From Human to Data to Dataset: Mapping the Traceability of Human Subjects in Computer Vision Datasets

MORGAN KLAUS SCHEUERMAN, University of Colorado Boulder, USA
KATY WEATHINGTON, University of Colorado Boulder, USA
TARUN MUGUNTHAN, University of California, Berkeley, USA
EMILY DENTON, Google, USA
CASEY FIESLER, University of Colorado Boulder, USA

Computer vision is a "data hungry" field. Researchers and practitioners who work on human-centric computer vision, like facial recognition, emphasize the necessity of vast amounts of data for more robust and accurate models. Humans are seen as a data resource which can be converted into datasets. The necessity of data has led to a proliferation of gathering data from easily available sources, including "public" data from the web. Yet the use of public data has significant ethical implications for the human subjects in datasets. We bridge academic conversations on the ethics of using publicly obtained data with concerns about privacy and agency associated with computer vision applications. Specifically, we examine how practices of dataset construction from public data—not only from websites, but also from public settings and public records—make it extremely difficult for human subjects to trace their images as they are collected, converted into datasets, distributed for use, and, in some cases, retracted. We discuss two interconnected barriers current data practices present to providing an *ethics of traceability* for human subjects: awareness and control. We conclude with key intervention points for enabling traceability for data subjects. We also offer suggestions for an improved ethics of traceability to enable both awareness and control for individual subjects in dataset curation practices.

CCS Concepts: • **Human-centered computing** → *Empirical studies in collaborative and social computing*; • **Computing methodologies** → **Computer vision**; • **Social and professional topics** → **Intellectual property**.

Additional Key Words and Phrases: Datasets, computer vision, data subjects, data ethics, machine learning

ACM Reference Format:
Morgan Klaus Scheuerman, Katy Weathington, Tarun Mugunthan, Emily Denton, and Casey Fiesler. 2023. From Human to Data to Dataset: Mapping the Traceability of Human Subjects in Computer Vision Datasets. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 55 (April 2023), 33 pages. https://doi.org/10.1145/3579488

1 Introduction

Conversations about AI, machine learning, and computer vision often focus on the potential for harm, including how datasets shape model outputs. However, arising before ethical problems caused by dataset use (e.g., bias, poor documentation, lack of explainability, insufficiency of retraction) are issues surrounding data sources and methods of collection. Many common computer vision datasets consist entirely of images depicting real people, their images scraped from the web, without their

Authors' addresses: Morgan Klaus Scheuerman, morgan.scheuerman@colorado.edu, University of Colorado Boulder, Department of Information Science, CO, USA; Katy Weathington, Katy.Weathington@colorado.edu, University of Colorado Boulder, Department of Information Science, CO, USA; Tarun Mugunthan, tarun_mugunthan@berkeley.edu, University of California, Berkeley, School of Information, CA, USA; Emily Denton, dentone@google.com, Google, NY, USA; Casey Fiesler, Casey.Fiesler@colorado.edu, University of Colorado Boulder, Department of Information Science, CO, USA.

2573-0142/2023/4-ART55
https://doi.org/10.1145/3579488

knowledge or consent. While it is difficult to quantify the pervasiveness, scraping publicly available data—and even collecting data from public physical spaces without subject knowledge (e.g., Duke MTMC [17, 79])—are common tactics for building computer vision datasets. Web scraping has been used to create some of the largest and most controversial examples, such as ImageNet, Tiny Images, and MS-CELEB-1M. Platforms like Flickr, YouTube, Instagram, have become a robust data resource for researchers across industry and academia.

We can imagine how this practice might impact the hundreds to thousands of humans subjects in those images. For example, consider the hypothetical Jordan, an events photographer; they upload a portfolio of their work to Flickr, an online image hosting website. Their account is filled with photos of weddings, family birthday parties, and live concerts—hundreds of images of people celebrating moments large and small. Both Jordan and their subjects are unaware that those images have been scraped by multiple researchers and aggregated with other Flickr users' images into multiple datasets. Datasets for facial detection, scene understanding, gender classification, and even facial beauty ratings all include Jordan's images, and the faces contained therein. We can imagine how Jordan's subjects go on to be used to fuel computer vision research across industry and academia. Some models may be deployed commercially, the data used to train them thus contributing to millions of dollars in sales. Years after, some of those datasets have disappeared, their creators silently retiring them; but copies still exist in other data repositories and the images still exist within models circulating in academic and production settings. Even if Jordan's subjects were aware of their images being used in one dataset, how could they trace even a single dataset's life to all the other places it has ended up?

Those whose likenesses are featured in computer vision datasets, like Jordan's subjects, not only have little control over their image once it has been collected and converted into a part of a dataset, they may not even have the awareness it is happening. Research ethics review bodies, such as the Institutional Review Board (IRB) for U.S. universities, typically do not require review for research using public data [65]. Prior research has also indicated that research use of "public data" is often unknown to the original creators of that data (e.g., [29, 33, 37]). For example, interviews with photographers whose Flickr photographs appeared in IBM's facial recognition dataset revealed displeasure with their content being used this way without consent [74]. While some recent projects, such as exposing.ai [41], attempt to make finding one's image in a computer vision dataset more transparent, such projects are often technically limited to matching usernames on the original site—in this case, Flickr—with those in the dataset. There is limited legal precedent to give data subjects agency over removal, and, even if a dataset developer explicitly offers a process of subject removal, the downstream use of an individual's likeness is opaque.

The concerns about public data use by researchers unearthed in prior work indicates a desire for data subjects to know if, how, and who is using their data. Computer vision, which is fundamentally shaped by data, presents a timely opportunity to understand challenges to tracing the use of public human data. Specifically, computer vision uses visual data, which is often more personally identifying and intimate than many types of textual data. Given that machine learning requires large amounts of visual data, particularly for those looking to build huge generalizable models [11], computer vision researchers collect hundreds to thousands of identifiable human faces. Moreover, the dataset development lifeycle presents unique challenges to data subject awareness and control. Identifiable human faces spread beyond their initial use for one study or model, to be used in many other studies and models, generally without subject knowledge.

Drawing from Olsen and Borit's definition of *traceability* as the ability to access recorded identifications of a piece of information throughout its lifecycle [75], we consider traceability in the context of datasets as the ability for one to trace a single piece of data throughout a dataset's lifecycle. Given concerns about data subject awareness in prior work on research ethics for public

data [92], we consider what best practices might be for the treatment of data subjects in computer vision datasets by examining current practices and the challenges they pose. Much like Peng et al. [81], we trace the practices of computer vision dataset development, from their original data to dataset dissemination and use. However, rather than focus broadly on the lifecycle of a few datasets, we systematically examine moments of transformation within the dataset development pipeline where the human data subject is fundamental—and also becomes increasingly difficult to trace. Specifically, we conducted a content analysis of 125 unique computer vision datasets that stem from public data, either from the web, from physical public spaces, or from public records. Employing both structured content analysis and qualitative content analysis, we present findings that describe dataset curation processes: where data is often collected from, what kind of data subjects are often featured in datasets, and how those datasets are disseminated to research communities.

We discuss how attending to the processes involving human data subjects problematizes the traceability of individual data subjects throughout the computer vision dataset lifecycle. In doing do, we aim to advance conversations of ethics and transparency for machine learning data beyond research practices focused on issues like reproducibility, trustworthiness, and stability (e.g., [1, 30, 61, 87]) to address what we call an *ethics of traceability*: the issues surrounding data subject *awareness* of their data usage and the possibility of *control* over their data. Beyond augmenting prior literature in social computing promoting better data subject privacy and agency (e.g., [25, 40, 51]) and the social implications of datasets (e.g., [22, 66, 90]), we contribute *key points of intervention* across the dataset curation pipeline for dataset authors to attend to issues of traceability. We propose considerations for enabling both awareness and control on behalf of the data subjects featured in datasets along these intervention points.

2 Related Work

2.1 Computer Vision Datasets

Computer vision is focused on building computer systems that have the capability to metaphorically "see"—to analyze, classify, and describe patterns of information in visual data. Most modern computer vision systems are built using machine learning methods and are thus reliant on datasets, collections of visual data for teaching specific tasks. For example, collections of face images to teach a model face detection. Some computer vision datasets have been developed for highly specific tasks, like medical image analysis and fishery classification. Other datasets are scoped much more broadly, with an aim of providing "comprehensive and diverse coverage of the image world" [20].

The data that a computer vision model is exposed to shapes how the model treats new unseen data. Given the centrality of datasets to computer vision, dataset bias has become a major focal point in fairness and ethics scholarship. Many examinations of biased or otherwise harmful model outputs (e.g. [10, 13, 23]) or the larger ecosystem of machine learning, from problem specification to domain shift post-deployment (e.g., [18, 77, 98]), also highlight issues at the data-level.

At a high level, research on datasets has largely focused on subject distribution, process documentation, and dataset values. Analyses of distribution are concerned with the balance of diverse subjects within a dataset: whether a certain subject group is represented [77] (e.g., is a certain ethnicity missing from a facial recognition dataset?) and how many subjects are in that group [24, 112] (e.g., do white faces far outnumber Black faces?). Further, the categories employed in datasets have been found to simplify the world and erase certain subject groups [53, 90].

Process documentation has focused on *how* datasets are being documented and areas which are opaque or poorly documented. Prior scholarship has identified an absence of consistency or best practices in documenting dataset processes, resulting in a lack of transparency, trustworthiness, and reproducibility [31, 87]. While a number of frameworks have offered potential standards

for datasets (e.g., [30, 47]), inconsistent standards has been found to stem from organizational constraints, including differing organizational priorities and power differentials [67, 68], leading to calls for interventions at the process level, rather than at the documentation level.

Finally, there has been increasing attention to dataset values: not de-biasing or balancing distributions nor improving documentation methods, but examining and questioning the meaning and the role of datasets. Such scholarship has scrutinized the moral and ethical implications at the annotation level, questioning how categories exclude or demean certain groups and whether those categories should be included in computer vision practices at all (e.g., [9, 10, 52, 88, 97, 103]). More broadly, dataset documentation has been found to communicate specific values, such as objectivity, which otherwise displace the human subject [14, 80, 87].

Much of the prior work cited above focuses specifically on implications for dataset authors to improve creating and documenting datasets. Yet how datasets become ethically problematic is also not stable, and changes over time as datasets are put to use and potentially retracted [17, 81]. In this work, we center the human subject in our analysis of documentation, focusing less on the distributions, processes, or values of the dataset itself, and zeroing in on the areas of the dataset lifecycle that introduce difficulty for subjects to understand if and how their data is being used.

2.2 Ethics, Traceability, and Public Data

Alongside the analysis of computer vision datasets are larger conversations about the ethics of using public data for research purposes, particularly as the use of data scraped from the web becomes more ubiquitous [83]. Research ethics continue to be a salient topic for the CSCW community (e.g., [21, 27, 105]). Social computing researchers have focused particular attention on the ethics of using online user data for research projects, particularly around issues of consent and expectations of privacy (e.g., [28, 78, 113, 114]).

Regardless of the specific purpose of scraping public data, there are no clear legal standards for if and when it is appropriate to do so. While sometimes resulting in a "Terms of Use" violation of specific websites, or a copyright violation on behalf of the copyright holder of an image, web scraping is unlikely to otherwise be a violation of the law in the United States [26], though case law is evolving (e.g., [108]). Moreover, web scraping also offers tools for auditing and making it illegal altogether might give content providers too much power [2]. Therefore, conversations in the research community have focused more on whether scraping for specific types of research is ethical. Many ethical review boards, like U.S. IRBs, do not consider scraping public web data to be human subjects research, and therefore exempt such studies from ethical review [104].

