AI models. This necessitates substantive and informed

299 See Peng et al., supra note 17; See Luccioni et al., supra note 254. Luccioni and co-authors have recently suggested a framework for retraction which centralizes data hosting and describes precisely why a dataset is retracted.
301 See Paul Schwartz, Internet Privacy and the State, 22 CONN. L. REV. 815, 824-25 (2000) (critiques notice and consent as ineffective for a number of reasons including information asymmetries, collective action problems); Helen Nissenbaum, A Contextual Approach to Privacy Online, 140 DAEDALUS 32, 36 (2011) (stating that there is a "transparency paradox. Achieving transparency means conveying information handling practices [however] [but] if notice . . . finely details every [relevant fact] . . . we know that it is unlikely to be understood, let alone read."); Kenneth Bamberger & Deirdre Mulligan, Privacy on the Ground: Driving Corporate Behavior in the United States and Europe (2015).
consent on part of the individuals who may be impacted in the ways we have elaborated above. However, not only would it be complicated to explain how data is being used in a given dataset, dataset exchange and distribution practices between commercial and non-commercial entities is difficult to track. Terms of Service agreements in their present form have limited scope when it comes to informing data subjects of how their data is being used.

Model subjects also face a substantial amount of opacity when it comes to understanding how datasets may impact them and how to consent to an AI model being used to make decisions about them. Due to the long and complex process of dataset development, a model subject can only understand a small part of the process by which the AI model has been developed. This means that substantive rights such as a right to explanation or transparency will need to take into account the dataset development process as a whole, and data distribution practices specifically, to create an effective accountability framework.

B. Normative and Regulatory Approaches to Dataset Accountability

1. Systemic and Individual Algorithmic Accountability

Thus far we have shown how various subjects are connected to one another and implicated at the different stages of dataset development. In this section, we situate this analysis of the dataset development within a discussion of current approaches to algorithmic accountability. We seek to explore how this framework – the taxonomies and the resulting matrix – fits into solutions that have already been proposed to regulate AI and automated decision-making systems. Legal scholars have suggested both individual- and system-level solutions in the United States, with the latter being the focus of recent regulatory proposals such as the Algorithmic Accountability Act.\footnote{302 The recently proposed Algorithmic Accountability Act (AAA) would mandate algorithmic audits and impact assessments. H.R. 6580, 117th Cong. (2022).} We examine both kinds of solutions and suggest where the dataset development framework may be useful.

Legislative and regulatory measures rely on particular normative values, prioritizing constitutional protections such as anti-discrimination and due process. These are foundational to building the normative basis of algorithmic accountability. Below, we examine how scholars have suggested incorporating these norms.
Danielle Citron calls for both individual and systemic regulation of algorithmic systems, especially of those used in government agencies. She notes that automated decision-making systems implicate the due process clauses of the Fifth and Fourteenth Amendments. Concerns such as deprivation of property interest without due process – for example, denial of food stamps based on inaccurate assessment of a person’s eligibility – may be at stake. The right to be given notice requires individuals to be informed of how they will be impacted, the evidence used to make the decisions, and the government agency’s decision-making process. Moreover, the opportunity to be heard is also implicated in automated decision-making systems being used in government settings.

In order for an individual to realize these due process rights and contest a decision, however, Citron notes that an individual would need meaningful access to the automated program’s source code. Systemic solutions, such as creating mandates around disclosure of the computer decision-making process and audit trails is one way of addressing these information gaps. Citron also notes that accountability and transparency need to be built into a system’s design and private vendors should make source code available to the public. Here we see that both individual and systemic solutions go hand in hand. In light of the dataset accountability framework, due process violations fall on the model subject, who then have an interest to interrogate the processes of data collection, cleaning, model training and inference, and data distribution.

In the context of the government’s use of AI systems, Aziz Z. Huq focuses on systemic solutions, discussing the need to recalibrate...
norms of due process and equality to machine learning.\textsuperscript{311} Huq notes that in order for a lawsuit to be successful, we need more clarity about how large aggregates of data are used, which indicates that we need to look closely at dataset development. He notes that, at the moment, there is no exploration of the risks arising from the creation of large data aggregates.\textsuperscript{312} He argues that there needs to be a due process analysis on system-level choices, because the problems are systemic in the first place.\textsuperscript{313} For example, we would need to examine how well the training data, model, and outcomes match with one another.\textsuperscript{314} Furthermore, due process concerns which arise in the calibration of due process means that the constitutional analysis must look at algorithmic design choices in time from the incident in which a human is subject to the classification – which means looking at data collection and annotation stages that we have elaborated on above.\textsuperscript{315} These observations suggest that an examination of how datasets are formed ought to be a central part of systemic solutions.