Researchers have thus sought to understand how people feel about their web data being used by academic researchers. Researchers have found that the public is largely unaware of their data being used for research purposes [29, 39, 92]. When made aware, how people feel about their data being used for research is largely dependent on everything from the data type to its intended use to who is doing the collecting [34]. Recent work by Zimmer and Logan surveyed the general public about use of public data for suicide risk prediction algorithms, indicating potential public concern for machine learning use cases [114]. However, there has been less "participant" focused research regarding perceptions uses of public data for computer vision, specifically. It has largely been academic researchers and journalists surfacing concerns about subject agency and informed consent [10]. Prior work has focused on perceptions of specific computer vision applications, unearthing concerns about uses cases like facial recognition (e.g., [12, 91]) and identity classifications (e.g., [8, 38]). Further, journalists have covered the use of public data for machine learning, likely raising the awareness of the public to data scraping for computer vision (e.g., [48, 59**?**]).

The synthesis of concerns about computer vision uses and the role of public data have led to explorations for improving the explainability of how data drives model outputs (e.g., [111]),

mechanisms of resisting deployed models (e.g., [54]), and local and national legislation aimed to protect people from non-consensual facial recognition datasets (e.g., [43, 95]). Some projects, like exposing.ai [41], have also attempted to make datasets which use Flickr data searchable by username, so that copyright holders can find their images.

Projects like exposing.ai touch on an area of ethics that has largely been absent from conversations about public data use and machine learning: *traceability*. Traceability is defined by Olsen and Borit as "the ability to access any or all information relating to that which is under consideration, throughout its entire life cycle, by means of recorded identifications" [75]. Traceability, as a practice in ethics and accountability, is well-established in the field of food science, where food products are traced across the supply chain (e.g., [94]).

Data provenance also has a rich history in library sciences and archival studies. Data provenance work has primarily focused on detailed documentation and linking of research artifacts for promoting experimental reproducibility and assessing scientific claims (e.g., [6, 60, 85, 109]), including increasingly in the age of big data and automation [19]. The concept of traceability has also been employed in software engineering, generally for the sake of accountability to established system requirements and the ability to examine data relationships [36]. Yet, there remains a need for tracing the people who are being represented in datasets, not for the sake of research credibility but for the sake of the data subjects themselves.

The traceablity of public data can also be seen in social media research on re-identification and de-anonymization (e.g., [4, 15]), designing notification-based opt-out systems (e.g., [115]), and participatory data mapping which links data instances to standardized documentation (e.g., [110]). Bates et al. proposed a methodology, data journeys, for tracing flows of data between different sites of practice, with a major focus on how social worlds become interconnected as data flows between them [5]. Data journeys, and associated concepts like data flows [63] and the "Follow the Data" interview protocol [63], provide methods for understanding the practices and relationships of socio-historical human actors in shaping, producing, and reusing data.

In this work, we adopt Olsen and Borit's notion of traceability as the ability to trace a single artifact—in this case, a data subject—through its entire lifecycle through examining publicly available documentation. We choose to adopt Olsen and Borit's approach to traceability in food science as it focuses on the safety of the consumer, rather than the reproducibility or trustworthiness of scientific experiments. We focus specifically on how dataset development shapes how human data subjects can be accessed, adopted, and used. By adopting a traceability approach to data, we augment prior work on the ethics of public data by focusing on the areas of the dataset construction pipeline that are particularly difficult for potential data subjects to trace and contend with. In particular, we delve more deeply on how the practices of dataset authors throughout the dataset lifecycle present unique challenges to the awareness and agency of data subjects.

3 Methods

3.1 Computer Vision Dataset Corpus

Our goal was to understand the points throughout the computer vision dataset lifecycle where visual data of human subjects gathered from public data sources is processed: collected, indexed, translated, used, adapted, etc. Therefore, we needed to establish a corpus of computer vision datasets that feature human subjects and specifically use public data: web data, publicly collected data, or public records. To create a corpus from which to sample, we first started with the corpus of image-based computer vision datasets compiled by Scheuerman et al. [87].[1] Scheuerman et al. compiled the corpus of 753 computer vision databases by manually sampling from computer vision

[1]available at https://zenodo.org/record/3735400#.Ybjd_L3MKF6

conferences in IEEE. We continued to build on this corpus for two reasons: to provide a more comprehensive computer vision corpus to other researchers and to increase the number of datasets collected from public data sources. We built on the corpus by augmenting it with two more data sources: https://exposing.ai/ and https://paperswithcode.com/.

exposing.ai is a project which allows the public to search for their image in six computer vision datasets[2] that scraped data from Flickr, a photo sharing website. We chose to use exposing.ai because its goal as a tool for improving traceability by allowing human subjects to find themselves in existing datasets which use Flickr data uniquely fits the context of the current study. Some of the databases from exposing.ai were already present in the Scheuerman et al. [87] corpus.

PapersWithCode is an open source and community-maintained repository of machine learning papers, code, datasets, and benchmarks. We chose to use PapersWithCode because it is considered an invaluable resource for both state-of-the-art methods and datasets in machine learning, and may be the largest and most up-to-date database of computer vision datasets online. We downloaded all of the image and video datasets from PapersWithCode, then removed those already present in the corpus. We then used the paper titles downloaded from PapersWithCode to automatically scrape the MLA formatted citation, year, citation count, and Google Scholar link.

The final corpus we compiled includes 2,227 datasets.[3] For each dataset, we include the MLA formatted citation for the original dataset paper, the Google Scholar link, the number of citations on Google Scholar, the date the number of citations were checked, the venue the original paper was published in, and the original date the paper was published.

3.2 Sampling for Analysis

We sampled 125 datasets for deeper analysis. As we intend to contribute to conversations around the ethics of using human data from public sources, we only sampled datasets which used public data. We defined public data as data scraped from the Internet, data taken from public records, data taken in real world public settings (e.g., on the street or college campuses), or data derived from other publicly available datasets. We also only sampled datasets which included human data, such as face or full body images. We sampled in four ways. First, we sampled the top ten most cited datasets that used public data.[4] Second, given the exposure a more general public might have to them, we included the remaining five exposing.ai datasets. Third, we kept the 22 datasets derived from public data that were coded for in [87]; we felt it would be useful to review those datasets through the lens of public data scraping for differing real world motivations. Finally, we randomly sampled the remaining 88 datasets, checking each one to ensure that the source dataset authors used to create their dataset was public data. We decided to round out our corpus with random sampling because it is a straightforward and simplistic method to obtain an unbiased selection of datasets, and we had already performed purposive sampling to ensure we got highly cited and critiqued datasets [99]. In cases where we randomly sampled from the larger corpus of datasets and got a dataset which did not include human data, we went back and randomly sampled a new dataset until we reached the goal of 88 randomly sampled datasets (for 125 datasets total). Our four sampling strategies ensured that our sample included: popular and commonly employed datasets; controversial datasets; datasets potentially familiar to a more general public; datasets which have been previously examined in the CSCW community; and less commonly used smaller datasets.

[2]DiveFace, FaceScrub, IARPA Janus Benchmark C (IJB-C), MegaFace, People in Photo Albums (PIPA), VGG Face
[3]available at https://doi.org/10.5281/zenodo.7535600
[4]MSRA10K, PASCAL Visual Object Classes (VOC), Caltech-101 Object Categories Dataset, VGG Face, CelebA, INRIA Person, MS-COCO, Labeled Faces in the Wild (LFW), Sports-1M, ImageNet

#	Dataset	#	Dataset
1	10K US Faces	64	LFWgender
2	300 FACES IN-THE-WILD CHALLENGE (300-W)	65	Lopes et al.
3	Abstract Paintings / Artistic Photographs Datasets	66	MALF
4	Acted Facial Expressions In The Wild (AFEW)	67	Market-1501
5	Activity Net	68	MegaFace
6	Adience	69	MS-Celeb-1M
7	APPA-REAL	70	ModaNet
8	BAVL (Blind Audio-Visual Localization)	71	MORPH
9	Beauty 799	72	MOT16
10	Berkeley Segmentation Data Set (BSDS300)	73	MPII Human Pose
11	Caltech-101 Object Categories Dataset	74	MS-COCO
12	CAMO++	75	MSR-VTT
13	CASIA-WEBFACE	76	MSRA10K
14	Celeb-DF	77	ND-IIITD Retouched Face Database
15	CelebA	78	Nis Web-Collected Database
16	Celebrities in Frontal-Profile (CFP)	79	NIST Mugshost Identification Database (MID)
17	Compaq Dataset	80	NPDI Pornography-800
18	CROSS-AGE CELEBRITY DATASET (CACD)	81	Open Images V4
19	CrossTask	82	OUI-Audience
20	CVC-14	83	Paper Doll
21	DeepFashion	84	PASCAL Visual Object Classes (VOC)
22	DiveFace	85	Penn Action
23	DPC-Captions	86	People in Photo Albums (PIPA)
24	DukeMTMC	87	Pilot Parliaments Benchmark (PPB)
25	EMOTIC (EMOTIons in Context)	88	People in Social Context (PISC)
26	EVVE	89	Planned Event Dataset
27	FACEBOOK100	90	PTZ Tracking
28	FaceForensics	91	Real-World Affective Faces Database (RAF-DB)
29	FaceScrub	92	Real-World Masked Face Dataset (RMFD)
30	FAD (Face Attributes Dataset)	93	SBU Captioned Photo Dataset
31	FAMILIES IN THE WILD (FIW)	94	SCUT-FBP-A
32	Family101	95	SOBA (Shadow-OBject Association)
33	Fashion144K	96	Social Event Dataset (SED)
34	FDST (Fudan-ShanghaiTech)	97	Sports-1M
35	Flickr1024	98	Stanford Region Labeling Dataset
36	Flickr30k	99	SUN
37	FVI	100	TGIF
38	Grand Central Station Dataset	101	The Mobiface Dataset
39	Gray Dataset	102	TikTok Dataset
40	Gun Detection Dataset	103	Tiny Images
41	Helen Facial Feature Dataset	104	TLL
42	HICO (Humans Interacting with Common Objects)	105	Tri-Subject Kinship Face Database (TSKINFACE)
43	Hipster Wars	106	Twitter100k
44	Hot-Or-Not Database	107	UAVDT
45	HowTo100M	108	UCF101 Human Actions Dataset "
46	HRT Transgender Face Database	109	UMass FDDB
47	Humans In 3D (H3D)	110	Unsplash2K
48	i-LIDS	111	VGG Face
49	IARPA Janus Benchmark C (IJB-C)	112	Viewpoint Invariant Pedestrian Recognition (VIPeR)
50	IdentifyMe	113	Vimeo Creative Commons Collection (V3C)
51	IG-1B-Targeted	114	Visual Genome
52	IIITD Plastic Surgery Face Database	115	WhoIsIt (WIT) Face Database
53	Illicit Drug Abuse Face Database	116	WIDER FACE

(Continued on next page...)