System-level choices in dataset development also impact the ways in which errors and choices in the machine learning process trigger equal protection concerns.\textsuperscript{316} For example, annotation or cleaning errors can lead to biased outcomes if the “designated outcome variable fails to track ground truth equally well for different groups, or based on race or

\textsuperscript{311} Aziz Z. Huq, Constitutional Rights in the Machine-Learning State, 105 CORNELL L. REV. 1875 (2019); \textit{Id.} at 1880-1881, 1883.

\textsuperscript{312} \textit{Id.} at 1905. See also State v. Loomis, 881 N.W.2d 749, 761-69 (Wis. 2016). In this case, the court rejected an individual due process challenge to the use of COMPAS, an algorithm used to determine sentencing evaluations. Although the challenge was rejected, the Court did note the proprietary nature of COMPAS and the inability to understand more about how group data was being used to calculate risks about individuals. See also Andrea Roth, Machine Testimony, 126 YALE L.J. 1972, 2026 (2017). Andrea Roth introduced a taxonomy that can be used to categorize the various kinds of machine-based evidence that may be introduced in court. This taxonomy includes standards for what constitutes sufficient information about a dataset, and documentation related to it.

\textsuperscript{313} \textit{Id.} at 1910-1911.

\textsuperscript{314} \textit{Id.} at 1917 (“determining whether a machine learning tool impinges on due process demands an examination of the fit between quality training data, the learning model, and outcome variable, and the match between the outcome variable and the latent variable.”)

\textsuperscript{315} \textit{Id.} at 1906.

\textsuperscript{316} Anytime a government distributes benefits or burdens based on individual racial classification, it will lead to the application of strict scrutiny. Parents Involved in Cmty. Sch. v. Seattle Sch. Dist. No 1, 551 US 701, 720 (2007); Jack M. Balkan & Reva B. Siegel, The American Civil Rights Tradition: Anticlassification or Antisubordination? 58 U. MIAMI L. REV. 9, 10 (2003) (the anticlassification principle is that a government may not “classify people either overtly or surreptitiously” based on factors such as race).
At a systemic level, there are two elements needed in solutions that address constitutional concerns: ex ante rules to create transparency and impose disclosure requirements, and making aggregate-level litigation remedies more available. Scholars have suggested that we can achieve better transparency by making design choices more available. However, saying we need more information about design choices is vague without a clear set of questions about the information we are seeking. Here, a deeper understanding of the stages of dataset creation, including the framework we have proposed and associated documentation interventions, such as datasheets, would allow us to identify what specific design choices we need more clarity on.

The dataset development framework is also useful when it comes to effectuating individual-based proposals for algorithmic accountability. The individual rights such as transparency and notice need to be incorporated. Kate Crawford and Jason Schultz have written about how individual due process protections could be developed in the context of automated systems. They note that in the context of large amounts of data being processed and the wide range of resulting applications and effects, we will need to maintain more flexible standards that emphasize values instead of fixed procedures. On the other hand,
elements of due process such as "the right to have the decision based only on the evidence presented; the right to counsel; the making of a record; a statement of reasons,"\textsuperscript{325} are more easily applicable if seen through a dataset development framework. Our framework can help individuals identify their relevant interests by clarifying their subject positions, and more clearly defining what is at stake for them. As we have elaborated, there can be critical differences between a data subject and model subject's interest, and the framework we have proposed can more clearly define redress for each one.