#	Dataset	#	Dataset
54	ImageNet	117	Win-Fail Action Understanding
55	IMDB-WIKI	118	WWW Crowd
56	INRIA Person	119	XD-Violence
57	JHMDB (Joint-annotated Human Motion Data Base)	120	Yahoo! Creative Commons 100 M Database (YFCC100M)
58	Kinetics-700	121	Yahoo's Safe for Work (SFW) or Not Safe for Work (NSFW)
59	KinFaceW	122	Yoga-82
60	Labeled Faces in the Wild (LFW)	123	YouTube-100M
61	Labeled Faces in the Wild-A (LFW-A)	124	YouTube-VOS
62	Large Age-Gap (LAG)	125	YT-BB (YouTube-BoundingBoxes)
63	Leeds Sports Pose		

Table 1. The table shows all of the datasets in our sample, listed in alphabetical order. Each dataset has an associated reference number to its left. In the Findings section, we often list each dataset that falls under a certain category. For ease of reading, we use the reference number for each dataset in parentheticals.

3.3 Codebook Development

Our initial codebook was also adapted from Scheuerman et al. We pared down the codebook in [87] to focus primarily on issues of data collection and dissemination practices, such as whether and how authors license their datasets when making them public. As we began coding the datasets around broad notions of collection and licensing, we discussed what other variables might be interesting to capture that were not already present in the codebook. As coding evolved, we began to focus more clearly on areas where the data subject's traceability becomes opaque or difficult. We began to add new categories to code for such as the status of the dataset (whether it was retracted or not), the type of organizations the authors did the work at (e.g., academic institutions), and, if the dataset authors licensed their dataset, what that license was meant to prohibit. Details of our codebook can be found at https://doi.org/10.5281/zenodo.7535600.

3.4 Analysis

We conducted a content analysis on the sample of 125 datasets for moments where the data subject becomes salient. We employed content analysis because it is a flexible method for examining documents both qualitatively and quantitatively [107]. We looked for specific parts of the dataset lifecycle where human data subjects are central to the creation and dissemination of the dataset. We examined their role in dataset collection, licensing, use, and retraction. We did not explicitly code from the perspective of a data owner or user. Rather, we focused on areas of the dataset lifecycle where issues of traceability might arise for a data subject who wants to track the use of their data and exercise control over it. We use the findings of our documentation analysis to theorize which moments of the dataset lifecycle would be difficult for a data subject to manage.

For each dataset, we examined a number of artifacts: the original paper the dataset was proposed or introduced in, the website the dataset was hosted on (if available), and the dataset itself if sufficient information could not be gleaned from other documentation. We employed both thematic coding and structured coding. The team divided the coding of the sample, with the first author coding 62 datasets and the second, third, and fourth authors coding 21 datasets. Given the qualitative and subjective nature of the coding process, which could then be disputed or interpreted differently by different research approaches, we decided against utilizing formal interrater reliability methods [62]. Instead, we met regularly to discuss questions, confusions, and salient themes.

Thematic coding was focused on identifying, understanding, and interpreting how authors documented aspects of data collection and dissemination. We also examined what potential gaps or vagueness might communicate about traceability issues. The thematic coding process included taking notes on pieces of text within the artifacts which fit into the codebook. For example, when

coding for licensing, the authors took notes on the intersections of both the license of the original data scraped from the web and the license proposed by the authors for their dataset (e.g., a dataset of Creative Commons images then being copyrighted by the dataset authors). Thematic coding largely informed the structured coding as themes arose we believed would be useful to measure.

Structured coding consisted of defining specific variables to code for. Structured coding was inductively derived from the process of thematic coding. One instance of structured coding might be absence-presence. For example, whether the dataset was retracted (yes/no, or N/A for those datasets which could no longer be found but were not explicitly retracted). Another instance might be fixed categories which we could then measure. For example, what types of restrictions each data license communicated. We defined bucketed variables by first thematically coding, then grouping like themes into larger categories. For example, a restriction like *"The MMLAB is not responsible for the content nor the meaning of these images"* (CelebA) would be bucketed into the larger theme of "no legal liability of authors." In the findings, structured coding is denoted by descriptive statistics. We then use the original thematic coding to further describe these statistics, and provide example quotes from documents.

Once all coding was completed by all team members, we then met to discuss and resolve disagreements. Finally, the first author went through each of the 125 datasets and checked that the coding of the entire team matched the expectations determined through regular meetings, as well as to normalize all structured codes. After writing up our findings, our team then met to synthesize each finding through the lens of traceability. We speculate about the barriers that a data subject would face, given an attempt to trace their data through the dataset lifecycle. We identified two major barriers to the agency of data subjects: (1) awareness of how their data is collected, transformed, disseminated, and used; and (2) the ability to enact control over these processes given they have awareness. We present the synthesis of our findings through our Discussion.

3.5 Access to Research Materials

Computer vision research datasets are generally open source or are otherwise readily available for people to download. While we problematize the notion of gathering human data without consent or knowledge, we also hope to encourage deeper critiques and engagements with existing computer vision datasets. Therefore, we feel it is ethically responsible to share our corpus, sample, and codebook, and encourage others to use and build on our work. We do not share the datasets themselves, but the names of the datasets, links to the datasets, and the publications they were originally proposed in, when applicable. We similarly encourage those concerned about their data use to use our corpus as a resource, albeit we acknowledge the structure of our corpus (links and names) and of the datasets themselves makes it, as we will reveal in this paper, incredibly difficult to trace individual subjects. The materials are available at: https://doi.org/10.5281/zenodo.7535600.

4 Findings

We summarize our systematic analysis of issues relating to data traceability in four sections. First, we focus on aspects of data collection, including data sources, subject types, and consent processes. Next, we examine considerations that relate to converting data to a dataset and releasing it, such as dataset availability and licensing. We summarize key findings relating to data collection and conversion in Table 3. Next we discuss issues relating to dataset use, including model usage and the development of derivative datasets. Finally, we offer a deep dive into the 5 retracted datasets in our corpus.

Data Collection		
Original data sources	Public websites	78.4%
	Prior datasets	20%
	Public spaces	8%
	Public records	2.4%
Subject type	Includes regular people	80.8%
	Includes celebrities and/or public figures	28.8%
Mention of consent process		1.6%
Mechanism for removal of data		2.4%
Mentions of original data licensing (for 109 web-scraped datasets)		33%

Packaging Data into Datasets		
Dataset availability	Dataset available to freely download	60%
	Download limited by access agreement	26.4%
	Retracted or unable to locate	11.2%
Datasets released with license or terms of use		49.6%
Common terms of use	Forbidding of commercial use	38.4%
	Forbidding of use without attribution	18.4%
	Absolving authors of legal liability	12.8%
	Forbidding redistribution	12%
	Forbidding ethics or privacy violations	2.4%

Dataset Use	
Datasets pubilshed with modeling contribution included in paper	7.2%
Datasets derived (in part or full) from prior computer vision datasets	20%
Derivative datasets that mention original data lisence	14%

Table 3. Summary of findings relating to dataset collection and packaging. Totals do not always add up to 100% because some datasets include multiple variables (e.g., both regular people and celebrities).

4.1 Data Collection

Data collection is the first step in creating a new dataset. Choices made at the data collection stage can make traceability more difficult after a dataset has been created and released. In this section, we outline four areas where dataset collection impacts traceability: (1) original data sources, or where dataset authors get the data from; (2) the subject type, who is being included in the dataset; (3) whether the data subjects were asked for their consent to be included in the dataset; and (4) the original licensing governing the data that dataset authors collected. The goal of this section is to showcase the current practices in dataset collection and how those practices make traceability more difficult for a data subject.

4.1.1 Original Data Sources

We examined the data sources to understand the variety of places that dataset authors are sourcing their data subjects from, especially given arguments that data subjects should be aware and able to enact control over their data pre-collection [44, 113]. We found a vast variety of data sources: datasets were derived from 82 unique sources (see Table 5). 98 datasets sourced data from public websites; 22 from prior computer vision datasets; 10 from public spaces; and 3 from public records. Many datasets sourced data from multiple sources. For example, 9 datasets collected data from both websites and prior computer vision datasets (3; 9; 12; 23; 25; 40; 77; 88; 95). 1 dataset used data from all three categories (Gun Detection Dataset). Table 5 shows the many data sources authors used to create datasets. The vast variety of sources, including those in public spaces, suggest that data subjects' awareness of collection is likely impossible. These sources also do not have mechanisms

for tracking who is downloading or scraping your data. In the cases of datasets that source from broad sources, like Google and Bing, what websites the data was actually hosted on is absent from the documentation. This lack of information is most extreme for datasets which only stated they sourced from "the Internet" or "the web."

Public Data Category	Type of Source	# Using Type of Source	Specific Source
Websites	Web search engines	27	Google (20); Bing (7); Unspecified (2); Yahoo! Images (1); Picsearch (1); Ask (1); Baidu (1); Cyrdral (1); Webshots (1); Altavista (1)
Websites & Public spaces	Other	27	Medical websites (4); Movies (3); Online shopping websites (Forever21, Mogujie) (2); Paintings (1); Author personal images (1); Crowdsourced images (1); memebase (1); facere-search.org (1); unnamed yoga website (1); deviantart (1); Prelinger Archives (1); unnamed pornography websites (1)
Websites	Video sharing sites	26	YouTube (24); TikTok (1); Vimeo (1); Google Videos (1)
Websites	Unknown	24	Unspecified entirely (4); "The Internet" broadly (20)
Multiple	Prior Datasets	22	MS-COCO (5); YFCC100M (3); LFW (2); ADE20K (2); IAPS (1); SBU (1); CelebFaces (1); Pond5 (1); YouTube-BoundingBoxes (1); YouTube-VOS (1); Video Anomaly Detection Dataset (1); MegaFace (1); Visual Genome (1); Names and Faces in the News (1); Shanghai Dataset (1); PubFig (1); Salido et al. Dataset (1); Corel (1); PaperDoll (1); Graz (1); XM2VTS (1); LVIS (1); PASCAL (1); University of Notre Dame, Collection B (1); Geometric Context (GC) (1); AVA-Plus (1); MSRA (1); Database-5 (1); MRSC (1); ImageNet (1)
Websites	Online photo albums	19	Flickr (19)
Public spaces	Public spaces	10	College Campus (3); Unspecified (3); UAVs (1); CCTV Camera Footage (1); Train station (1); Office (1)
Websites	Social media	7	Twitter (3); chictopia (3); Tumblr (1); Instagram (1); Facebook (1)
Websites	Informational websites	4	IMDb (2); IMFDb (1); Wikipedia (1)
Public records	Public records	3	Mugshots (2); Yahoo! News (1)
Websites	User rating websites	3	Hot-or-Not (2); AgeGuess (1)
Websites	Stock image websites	2	Unsplash (1); Getty Images (1)

Table 5. The twelve Types of Sources, the number of datasets using each Type of Source, and the 82 unique Specific Sources each dataset used (with counts in parentheses). The left column lists which of the three Public Data Categories that the data came from (as discussed in 3.1). We note that it is difficult to dilineate Public Data Categories cleanly. Prior datasets may include every category; "other" datasets encompass multiple categories; and "unknown" sources could include any of the categories.

4.1.2 Subject Type

We wanted to understand what types of human subjects were commonly used for computer vision datasets. In particular, we wanted to know whether regular people or celebrities/public figures were more commonly used. The majority of human subjects in the datasets were regular people (101 datasets; 80.8%). 28.8% of datasets included celebrities (e.g., movie actors, singers) (19 datasets; 15.2%) or public figures (e.g., politicians) (17; 13.6%). Three datasets featured people of unknown origin (17; 89; 94) and two featured fictional (video games or animation) characters (40; 104). Only one dataset (UAVDT) did not feature recognizable human faces, given it was collected using unmanned aerial vehicles (UAVs) from above. Given that regular people were most commonly used for data subjects, the traceability of data subjects may be a more pertinent issue. It is more expected for celebrities likeness' to be used without it being a violation of their privacy; part of their role is being visible in the public eye. On the other hand, regular people may have a higher expectation of privacy. However, what distinguishes a public figure from a regular person is difficult to discern, given dataset authors employ the concept of public figure variably. Some people labeled public figures by dataset authors (e.g., journalists) may not consider themselves public figures.