Margot Kaminski and Jennifer Urban write about enacting a more precise and actionable individual right to contest AI, and the necessity of such a right in accordance with constitutional ideals of due process.\textsuperscript{326} Their basic premise is that a right to contest AI would give individuals the ability to exercise some control over unjust outcomes and increase the predictability of decisions and also preserve values like rule of law, fairness, and justice.\textsuperscript{327} Furthermore, procedural justice is an essential quality of a legal system which shows respect towards its subjects, builds citizens' trust in the legal system, and protects values such as transparency and consistency.\textsuperscript{328}

Furthermore, they emphasize that an effective individual right requires attention to the "entire decision-making system – human, machine, and organizational."\textsuperscript{329} This means that designing an effective legal framework has to interact and be attuned with the technical and organizational structures creating the AI system. Here, attention to the dataset development process can not only be a source of key information about these structures, but also a way to identify specific informational harms that may arise during the development process. This framework would thus allow individuals to have more clarity about what actions and practices to contest.

Other scholars have attempted to recalibrate individual remedies such as the right to explanation. Andrew Selbst and Solon Barocas detail the ways in which machine learning evades explanation and how to work around the barriers to better explanations.\textsuperscript{330} They note

\textsuperscript{325}Id. at 1284-92.

\textsuperscript{326}Kaminski & Urban, supra note 30.

\textsuperscript{327}Id. at 1974-1975.

\textsuperscript{328}Tom R. Tyler, What is Procedural Justice?: Criteria Used by Citizens to Assess the Fairness of Procedures, 22 LAW & SOC’Y REV. 103, 132 (1988).

\textsuperscript{329}Kaminski & Urban, supra note 30, at 2031.

\textsuperscript{330}Selbst & Barocas, supra note 232, at 1087; see Lilian Edwards & Michael Veale, Slave to the Algorithm? Why a ‘Right to an Explanation’ Is Probably Not the Remedy You Are Looking for, 16 DUKE L. & TECH. REV. 633, 657-60 (2017) (There are significant
that the inherent inscrutability and non-intuitiveness of machine learning algorithms are challenges to designing an effective implementation of a right to explanation. Moreover, source code that is released is often not helpful because it requires specialized knowledge to understand, and the more sophisticated an automated decision-making system becomes, the harder it is to explain. In order for this right to be effective, it also has to be incorporated into system-level solutions. They also note that explaining the outcomes of a single case do not provide enough information about the logic or normative values of a system, and that is difficult to provide explanations of causal results for each individual case.

Our framework may provide an analytic to be applied to a particular model and allow at least a partial solution to the problem of model inscrutability. Beginning with the processes of data collection, cleaning, and annotation does not provide an explanation, per se, but provides an entry point to understand what may shape model decisions. When explaining decisions, or laying out the justifications for a decision, the kind of information needed include “values and constraints that shape the conceptualization of the problem”, – an issue we note in the problem formulation stage discussed above – and how these values feed into the development of the models. Effective explanations make it imperative to learn about the decisions that lie behind and become part of a model, which we contend includes how the dataset is created.

Selbst and Barocas further note that accounting for decisions made in the process of model development reveals subjective judgments that should be evaluated, and these kinds of explanations are important because they are not immediately obvious from a model. The dataset

barriers to an individually based right to explanation, such as lack of individual capacity and access to justice).

331 Selbst & Barocas, supra note 232, at 1087
332 Id. at 1093
333 Id. at 1094; see also Jenna Burrell, How the Machine thinks: Understanding Opacity in Machine Learning Algorithms, BIG DATA & SOCIETY 1, 3-5 (2016).
334 Selbst & Barocas, supra note 232 at 1105.
335 Id. at 1104-1105; See also Riccardo Guidotti et. al., A Survey of Methods for Explaining Black Box Models, 51 ACM COMPUTING SURVEYS (2018); Michael Gleicher, A Framework for Considering Comprehensibility in Modeling, 4 BIG DATA 75, 82-83 (2016).
336 Selbst & Barocas, supra note 232 at 1130.
338 Selbst & Barocas, supra note 232.
339 Id. at 1133.
340 Id. at 1137.
development framework reveals who is impacted by these subjective judgments and contribute to a roadmap for answering questions about what to include in an explanation in the first place.

Both individual and systemic solutions also need critical information from private companies and vendors developing the automated decision-making systems. An important requirement for these proposals to work is good faith participation by private companies, since key information is often protected under trade secrecy. Andrew Selbst explores the scope of system-level proposals such as Algorithmic Impact Assessments (AIAs) and Audits. Companies are also best placed to provide more information on the impacts of their technology due to the complex nature of how they are developed. As Selbst notes, it becomes difficult to develop liability regimes in the absence of knowledge about the development process. Our dataset development framework can serve as a starting point for what information needs to be disclosed in the first place, and who may be best placed to provide that information.