4.1.3 Consent

Though all the datasets in our sample were derived from public sources, the dataset creators may still have been able to gain consent for image use. Therefore, we sought to understand whether dataset authors asked their subjects for consent when using their data. When describing the process of data collection, only two dataset authors (1.6%) mention consent: FACEBOOK100 and Flickr1024. FACEBOOK100 got permission from the data subjects themselves (50 subjects and their friends; it is unclear whether the friends gave permission), while Flickr1024 got permission from the image copyright holders. However, the copyright holders may not be the same as data subjects, and there is no guarantee that the data subjects themselves were aware and consenting. The remaining 123 datasets do not mention the consent of either subjects or copyright holders. In one case, consent to use images was obtained by the site administrators of the website from where the data was sourced,[5] but not from either copyright holders or subjects (TLL).

Beyond those explicit examples of consent, whether data subjects or copyright holders were informed of the data collection process was not mentioned by any dataset authors. The authors of the Gray Dataset, Illicit Drug Abuse Face Database, and IIITD Plastic Surgery Face Database justify the scraping of images from the web since people upload their images voluntarily (although, whether people upload their images to websites like Hot-or-Not voluntarily is debatable):

> *"Users who submit their photo to this site (Hot-or-Not) waive their privacy expectations and agree to have their likeness criticized."* —Gray Dataset

Three datasets had a mechanism for data subjects to be removed. Two datasets allowed the copyright owners to contact them to remove images from the dataset: Cross-Age Celebrity Dataset (CACD) and IMDB-WIKI. One dataset explicitly mentions allowing data subjects to have their images removed (Celeb-DF), stating: *"If you feel uncomfortable about your identity shown in this dataset, please contact us and we will remove corresponding information from the dataset."* However, the lack of processes for consent or notifying data subjects of their inclusion in a dataset in the first place would make it difficult for data subjects to know they are in a dataset to request removal. The majority of datasets did not have clearly outlined processes for data removal.

[5] http://memebase.cheezburger.com/totallylookslike

4.1.4 *Original Data Licensing*

Original data licensing refers to the copyright or licenses associated with the images scraped for datasets. We examined whether dataset authors made any references to the licenses governing the original data, as licensing is currently viewed as the most reliable method for ethical and legal data use. Of the 109 datasets using scraped data from the web (not public records or publicly taken photographs), 36 (33%) mentioned the the licenses or copyright pertaining to the original data. The remaining 73 (67%) did not mention the original licenses or rights governing the data used. While the lack of mentioning original licensing does not necessarily mean that dataset authors did not correctly adhere to licenses, it provides no information for dataset users to know if the dataset complied with the data's copyright or licensing.

Of the 36 that mentioned the original data's licensing, 22 (1; 3; 6; 13; 14; 18; 21; 23; 26; 29; 35; 36; 52; 54; 55; 56; 74; 80; 91; 111; 122; 124) mention that the original copyright belongs to the image owners. 7 (21; 23; 26; 29; 52; 54; 111) chose to link to the original image URLs or IDs instead of providing a copy of each image. Providing URLs or IDs over copying the images and reproducing them seemed to be a means for dataset authors to avoid copyright violations; however, how such violations might apply to using the data for modeling did not come up. By linking to URLs (rather than providing images directly), people are theoretically more in control of their data because they can remove the source image. However, in practice, researchers frequently download a full copy of the dataset once and then do not have a practice of checking to remove local copies of images that have been removed from the source. Further, the traceability of image use is increasingly difficult when the data in datasets is unstable. The authors of FAD, which use images from PubFig, describe how some of the images from PubFig were lost due to the original URLs being removed:

> *"Due to copyright issues, original images were never provided for the PubFig Dataset, and only the respective internet addresses (URLs) were given. Since the release of PubFig, many of those URLs have become invalid, so we focused on the subset of images of the original data which are still available online."* —FAD (Face Attributes Dataset)

8 datasets (82; 49; 42; 113; 120; 68; 86; 81) chose to use only data licensed under Creative Commons. Creative Commons provides a number of licenses for copyright authors to choose from, giving instructions on how that data can be used. For example, a copyright author may allow their images to be remixed with attribution. We note that Creative Commons licenses also have individual requirements for use and attribution which may not have been followed for each individual piece of content. The authors of V3C explicitly sought Creative Commons videos for the dataset. They were also the only ones to mention Terms of Service restrictions as well:

> *"Vimeo was chosen over YouTube because while YouTube offers its users the possibility to publish videos under a creative commons attribution license which would allow the reuse and redistribution of the video material, YouTube's Terms of Service ... explicitly forbid the download of any video on the platform for any reason other than playback in the context of a video stream."* —Vimeo Creative Commons Collection (V3C)

While mentioning the original license enables some trust in being able to use the dataset ethically, we also note that expectations of use of Creative Commons images by the original data owners may differ from machine learning use cases (e.g., [45]). Further, properly using licensed images may adhere to principles of legality, but may still violate data subject awareness and consent.

4.2 Converting Data into Datasets

Once the data has been gathered, it is then converted into a dataset, often with some form of annotation (labels, bounding boxes, facial points, subject identifiers, and so on). The collection of data and annotations—the dataset—is generally treated as a separate entity from the data itself, in

that those who created the dataset, what we have referred to as dataset authors, claim ownership of the dataset. They then license and distribute the dataset for specific purposes. In this section, we describe the variety of availabilities, licenses, and prohibitions associated with datasets. Specially, we discuss (1) how authors make their new dataset available and (2) the licensing and limitations authors then put on the available dataset. Converting data into datasets is a moment of transformation, in which individual data subjects become available to others for use in research and commercial projects. This moment of transformation, from data to dataset, makes tracing individual data subjects increasingly opaque and unwieldy, largely because dataset authors have not set up any mechanisms for tracing dataset use and license violations.

4.2.1 Dataset Availability

It is common practice in computer vision for datasets to be developed as community resources for model development and benchmarking. As such, once the data has been converted into a dataset, dataset authors frequently make it available to others to download and use.

How a dataset is made available influences whether the dataset's use can be easily tracked. Dataset access is sometimes limited by an access agreement—a form which a potential user of the dataset must agree to abide by in order to access the dataset. For example, the WhoIsIt (WIT) Face Database requires a user to fill out a digital form with a name, signature, date, organization, and address then email it to the dataset authors. This form stipulates a number of items the user must agree to in order to gain access to the dataset, such as that the dataset is *"valuable intellectual property"* and thus *"the researcher(s) shall have no rights with respect to the Database or any portion thereof."* Theoretically, having to fill out an access agreement form to access the data would leave a paper trail of who has used the dataset. However, that paper trail is in the hands of the dataset authors. We did not come across any lists of who has accessed the data and for what. Obtaining a list would require requesting one from the original dataset authors (e.g., [41]).

In our sample, 75 datasets were available to download freely, without requiring the user to fill out any form of access agreement. 33 datasets required some form of access agreement to access all of the dataset. 3 datasets (5; 15; 55) allowed some data in the dataset to be downloaded freely while other data required an access agreement. Freely being able to download a dataset would leave no paper trail, and thus tracing a data subject to every endpoint of the data's use would be impossible. Further, these datasets are hosted on a range of platforms, from custom websites to Google Drive to GitHub. These websites are not persistent and thus the data may be removed or disappear, as we discovered during our analysis. 14 datasets were unavailable, either due to retraction or the inability to locate the dataset. Yet missing and retracted datasets may still be in use or stored in a user's personal repositories. There are no reliable or systematic methods for tracing where freely available, missing, or retracted datasets end up.

4.2.2 Dataset Licensing and Prohibited Uses

Dataset licensing indicates the terms that dataset users must abide by when using a dataset. Examining dataset licenses provides insight into what rules dataset users were expected to abide by when using data. The datasets in our corpus were licensed in a variety of ways. Of the datasets which we could locate online, 62 datasets had a license which posed some restrictions of use and 49 datasets mentioned no restrictions of use.

Some licenses were standardized, such as: Creative Commons (17 datasets: 12; 20; 29; 36; 47; 58; 68; 70; 74; 81; 88; 97; 102; 111; 114; 124; 125), BSD (3 datasets: 19; 26; 73), Apache (1 dataset: 97), and MIT (1 dataset: 25) licenses. Other datasets used customized licenses created by the dataset authors. Customized licenses varied in their terms. These terms imposed restrictions through a set of prohibitions specific to each dataset. The most common terms were: prohibiting commercial

use (48 datasets), prohibiting usage without proper attribution (23 datasets: 2; 14; 21; 29; 40; 46; 47; 50; 52; 58; 70; 74; 77; 79; 80; 81; 82; 87; 88; 97; 111; 115; 125), absolving authors of legal liability (16 datasets: 13; 14; 15; 22; 31; 54; 56; 65; 68; 74; 77; 80; 82; 95; 102; 124), and forbidding redistribution (14 datasets: 15; 21; 22; 25, 31; 35; 46; 50; 52; 79; 80; 87; 91; 115). Licenses generally act as a mechanism for transferring ownership and credit of the original images to the dataset authors.

The number of restrictions imposed also varied between datasets. For example, WWW Crowd dataset simply stated, *"These data can only be used for University research purposes,"* while the ND-IIITD Retouched Face Database explicitly prohibited redistribution without permission, commercial use, privacy and ethics violations, using more than a certain number of images from the dataset, and any legal liability of the authors. Some derivative datasets borrowed license terms from their sources; e.g., the Labelled Faces in the Wild- A (LFW-A) dataset required users to agree to the terms of use of its parent dataset, LFW. Flickr30k required users to abide by Flickr's terms of use.

Only three datasets had terms of use prohibiting ethics or privacy violations. The Pilot Parliaments Benchmarks terms of access prohibited use that "*violates the rights or privacy of the subjects depicted*"; ND-IIITD Retouched Face Database prohibits use that "*could cause the original subject embarrassment or mental anguish.*" The now-retracted HRT Transgender Face Database similarly prohibited usage causing the subjects "*humiliation, harassment, or mental anguish, or be perceived in a false light.*"

Licensing practices indicate a focus on attribution and absolving the authors of any legal liability, rather than on the ethics or privacy of data subjects. Of the 23 datasets requiring attribution, 10 mentioned the original data's licensing (14; 29; 47; 52; 74; 80; 81; 82; 87; 111); 13 (56.5%) did not (2; 21; 47; 50; 58; 70; 77; 79; 88; 97; 115; 125). Of the 16 datasets absolving authors of legal liability, 9 mentioned the original licensing (13; 14; 54; 56; 68; 74; 80; 82; 124); 7 (43.8%) did not (15; 22; 31; 65; 77; 95; 102). Dataset authors, who may have violated the consent, attribution, and the legality of the original data subjects and dat owners, are interested in maintaining their own rights in regards to the dataset.

Dataset licensing can aid in creating a paper trail (e.g., required attribution leading to traceable citations) and limiting places the dataset might be found (e.g., a dataset should not be used in commercial systems if the license forbids it). While the approach to licensing may violate awareness and control during the data collection process, licensing may help to create awareness after the collection process. Of course, the benefits of licenses rely on users abiding by them, and that is not always the case (e.g., [41]). Ideally, violations of licenses (e.g., redistribution without permission, ethical and privacy violations) would also allow dataset owners—and potentially dataset subjects—some recourse, though none of the licenses state how recourse for violations is attained.

4.3 Dataset Use

Once a dataset has been created and disseminated, it can be used by others for a variety of reasons. We focus on two major use cases in this section: (1) the use of the dataset in machine learning models; and (2) the use of the data in the dataset to create new datasets, which we call dataset derivatives. We also highlight that dataset derivatives may be created for entirely new domains, separate from the domains intended by the original dataset authors. Once a dataset is in use, the data instances of individual data subjects may branch exponentially, depending on the dataset's popularity. Use of the data in modeling also transforms the data from a visual object to model features, which can be deployed on new unseen subjects. Dataset derivatives replicate data in new datasets, which may be governed by entirely different availability mechanisms and licenses. These derivatives can then be used in new ways and new domains.

4.3.1 Model Use

Once a dataset is used for modeling purpose, its reach extends beyond simply containing data subjects' information to actively using that information on new unseen data subjects. Only 9 (7.2%) datasets in our corpus were not used for modeling in the papers they were originally introduced in (60; 67; 72; 79; 92; 96; 113; 120). Most datasets are built by the authors for specific modeling problems, and are associated with either a model or a challenge (an open call to use the data to achieve the best possible modeling results on a specific problem). Those datasets are then used by other researchers and practitioners in models both small and large. Use is difficult to track via citations, given papers citing datasets may simply be describing the dataset or building on their work without utilizing it. Looking at citations showcases the difficulty of parsing how datasets are used in modeling and whether the data are being used as intended by the dataset authors and their licensing terms. For example, ImageNet has been cited 34,630 times on Google Scholar, as of January 2022. Many uses may not even be documented in academic publications; datasets like ImageNet are also employed outside academic contexts and also commercially [32], making it impossible to track use through mechanisms like citations. Further, a single data subject's influence on a model is opaque. Even if one were to trace themselves to a specific model, actual data removal is difficult, especially for larger or commercial models [35].

4.3.2 Dataset Derivatives

We examined dataset derivatives because they add a new "branch" to tracing a data subject in one dataset and its uses to tracing a data subject to an additional dataset and its uses. 22 datasets sourced data—partially or in full—from prior computer vision datasets. Only 3 (14%) datasets that derived data from prior datasets explicitly discussed the licenses governing use of the parent dataset (22; 61; 68). While some derivative datasets inherit the same licenses as their parent datasets (e.g. all datasets in our corpus derived from YFCC100M maintained the Creative Commons licensing that YFCC100M was released under), many derivative dataset licenses differed from the parent dataset licenses. Differing licenses add complexity to how and for what purposes datasets can be downloaded and used.

Some derivative datasets are sourced from pre-existing datasets but introduce new annotations or new transformations of the original images. For example, several offshoot datasets have been derived from Labeled Faces in the Wild (LFW): Labeled Faces in the Wild-a (LFW-a) contains LFW images that have been aligned using a commercial face alignment software; and LFW-gender contains LFW images that have been annotated with machine-produced binary gender labels. In another example, MSRA10k offers pixel-level semantic segmentations for images from the MSRA dataset. Other datasets source their images from a dataset but filter, process, and clean images so as to achieve certain desirable dataset characteristics in the new dataset. For example, MegaFace images are all sourced from the Yahoo Flickr Creative Commons 100 Million (YFCC100M) dataset. An automated face detection system was used to filter out images not depicting faces, and images posted to accounts with a small number of images were filtered out to maximize the likelihood of having multiple faces of the same identity.

Several derivative datasets sourced images from an amalgamation of sources, including multiple pre-existing computer datasets, image search engines, and social media websites. These types of derivative datasets also add new data subjects who were not present in the prior datasets. Further, those that derive from multiple sources may have to contend with multiple data licenses and differing data subject expectations. For example, PISC (People in Social Context) sourced images from three pre-existing datasets (Visual Genome; MS-COCO; YFCC100M), Flickr, Twitter, and image web search engines (e.g. Google Images and Bing). Some derivative datasets relied on pre-existing datasets to source images for a particular category. For example, Gun Detection Dataset sourced

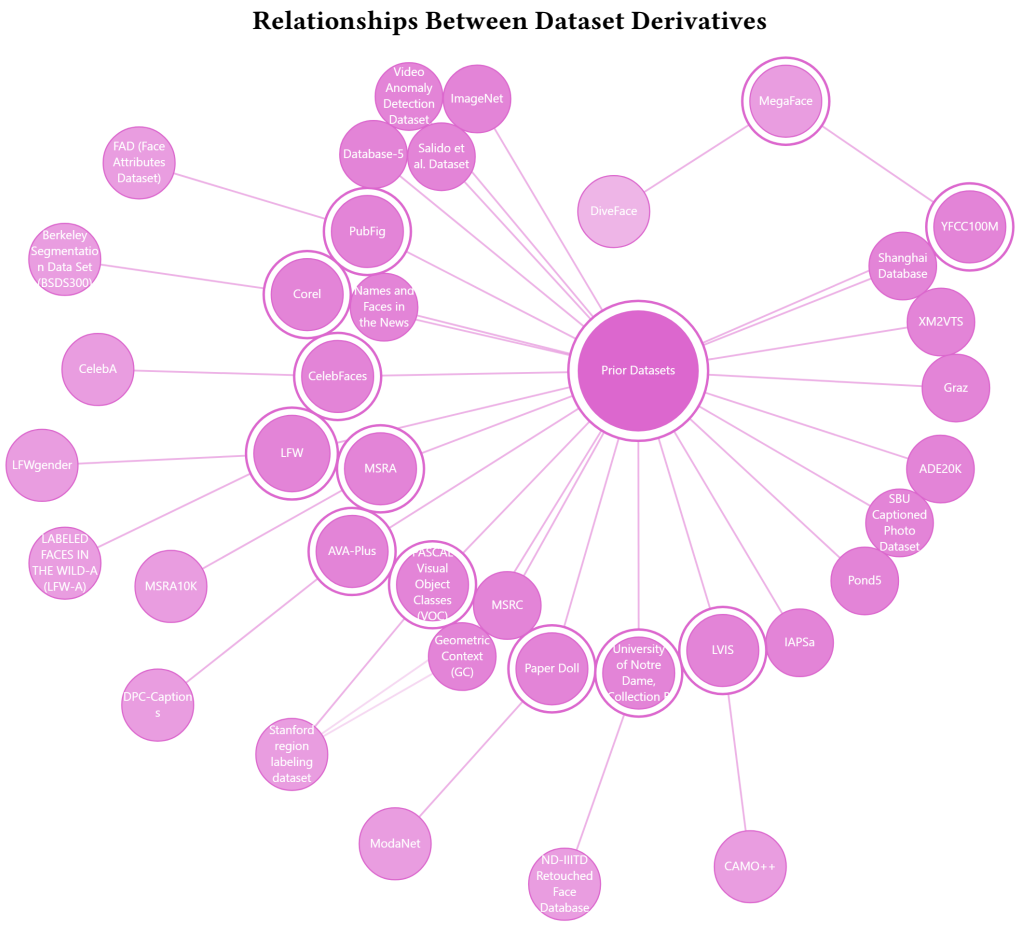

Fig. 1. The figure above shows the relationships between the datasets in our corpus which use data from prior datasets. Notes surrounded by a circular border represent "parent" datasets which other "child" datasets are derived from (e.g., the child ModaNet is derived from the parent Paper Doll.) There are also datasets with multiple levels of derivation (e.g., YCC100M is the parent to MegaFace which is the parent to DiveFace). Finally, there are datasets which act as parents to multiple other datasets (e.g., LFW is the parent to children LFW-A and LFWgender). A full interactive chart of dataset connections can be found at: https://www.morgan-klaus.com/traceability/dist/index.html

gun images from prior weapon and anomaly detection datasets outside our corpus and sourced non-gun images from the well-known image classification datasets ImageNet and MS-COCO.

We also uncovered several examples of multiple levels of derivation. For example, DiveFace is sourced from MegaFace which is itself sourced from YFCC100M; PISC is sourced from Visual Genome which is itself also sourced from YFCC100M; LFA, which has been the source of two datasets in our sample, was itself derived from an earlier dataset (not in our sample), Faces in the Wild, that was released with noisy and incomplete labels. Broadly, dataset derivatives make the traceability of a data subject more difficult. Multiple levels of derivation further compound the issues of traceability that occur with dataset derivatives.

4.3.3 *Derivative Domain Shift*

As previously stated, many datasets are proposed for specific computer vision tasks. For example, Adience was proposed for age and gender classification, while Beauty 799 was proposed for face beautification. We observed derivative dataset tasks and source datasets tasks may be closely related. For example, DiveFace was derived from MegaFace and both are face recognition datasets. FAD, proposed for gender, ethnicity, and race classification, sourced images from PubFig, a facial analysis dataset which analyzes gender, but also other categories not present in FAD. LFWgender, intended for gender classification, was entirely derived from LFW, intended for facial recognition.

However, in other cases, tasks of derivative datasets differ significantly from the intended task of the source dataset. For example, the object recognition and scene understanding dataset, MS-COCO, has been used to source images for 5 other datasets in our sample (25; 40; 88; 95; 114). The use cases of these 5 derivatives ranged from emotion classification to shadow detection. The object recognition dataset, YFCC100M, has been used to source images for 3 other datasets in our corpus (68; 88; 114), with use cases including face recognition and the classification of social relationships between people in images. Domain shift may result in datasets being used in new modeling domains, again creating further branches of dataset and modeling relationships to attempt to trace. In the hypothetical cases where data subjects consented to their data's use in a specific dataset or domain (given we only found one dataset in our sample), domain shift may also violate data subject expectations, perhaps without their awareness.

4.4 Dataset Retractions

Dataset retraction is when authors remove a dataset so that it can no longer be downloaded from its original source. This is sometimes accompanied by a statement of why the dataset is being retracted and that people should no longer use it, but not always. Of the 125 datasets we examined, 5 had been retracted: HRT Transgender Database, MegaFace, MS-CELEB-1M, DukeMTMC-reID, and Tiny Images. Retracted datasets are being increasingly discussed because of ethical issues around the data still being used: so-called "runaway data" [81] or "'zombie datasets" [17]. The traceability of retracted datasets becomes difficult as the original datasets are removed, but copies continue to crop up from third parties and be put to use. The scenarios surrounding the development, use, and retractions of these datasets are highly varied. In this section, we detail the retraction and issues of continued use of the five datasets in our sample.

4.4.1 *MegaFace*

The MegaFace Database was made public by authors at University of Washington. It was created using images from the dataset YFCC100M, which used Flickr data, in order to collect faces beyond the common use of celebrity images. The original purpose of the dataset was to accompany a "challenge." Once sufficient accuracy on the challenge benchmark was satisfactorily reached, the authors of the dataset no longer desired to maintain the servers it was hosted on. Citing time and monetary costs of maintaining the infrastructure of MegaFace, the dataset was decommissioned in June 2020 and is no longer available for use. Prior to its retraction, there was also critical media coverage in 2019 over Creative Commons license violations, as well as use by corporate and government entities using MegaFace for surveillance and policing [41].

The images were used in a dataset derivative called DiveFace, meant for ethnically diverse facial recognition; since MegaFace's retraction DiveFace now uses the original images from YFCC100M,

which are the same images MegaFace. On Google Scholar, 180 papers cited MegaFace in 2020[6] and 166 papers cited it in 2021.