In addition to gradual participation and disclosure standards for AIAs, other systemic proposals directed at private companies recommend testing, design and documentation requirements, whistleblower protections, and public interest course of action. Sonia Katyal calls for both ex ante and ex post evaluation of training data, as well as whistleblower protection for employees working at the companies developing these algorithms. Whistleblowers have been instrumental in shedding light on critical practices inside corporations and have received protection under US law to some extent. Katyal makes a case for whistleblower protection for employees working on AI, as these individuals could provide important information on

341 Selbst, supra note 28 at 4-5.
342 Id.
343 Id. at 11.
344 Id. at 12.
345 Sonia Katyal, Private Accountability in the Age of AI, 66 UCLA L. REV.; Desai & Kroll, Trust but Verify: A Guide to Algorithms and the Law, 31 HARV. J. L. & TECH., 1, 43 (2017); Huq, supra note 311 at 1940 (calling for class actions and aggregate challenges to identify system wide causes of harm).
346 Katyal, supra note 345, at 126-129.
discrimination, bias and other AI-related harms. We have shown that many of the practices in data collection, cleaning, and annotation may be arbitrary and difficult to track. In such instances, insider access and whistleblower protections may be necessary to encourage employees to come forward with information about harmful practices.

It is clear from these various proposals that both individual and systemic solutions are necessary and intertwined. We see the dataset development framework as serving two main functions: as a standard to deepen dataset documentation requirements, and as an analytical framework to identify interconnected informational harms and ethical concerns. This framework makes the decision-making process around dataset creation more comprehensible and identifies the individuals who are impacted. At the system-level, the public would be better informed about processes that are largely opaque and unchecked. It would aid the right to explanation and other transparency requirements, as well as help specify how individual rights such as contestation and due process may be upheld.

It is important to note that while the dataset development framework can be helpful as a documentation standard, it also sheds light on the harms that arise from the development process itself, such as the subjective and objective harms we have noted. Therefore, paying attention to the dataset development process also leads to avenues for individuals to contest the use of AI systems, and highlight the novel harms they may be subject to.

2. Proposed Legislation

In the past year, many new laws have been introduced, and some have passed, at the state level. Some of the laws that have been passed address a number of applications, such as hiring and the use of facial recognition in criminal justice. A number of pending laws would

---

348 Katyal, supra note 345 at 126-129.
349 Selbst, supra note 30, at 37.
350 *Infra* Part II.
352 H.B. 2022-420, Reg. Sess. (Ala. 2022) (limits the use of facial recognition to make arrests); COLO. REV. STAT. § 2-3-1707 (2022) (restricts use of facial recognition); Artificial Intelligence Video Interview Act, 820 ILL. COMP. STAT. 42/20 (mandates reporting requirements for employers using AI for hiring); see Caroline Kraczon, *The
mandate documentation practices like impact assessments. Proposed laws will also establish specific government agencies to oversee the use and deployment of AI.

These are welcome steps towards creating accountability structures for the use of AI by state agencies. There is still a gap when it comes to regulating how AI is developed in the first place, which is crucial if we are to address the informational harms highlighted in this paper. In section, we look at a law introduced at the federal level that attempts to establish overarching accountability for how automated systems are designed and used. Incorporating the dataset development matrix could be beneficial to how these laws are interpreted and implemented on the ground.

The Algorithmic Accountability Act (AAA) was recently reintroduced in Congress. This law, if passed, would empower the Federal Trade Commission (FTC) to institute rulemaking related to algorithmic impact assessments and other obligations on companies making use of AI. We focus on this regulatory proposal because it has several provisions that could both incorporate an understanding of the dataset development process, and create a comprehensive framework that includes many of the subjects and stages we have identified in this paper.

Section 3 requires the FTC to promulgate regulations to impose certain obligations on companies. These regulations will address the ways consumers are impacted in areas such as education, employment, healthcare, and housing. It also states that the FTC may consider what is an appropriate assessment at each specific point in the technology


California Victim Compensation Board: payment of claims, 2022 CAL. STAT. 92 (establishes a number of requirements before use of automated decision-making systems by state agencies).


development life cycle. This is useful because it is necessary to identify what kind of impact assessment is needed at a specific stage. For example, assessments of pre-employment screening technologies will be distinct from post-employment worker management technologies. Each stage implicates a specific set of rights, and it is important that the FTC break down the stages of development so that a one-time assessment or one-size-fits-all approach is not seen as the norm.