4.4.2 Tiny Images

Tiny Images was a dataset of 80 million images, each 32x32 pixels. It was created because small or low resolution images present a challenge to computer vision due to the limited visual information they can contain. The authors argue that, in order to compensate for the lack of information in any singular image, a very large number of images was required. The dataset authors therefore took a highly automated approach, first scraping nouns from WordNet as categories, then scraping images based on those nouns.

However, the reliance on automation with lack of oversight allowed derogatory terms to be used at categories, and the inclusion of offensive or harmful imagery [10]. Following the audit of Tiny Images from [10], the dataset's creators released a statement of retraction. They explained that due to the high number of images and small size, manual review of images for problematic content would be unfeasible, and that they could provide no guarantee of removing every offensive image. Thus, the creators made the decision to fully retract the dataset, stating *"biases, offensive and prejudicial images, and derogatory terminology alienates an important part of our community – precisely those that we are making efforts to include"* [101].

The retraction statement specifically asked that people refrain from using Tiny Images in the future and asked that any saved copies of the dataset be deleted. However, the dataset was created in 2006, and was not retracted until June 29th, 2020, so there existed a significant window during which the dataset was accessible and thus could still exist on a number of machines and within working models. The dataset was cited 148 times in 2020 and 145 times in 2021.

4.4.3 HRT Transgender Database

The HRT Transgender Database contained images of transgender people before and after transitioning[7], scraped from YouTube transition timelines without explicit consent from the individuals in the videos. Following a researcher posting about the dataset on Twitter, there was a slew of critical media coverage about the collection of it [48]. While no statement was made by the authors, the dataset disappeared from the web. Given there was no explicit retraction, we could not identify the date it was removed, and thus, how many academic papers have cited it since.

4.4.4 MS-CELEB-1M

MS-CELEB-1M contained images of celebrity faces, with some non-celebrity distractors mixed in. However, the definition of "celebrity" used here was broad, to the point where they included activists, researchers, journalists, and individuals with a professional online presence of any kind. MS-CELEB-1M was also found to have violated some copyrights according to exposing.ai [41].

MS-CELEB-1M was found to be used in Chinese surveillance programs, including the mass detention of Uighurs in Xinjiang, in April 2019 [70]. The dataset was quietly retracted in June 2019 [69]. As documented by other scholars, MS-CELEB-1M was used in a number of derivative datasets and models, and can be torrented from third parties [17, 81]. It had also been used internally by Microsoft on its own proprietary project after retraction [41]. It was cited 239 times in 2019, 337 times in 2020, and 394 times in 2021.

[6]We did not filter papers by publication *after* the specific date datasets were retracted, so some papers were citing it before its retraction. We revisit the gap of interpretation the lack of specific dates leaves in the Discussion.

[7]The process transgender individuals undertake to shift from a gender role and presentation separate from the one they were assigned at birth.

4.4.5 DukeMTMC

The DukeMTMC dataset consisted of videos of students walking through campus taken from a variety of perspectives. The collection process was found to have violated the submitted and approved IRB [86]. Like MS-CELEB-1M, the DukeMTMC dataset was found to be used in anti-Uighur surveillance programs [86]. Also like MS-CELEB-1M, the dataset was retracted without public statement in June 2019. However, the faculty supervisor of the published work issued an apology for violating the IRB [100]. DukeMTMC was cited 337 times in 2019, 405 times in 2020, and 552 times in 2021.

5 Discussion

Through our findings, we have showcased the documentation of current dataset curation practices for a sample of computer vision datasets that use public data, whether from the web, public records, or public physical spaces. We highlighted the steps taken in the dataset curation pipeline, from data collection, converting data to a dataset, dataset use, and the occasional dataset retraction. Using the areas of the dataset pipeline that we identified as pertintent to human data subjects, we will now theorize how the practices documented along the pipeline present challenges to data subjects. To do this, we shift perspective from researcher to data subject. The areas of the pipeline in our findings that presented issues of opacity are reimagined as barriers to tracing one's own data.

The ability for a data subject to trace their own data throughout the dataset lifecycle is what we call *traceability*. We highlight two major issues in current dataset practices preventing data subject traceability: *awareness* and *control*. We argue that these barriers present an issue to an *ethics* of traceability, an approach to dataset creation which would give data subjects information and agency over how their data is used. We conclude by providing suggestions for data practices throughout the human-to-data-to-dataset pipeline that better enable both subject awareness and control.

5.1 Data of the Cave: Barriers to Awareness

In mapping the current landscape of empirical work on research ethics for public data, Shilton et al. detailed evidence that (1) data subjects largely lack awareness of how their data might be used for research, and (2) many are alarmed or upset when informed about research uses of their data [92]. Even when a computer vision dataset is publicly available, there are essentially no mechanisms in place for a data subject to know when data has been collected, where it has been collected from, or for what purposes their likeness is being used. Authors also do not report on alerting data subjects of data collection, which is further confirmed by reports on datasets like MegaFace [45] and DukeMTMC [86]. The lack of mechanisms in place to create awareness creates a sort of Allegory of the Cave,[8] in which a lack of awareness shapes that their online data is being collected and used in computer vision creates an inaccurate and incomplete understanding of reality. Expecting potential data subjects to dig through thousands of datasets and their derivatives to know whether their data is being used is not only an unreasonable task, it is a virtually impossible one. In this section of the Discussion, we highlight various barriers to awareness on the part of a data subject.

5.1.1 Expectations of Data Use and Data Licensing

Given that only 2 datasets in our findings referenced obtaining data subject consent (and not simply copyright holder consent, which may still indicate a lack of awareness of the person featured in

[8]The Allegory of the Cave by Plato is an allegorical story about how knowledge, or lack thereof, changes our perception of reality. In the story, people are chained to face a wall their entire lives. They can see shadows on the wall due to a fire behind them. The shadows represent the prisoner's perceptions of the real world, but they are incomplete and inaccurate representations of reality. Understanding true reality would require observing directly the forms casting the shadows.

the image), current data collection procedures indicate it is unlikely that subjects are aware of when data is being collected and where it is being collected from. Our findings also revealed a large diversity of sites for data extraction, where that extraction would likely violate user expectations.

Conversations around the ethics of using scraped data for research often include an assumption that data subjects are aware that by sharing content such as photos online, that content might be used in a variety of ways, including by researchers. However, prior work has shown that people are not only unaware of the general information flow of their social media content [82] and how it might be used by third parties, but also of research uses of data generally [92]. For example, in a survey of Twitter users, about 2/3 of respondents were unaware that tweets might be used for research purposes and almost half thought that this practice was not allowed [29]. This lack of awareness occurs despite a relevant provision in Twitter's privacy policy.

Indeed, researchers often rely on terms of service to justify data scraping when the terms do not explicitly prohibit it. However, it is unlikely that these policies have a strong influence on user expectations of how their data might be used, considering how rarely they are read and how difficult they tend to be to understand [26]. Moreover, it is extremely rare for such policies to mention external research uses explicitly; in one study's analysis of 100+ data scraping provisions, only one site mentioned academic research [26]. Therefore, even for datasets that rely on licensing or mention terms of service, this should not be relied on as a proxy for awareness.

We found that the vast majority of web-based datasets did not mention initial data licenses, either established by the platform itself or from the copyright holder. However, even in cases where licenses are established by the copyright holder, such as Creative Commons, prior research suggests that understandings of those licenses do not necessarily encompass research uses. Subjects might not understand that Creative Commons images may be used for machine learning when uploading their images to, for example, a family album on Flickr [**?**]. Other datasets use copyrighted material, but consist only of links to that material for others to download in order to avoid a copyright violations. The practice of linking to image URLs also indicates that copyright owners are likely unaware their data is being used in datasets. With respect to data collection in public physical spaces, legal rules around such collection of images are ambiguous or not well understood [55, 73]. There have also been examples of such data collection in public life inciting controversy [17, 86], implying expectation violation and a lack of awareness.

5.1.2 Dataset Use and Tracking

Even if licensing agreements or similar mechanisms did serve as means to awareness or even consent, the creation and dissemination of datasets presents a further challenge in that uses may be decoupled from collection. Once data has been converted into a dataset, issues of awareness shift from whether data is being collected to whether, how, and by whom it is being used. Dataset use has been increasingly scrutinized for a lack of standard mechanisms for dataset authors to follow who is using their data and for what purpose (e.g., [81, 87]).

Even in cases where dataset authors require a user fill out an access agreement (26% of our sample), whether authors track dataset use after granting access is unclear. Who was given access to the data is also in the hands of the dataset authors, who may or may not log the access agreement forms. Therefore, for data subjects, even if they become aware of the original collection of their data, it is possible that even the dataset authors could not keep them updated on the current status of that data. How data is disseminated through availability, licensing, and prohibitions shapes how easily data subjects can potentially trace their data. If a dataset is simply available with no access agreements, like the majority of datasets in our sample, then anyone can download the dataset and it is unlikely the authors have implemented mechanisms for tracking those downloads.

Access awareness also presents an issue of accountability to licensing agreements and whether terms are being violated or not. As Peng et al. discuss in [81], not only are licensing terms poorly defined in regards to commercial use, derivatives do not necessarily inherit the same license terms as their parent datasets. This means that if an original dataset outlines prohibitions, a derivative dataset may not prohibit those same actions.

Therefore, awareness of where data has ended up—whether in models or dataset derivatives—becomes extremely opaque to data subjects. As others have demonstrated, even when datasets are retracted, they are still found to be in use [17, 81]. For example, we described how MegaFace is cited 328 times on Google Scholar post-retraction. However, we cannot be sure that each paper citing MegaFace actually utilized the dataset itself without conducting an in depth review of each paper. This highlights another area of opacity on behalf of the data subject to identify where their data may be being used; manual inspection of citations is not a scaleable way to raise awareness.

Derivative datasets also extend the issue of tracking data use, as data subjects become embedded in new datasets. For example, YFCC100M (which is still available) was used as a basis for MegaFace (which is now retracted) and then MegaFace was used as a basis for DiveFace (whose annotations are still available and now rely on YFCC100M's images). Tracing a single data subject through each dataset and its uses would be untenable. Further, as datasets are used and derived there is often a domain shift—a dataset intended for one task begins to be used in another task. YFCC100M was intended for object recognition, but MegaFace authors began to appropriate the data for facial recognition; Duke MTMC-reID was stated as being intended for motion analysis, but was found being used by commercial Chinese companies for person re-identification [86]. Moreover, collectively these structures that produce a lack of awareness also prevent data subjects and copyright owners from taking action.

5.2 Pandora's Dataset: Barriers to Control

Once data has been collected, control of that data is difficult to exercise. Control refers to the data subjects' ability to exercise agency over their data and how it is used. We posit awareness as a precursor to control, both pre- and post-data collection, as a lack of awareness of data use would preceded an ability to enact control over that use. In other words, if someone does not even know that their photo is in a dataset, they cannot do anything about it. Awareness is a precursor to control in the same way that "informed" is a precursor to "consent." One must be aware of and understand what they are consenting to.

Given the lack of awareness present throughout the dataset lifecycle, agency becomes increasingly difficult to enact. For a data subject to attempt to exercise control over their data prior to its collection, they must first consider the reality that their data could potentially be collected. However, as we will discuss in this section, just because a data subject is aware of their data's use, does not mean they can actually enact control. Even in cases where a data subject is aware—even if not *informed*—we found there are largely no mechanisms for data subject control built into the dataset lifecycle.

In this section, we assume that a subject is somehow aware of their data being used. We then highlight how awareness does not resolve barriers to control over one's data. Once data is converted into a dataset, it becomes a Pandora's Box: once opened, unable to be contained. Thus, it becomes virtually impossible to enact full control over one's data. We discuss how laws that implicate data ownership provide little guidance for computer vision data, how licensing prioritizes dataset authors, and how the lack of standard procedures for data removal disempower subjects.

5.2.1 Content Ownership and Licensing

We previously discussed how the existence of licenses or similar legal mechanisms are unlikely to contribute to data subject awareness. However, these same mechanisms may currently be the

clearest path to subjects maintaining control of their data if they do gain that awareness. As legal scholar Amanda Levendowski puts forth specifically in the context of facial recognition [57], while awaiting laws that might regulate uses of the technology, copyright law may be the most effective effort of resistance for people who do not wish for their faces to be used in this way—and the same holds for the use of photos in any computer vision dataset. Due to the current state of legislation in the United States, copyright has frequently been the mechanism for addressing content used or shared without permission, even when privacy harms are clearly more significant—for example, in the case of nonconsensual pornography [42, 50, 57].

Some computer vision datasets have paid attention to issues of copyright and image ownership. For example, the Pilot Parliament Benchmarks dataset uses public domain photos [13] and the Flickr dataset uses only photographs that used Creative Commons licenses [57]. However, as we saw in our sample, dataset authors may simply link dataset users to copyrighted data, shifting away the responsibility of copyright violations to those who choose to download the data. Some dataset authors may have downloaded copyrighted data anyway, given most did not mention the original rights of the data (e.g., [**?**]). For example, one of the critiques of MS-CELEB-1M was its violation of copyrights in its data collection.

However, it is important to point out that copyright law provides control to the copyright *owner* and not to the *subject* in the image−−which is rarely the same person (except, for example, in the case of a selfie). Copyright protects ownership, not privacy. In the case that the copyright owner is different from the person featured in the image, attempting to exercise control over how one's likeness is being used in machine learning becomes an interpersonal conflict which, legally, prioritizes the rights of the copyright owner. Creative Commons themselves responded to critiques of the Flickr dataset with a statement that copyright is not the appropriate avenue to protect privacy, address research ethics, or regulate surveillance [64]. Levendowski, along with other legal scholars, have also put forth that most uses of copyrighted works to train AI systems likely constitute fair use, which means that a copyright owner would not have control over such third party use [56, 57]. These realities of copyright law limit the autonomy of both data subjects and copyright holders.

Even in the case of datasets that require access agreements, the majority of licenses in our corpus did not include clauses protecting data subjects from specific uses or potential misuse. Licenses were primarily designed to protect dataset authors, and were most focused on dataset ownership, commercial and non-attributed uses, and avoiding legal liability. Yet there were also no mechanisms—at least, publicly stated by the authors—for ensuring even these violations do not occur. While implementing licenses focused on protecting data subjects from misuses foreseen by the authors and developing means of tracing actual use would be an improvement, it would still not give control to the data subjects in defining data misuse or license terms. It is possible that a data subject may not be wholesale against their data being collected and used in all instances (e.g., as seen with Twitter data [29] or IBM Diversity in Faces [**?**], but are against specific uses.

To allow for subjects to enact control over their data's use, practices in dataset curation would have to incorporate subject input on final outcomes of licensing and availability. Implementing subject control might include gathering data on what each subject feels their data should and should not be used for, and creating contextual licensing around specific subject identifiers—even potentially dividing the dataset into different files and licensing each accordingly. While not included in our corpus, an example of such contextual licensing is in the Iranian Face Database. While it is unclear whether the authors consulted the data subjects themselves, the license agreement prevents publishing any images of women present in the database [72].

Yet, in the case of licensing, control remains largely in the hands of the dataset authors, who must make decisions about whether to incorporate subject input. Given subject input would likely limit the reach and size of the dataset, dataset curation practices would require a shift in the

values embedded in the discipline of computer vision—what Scheuerman et al. would label as valuing care over the efficiency of non-consent and contextuality over universal licensing [87]. Even the enforceability of licenses remains a challenge, especially when the violation of the license is unrelated to copyright. Beyond improving the licensing practices of dataset authors, closing the gaps in legal systems to protect subject rights is also necessary.

5.2.2 Procedures for Data Subject Removal

Most dataset authors in our sample did not have explicit procedures for removing data, particularly for the data subject and not the copyright owners. Data subjects might instead have to attempt informal procedures for having their data removed from original datasets, such as emailing lead authors listed in papers or contacting those maintaining the data on the web, like a website administrator or GitHub user. Otherwise, as stated earlier, in the case of datasets made of URLs from copyrighted images, subjects might have to remove their own images from the source.

Yet these mechanisms of control are also largely limited in the case of dataset derivatives and model use. Like with awareness, when a dataset is published and put in use, exercising control over its uses and data replications becomes increasingly difficult. Having one's data removed from the original dataset it was collected for does not guarantee it will be removed from models using that data, downloaded copies, or derivative datasets. Once a dataset has been retracted, ideally the data subjects' data in it is no longer in use. However, this has been found to be untrue by numerous scholars [17, 81]. The data subjects themselves might still have to reach out to original dataset authors, derivative authors, and those using the dataset (perhaps by analyzing citations) to ensure the data is actually no longer in use and, if they desire, has been destroyed. Understanding whether data is used and thus how and who to contact can be even more concerning for commercial models, like those in Yahoo's Safe for Work (SFW) or Not Safe for Work (NSFW) or MS-CELEB-1M, because it's possible the use of data made it to production models.

The strongest protections for data instance removal, in either datasets or models, would be for EU citizens under GDPR, given non-consensual identifying data is not allowed to be used and consent can be retracted at anytime [102]. However, removing data from trained models can be particularly difficult. There is the technical limitation that data removed from a model may potentially remain present in the model—what Papernot et al. call "implicit memorization" [76, 93]. Features learned from a person's data may be retained by a model initially trained on that data, even if the data itself is removed from the model. Such memorization is especially problematic when that data contains sensitive information, like names or personal records. The only way to ensure features learned from the data is fully removed is to entirely retrain a model, a process that is argued to be inefficient, costly, and undesirable for companies and researchers [35, 49]. Such limitations raise interesting questions about data ownership and control when a model has already gleaned useful information from a person's data, even when that data is removed.

5.3 Considerations for an Ethics of Traceability

Our findings show the steps of the dataset curation pipeline that are pertinent to human data subjects, and we discussed how current practices in this pipeline inhibit data subject traceability. We defined traceability as the ability to access specific data instances through a dataset's lifecycle. As showcased by the non-standardized practices of documentation in our findings, the difficulties data subjects face are not linear. Issues of awareness might arise after a dataset has already been created, and issues of data control might arise before data has been collected. In this section, we promote developing a research ethics for dataset subjects that aligns more closely with expectations for human research subject awareness and autonomy.

We argue that both awareness and control must be present to properly incorporate an *ethics of traceability*—allowing data subjects to be informed and aware of their data use and to exercise control over their data throughout the dataset curation pipeline. We propose that awareness comes before control; in order for a dataset subject to exercise control over their data's use in computer vision datasets, they must first be aware their data might be used, and then how and where it is being used throughout the pipeline.

We thus augment prior work—largely focused on making data practices more transparent, ethical, and reproducible by centering dataset authors (e.g., [30, 68, 79, 81, 87])—by focusing on potential interventions in the dataset curation process that centers data subjects. Even while we center the data subject in this work, dataset authors still maintain the power to shape how data practices enable or disable subjects' awareness and control. Given the nonlinear nature of awareness and control, and the need for awareness before being able to enact control, we outline potential interventions that dataset authors can incorporate into their dataset curation practices for every step of the pipeline outlined in the Findings. Considerations for intervention are listed in Table 6.

5.3.1 *Current Limitations to an Ethics of Traceability*

The public discourse and regulatory landscape regarding the acceptable use of public data is currently evolving. As such, our considerations for intervention should not be understood as concrete; rather, we offer opportunities for thinking about an ethics of traceability. We acknowledge that these considerations are both idealistic (e.g., getting consent from every data subject) and limited in scope (focused on implementation by dataset authors). We also recognize that the issues raised in this work require solutions that are beyond the scope or influence of a singular dataset author. Dataset authors may have difficulty handling data misuse and data retention on behalf of data users. Even more broadly, the lack of international standards around data licensing [81]; the nebulous and vague landscape of legal recourse for copyright infringement and data misuse [96]; and the rapidly changing landscape in defining fair use, misuse, and harm in the context of machine learning [58] are all major barriers that dataset authors cannot change on an individual level. That said, individual researchers and practitioners can also begin to shift the norms of public data use within computer vision.

Indeed, several efforts to mitigate the concerns raised in this work have been proposed within the machine learning community. For example, recently developed dataset documentation frameworks (e.g. [30]) ask dataset authors to clearly document licensing details and consent processes. The machine learning conference Neural Information Processing Systems required authors to describe whether and how consent was obtained from data subjects (for both pre-existing and new datasets) [9]. While not currently integrated into standard practice, established ethical guidelines, like the Belmont Principles, can potentially provide ethical guidance for machine learning practitioners [71]. There have also been efforts to standardize the licensing of machine learning datasets [7]. Finally, there have been efforts to sidestep or minimize the use of image data depicting people entirely. For example, Asano et al. develop a new image dataset depicting no people that is suitable for model pre-training in a manner comparable to ImageNet [3]. Moreover, the authors seek to address copyright issues by ensuring their dataset only contains images licensed under creative commons and with complete attribution metadata. Other efforts have focused on the development and use of machine-generated datasets in an effort to address privacy and data access concerns (e.g. the ICLR workshop on Synthetic Data Generation[10].

[9]https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist
[10]https://sdg-quality-privacy-bias.github.io/

Considerations for an Ethics of Traceability in Datasets

Phase	*Step*	*Considerations*
Collection	Data Source	**Awareness**: Make it clear to all data subjects where the data is being collected from and when. **Control**: Develop opt-in recruitment methods (e.g., build a tool for users to upload their photos directly from the target source.)
	Subject Type	**Awareness**: Strictly define what subject types are included in a dataset (e.g., what makes someone a celebrity or public figure). **Control**: Allow subjects to self-tag classification (e.g., subject type, gender, hair color; methods could include drop-downs, or open-text that is then normalized).
	Consent	**Awareness**: Obtain consent from the data subject, regardless of copyright ownership. If consent is not possible for all subjects, remove or anonymize those subjects who did not consent (e.g., blurring faces of bystanders). **Control**: Build in explicit and clear methods for data subjects to have their data removed from the original dataset.
	Original Data Licensing	**Awareness**: **Control**: Do not link to copyrighted data for others to download. Do not violate copyright authors' or data subject's licenses.
Converting	Dataset Availability	**Awareness**: Track and publish who is accessing the dataset and for what purposes. **Control**: If a data subject wishes to be removed from a dataset, communicate to those using the dataset to remove that subject as well.
	Dataset Licensing	**Awareness**: Make terms of licenses clear. Incorporate terms of data subject privacy and downstream mechanisms of control into dataset licensing. **Control**: Co-create licenses with data subjects. Allow data subjects to opt into different license types (e.g., a dataset divided into subsets of license terms, like those who allow domain shift).
	Prohibited Uses	**Awareness**: Regularly check in on those using the dataset to ensure compliance with terms. Track and publish any license or terms of use violations. **Control**: Source prohibited uses from data subjects.
Use	Model Use	**Awareness**: Make obvious the current state of models using the data (e.g., use in commercial systems, decommissioned after publication). **Control**: Design a tool which not only alerts data subjects of the misuse but gives them a template for contacting the user.
	Dataset Derivatives	**Awareness**: Track and publish all dataset derivatives. Ensure their terms incorporate the same privacy and control mechanisms as the original dataset. **Control**: Require consent of data subjects for each derivative author.
	Domain Shift	**Awareness**: Define appropriate domain shifts in dataset documentation. Track and publish uses of original dataset that shift domains. **Control**: Co-create or empirically source approved domain shifts from data subjects.
Retraction		**Awareness**: Publicly announce retractions and the reasoning behind them. Reach out to tracked uses and dataset derivatives about retractions to ensure data is deleted. **Control**: Build a tool for subjects to publicly flag derivatives that have yet to retract data.