Here, the dataset development matrix may be useful to identify the stages at which assessments should take place. Similarly, a recognition of the different stages of development is important when it comes to identifying covered entities. Recognizing that there are a variety of entities and stakeholders involved in the technology development process—starting from the curators who build the datasets used to train the models and entities which deploy those models, there is an entire ecosystem for the development of that technology. Each covered entity would then have a distinct set of capabilities when it comes to conducting impact assessments and clarifying how an application was deployed. The dataset development matrix could help covered entities and regulators identify both stages and subjects involved in the process.

Section 5 requires covered entities to include in the assessment an explanation of why a "critical decision being made and the purpose" for it. Here it is important to clarify the criteria that the FTC will use to evaluate whether such an explanation is adequate. For example, in the employment context, if the use of a machine learning is being made to make the hiring process more "efficient", is that considered enough of an explanation? Moreover, what is the criteria for assessing what efficiency means, and at whose expense? Relatedly, the problem formulation and data collection stages of dataset development could shed important light here for what drove the creation of a critical decision-making system in the first place.

This section stipulates a "transparency, explainability, contestability, and opportunity for recourse" for consumers. These are distinct mechanisms for AI accountability; implementing each would require more careful thought. For example, transparency is tricky to implement because it is not always clear what a consumer can ask for, and what exceptions to disclosure may apply such as trade secrets.355

355 See Mike Ananny & Kate Crawford, Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability, 20 NEW MEDIA & SOC’Y 973 (2016); Burrell, supra note 333.
However, stipulating more clear avenues for individual involvement through explainability and contestability is a welcome step. The individual course of action to allow contestation is rooted on the principle of anti-discrimination based on a protected trait. Therefore, it is important to consider how groups may be able to leverage their protected status to use the right to contest. Oftentimes, distinct marginalized groups are impacted by machine learning systems as a class, especially in cases of bias and stereotyping. To that end, they should be able to act as a group to contest them, rather than as individuals.

Section 6 leaves the publishing of impact assessment results to the discretion of covered entities. Since information about an AI system could be proprietary, often protected by trade secret law, this protection would have to be balanced with the public need for information and transparency, there should be a provision to create a mechanism for requests for information, especially for research purposes, by the public.

Section 7 would establish a public registry on AI. This registry would need to have clear guidelines on how datasets are shared, especially since many large datasets are made available to the public, and as discussed above, often without any terms or conditions. This registry would be a way to systematically include information about the dataset development process, and could also be a way to keep track of how open datasets are used and deployed in private corporate settings.

C. Summary

Some common themes emerge in the normative and legislative proposals that we have analyzed here. First, both legal scholars and the AAA attempt to work around the inscrutability of AI systems by compelling entities to make information about them more readily available and accessible. Second, they work to empower both individuals and marginalized groups to demand more attention and action to violations based on constitutional rights, due process, and discrimination. Lastly, these proposals desire to raise the standards on what the bare minimum requirements ought to be for companies developing and deploying AI systems. In this Part, we have made

connections between how these normative and legislative updates may incorporate the dataset development framework as a starting point for individual contestation, explainability, and systemic impact assessments.

CONCLUSION

In this article, we provide an overview of the dataset development process and why it should be central to discussions about AI regulation. Our analysis shows that developing a dataset accountability framework should be attentive to the subjects involved, the stage of the dataset development process, and a combination of constitutional, privacy, and copyright law.

There are four key takeaways we offer for further analysis. First, we highlight the overall opacity of data collection, cleaning, training, and distribution practices, which leads to lack of clarity about the liabilities and rights of different subjects. Second, we underline the need for legal and ethical assessments at various stages of development and an appreciation of the connectedness of the stages and subjects. Third, we need to expand how we encompass and understand harms, particularly issues emerging from copyright practices, novel privacy harms, and who may be held accountable for each type of harm. Lastly, we discuss the application of our framework in current normative and legislative proposals. The dataset development framework provides an analytic frame to understand what these system-level stages are, what the issues may be, and who is implicated.

***