Table 6. Potential considerations for Awareness and Control for each phase and step of the dataset pipeline.

We also acknowledge some instances where crucial research might benefit from public data, so we encourage both deeper thought about how the benefits might outweigh the harms (e.g., [13, 84, 89]) and how to implement mechanisms for data subjects to opt out when datasets and/or models are deemed beneficial. A great deal of ethics and fairness in computer vision has focused on improving poor data diversity, particularly to counteract model bias (e.g., [13]. Yet practices for the diversification of data awarded by public data sources aren't without flaws. Even while there are proposed measures for dealing with biased data when building a dataset (e.g., [106]) or counteracting bias once that data is used for training algorithms (e.g., [16]), the overrepresentation of minority groups in public facing datasets may replicate perceptions of some groups as inherently more criminal. As Anna Lauren Hoffman points out, more inclusive datasets as the solution to biased technologies can mask or concede the harms inflicted by those technologies and neutralize criticism of additional harms that might be caused by data collection [46]. Raji et al. similarly note that efforts to increase the representation of people in datasets disproportionately impact marginalized groups, leading to tokenism, exploitation, and privacy violations, which can perpetuate marginalization [84]. Dataset authors should consider the experiences of their data subjects, like marginalization and vulnerability, when considering if using public data brings more benefit than good.

6 Conclusion

Scholarship has increasingly engaged with the ethics of using public data—data taken from online resources, which, in this paper, we also extend to data collected from public records and public real world settings. Prior research (e.g., [29, 33, 37]) and popular press (e.g., [48, 59, 74]) have noted the lack of awareness of data subjects that they are even subject to research. Alongside general scrutiny of using public data in research is the increased concern with using public data in computer vision, especially due to the propensity of harm associated with computer vision research and commercial products. Given the concerns about computer vision, a domain fundamentally shaped by its use of data, we examined how human subject data is collected, converted into a dataset, and then disseminated in 125 computer vision datasets. In our findings, we highlight the practices at each stage of the dataset curation process—collection, conversion, use, and retraction—that make tracking an individual data subject particularly opaque. In the discussion, we highlight how current practices undermine individual data subjects' *awareness* of and *control* over when, where, and how their data is being used throughout the entirety of the curation pipeline. We contribute intervention points for which dataset authors can better enable data subject traceability. Using these intervention points, we offer some potential considerations for dataset authors to better integrate mechanisms of data subject awareness and control.

Acknowledgments

This work was funded by the National Science Foundation (#1704303). We'd like to thank Anthony Pinter, Michael Ann DeVito, Michael Zimmer, and Alex Hanna for reviewing versions of this work.

References

[1] [n.d.]. Playbook - Data Cards Playbook. https://pair-code.github.io/datacardsplaybook/playbook

[2] ACLU. 2020. Federal court rules 'Big Data' discrimination studies do not violate Federal anti-hacking law. https://www.aclu.org/press-releases/federal-court-rules-big-data-discrimination-studies-do-not-violate-federal-anti

[3] Yuki M Asano, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. 2021. PASS: An ImageNet replacement for self-supervised pretraining without humans. https://doi.org/10.48550/ARXIV.2109.13228

[4] John W. Ayers, Theodore L. Caputi, Camille Nebeker, and Mark Dredze. 2018. Don't quote me: reverse identification of research participants in social media studies. *npj Digital Medicine* 1, 1 (aug 2018), 1–2. https://doi.org/10.1038/s41746-018-0036-2

[5] Jo Bates, Yu Wei Lin, and Paula Goodale. 2016. Data journeys: Capturing the socio-material constitution of data objects and flows. *Big Data and Society* 3, 2 (jul 2016). https://doi.org/10.1177/2053951716654502

[6] Anat Ben-David and Adam Amram. 2018. The Internet Archive and the socio-technical construction of historical facts. *Internet Histories* 2, 1-2 (apr 2018), 179–201. https://doi.org/10.1080/24701475.2018.1455412

[7] Misha Benjamin, Paul Gagnon, Negar Rostamzadeh, Christopher Joseph Pal, Yoshua Bengio, and Alex Shee. 2019. Towards Standardization of Data Licenses: The Montreal Data License. *ArXiv* abs/1903.1 (2019).

[8] Cynthia L. Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P. Bigham, Anhong Guo, and Alexandra To. 2021. "It's Complicated": Negotiating Accessibility and (Mis)Representation in Image Descriptions of Race, Gender, and Disability. In *CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM.

[9] Sebastian Benthall and Bruce D. Haynes. 2019. Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. ACM Press, New York, New York, USA, 289–298. https://doi.org/10.1145/3287560.3287575 arXiv:1811.11668

[10] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021*. Institute of Electrical and Electronics Engineers Inc., 1536–1546. https://doi.org/10.1109/WACV48630.2021.00158 arXiv:2006.16923

[11] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S Buch, Dallas Card, Rodrigo Castellon, Niladri S Chatterji, Annie S Chen, Kathleen A Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E Gillespie, Karan Goel, Noah D Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O Khattab, Pang Wei Koh, Mark S Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D Manning, Suvir P Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J F Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W Thomas, Florian Tramèr, Rose E Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. *ArXiv* abs/2108.0 (2021).

[12] Adelaide Bragias, Kelly Hine, and Robert Fleet. 2021. 'Only in our best interest, right?' Public perceptions of police use of facial recognition technology. *Police Practice and Research* 22, 6 (2021), 1637–1654. https://doi.org/10.1080/15614263.2021.1942873

[13] Joy Buolamwini and Timnit Gebru. 2018. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* *. Technical Report. 1–15 pages.

[14] Stevie Chancellor, Eric P.S. Baumer, and Munmun De Choudhury. 2019. Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (nov 2019). https://doi.org/10.1145/3359249

[15] Emil Chiauzzi and Paul Wicks. 2019. Digital trespass: Ethical and terms-of-use violations by researchers accessing data from an online patient community. , e11985 pages. https://doi.org/10.2196/11985

[16] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. (jul 2018). https://doi.org/10.1063/1.3627170 arXiv:1808.00023

[17] Frances Corry, Hamsini Sridharan, Alexandra Sasha Luccioni, Mike Ananny, Jason Schultz, and Kate Crawford. 2021. The Problem of Zombie Datasets: A Framework For Deprecating Datasets. (2021). arXiv:2111.04424 http://arxiv.org/abs/2111.04424

[18] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2020. Underspecification presents challenges for credibility in modern machine learning. arXiv:2011.03395 https://arxiv.org/abs/2011.03395v2

[19] David De Roure. 2014. The future of scholarly communications. *Insights: the UKSG Journal* 27, 3 (nov 2014), 233–238. https://doi.org/10.1629/2048-7754.171

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[21] Melissa Densmore, Casey Fiesler, Cosmin Munteanu, Michael Muller, Janet C Read, Katie Shilton, and Özge Subasl. 2020. Research ethics roundtable. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. ACM, New York, NY, USA, 195–198. https://doi.org/10.1145/3406865.3419015

[22] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. Bringing the People Back In: Contesting Benchmark Machine Learning Datasets. (jul 2020). arXiv:2007.07399 http://arxiv.org/abs/2007.07399

[23] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does Object Recognition Work for Everyone?. In *CVPR Workshops*. 52–59. arXiv:1906.02659 http://arxiv.org/abs/1906.02659

[24] Chris Dulhanty and Alexander Wong. 2019. Auditing ImageNet: Towards a Model-driven Framework for Annotating Demographic Attributes of Large-Scale Image Datasets. (may 2019). arXiv:1905.01347 http://arxiv.org/abs/1905.01347

[25] Chris Dulhanty and Alexander Wong. 2020. Investigating the Impact of Inclusion in Face Recognition Training Data on Individual Face Identification. *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (jan 2020), 244–250. https://doi.org/10.1145/3375627.3375875 arXiv:2001.03071

[26] Casey Fiesler, Nathan Beard, and Brian C. Keegan. 2020. No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. In *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020*, Vol. 14. 187–196. https://ojs.aaai.org/index.php/ICWSM/article/view/7290

[27] Casey Fiesler, Melissa Densmore, Michael Muller, and Cosmin Munteanu. 2021. SIGCHI Research Ethics Committee Town Hall. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. ACM, New York, NY, USA, 232–233. https://doi.org/10.1145/3462204.3483283

[28] Casey Fiesler and Blake Hallinan. 2018. "We are the product": Public reactions to online data sharing and privacy Controversies in the media. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2018-April. https://doi.org/10.1145/3173574.3173627

[29] Casey Fiesler and Nicholas Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media and Society* 4, 1 (jan 2018), 205630511876336. https://doi.org/10.1177/2056305118763366

[30] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. , 86–92 pages. https://doi.org/10.1145/3458723 arXiv:1803.09010

[31] Stuart R. Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 325–336. https://doi.org/10.1145/3351095.3372862 arXiv:1912.08320

[32] Dave Gershgorn. 2017. The data that transformed AI research - and possibly the world. https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/

[33] Sarah Gilbert, Jessica Vitak, and Katie Shilton. 2021. Measuring Americans' Comfort With Research Uses of Their Social Media Data. *Social Media + Society* (2021).

[34] Sarah Gilbert, Jessica Vitak, and Katie Shilton. 2021. Measuring Americans' Comfort With Research Uses of Their Social Media Data. https://doi.org/10.1177/20563051211033824

[35] Antonio A Ginart, Melody Y Guan, Gregory Valiant, and James Zou. 2019. Making AI forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems*, Vol. 32. arXiv:1907.05012

[36] Orlena Gotel, Jane Cleland-Huang, Jane Huffman Hayes, Andrea Zisman, Alexander Egyed, Paul Grünbacher, Alex Dekhtyar, Giuliano Antoniol, Jonathan Maletic, and Patrick Mäder. 2012. Traceability fundamentals. In *Software and Systems Traceability*. Vol. 9781447122. Springer-Verlag London Ltd, 3–22. https://doi.org/10.1007/978-1-4471-2239-5_1

[37] Blake Hallinan, Jed R. Brubaker, and Casey Fiesler. 2020. Unexpected expectations: Public reaction to the Facebook emotional contagion study. *New Media and Society* 22, 6 (sep 2020), 1076–1094. https://doi.org/10.1177/1461444819876944

[38] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender Recognition or Gender Reductionism? The Social Implications of Automatic Gender Recognition Systems. In *2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*.

[39] Jeffrey T. Hancock. 2020. The Ethics of Digital Research. In *The Oxford Handbook of Networked Communication*. Oxford University Press, 512–519. https://doi.org/10.1093/oxfordhb/9780190460518.013.25

[40] Margot Hanley, Apoorv Khandelwal, Hadar Averbuch-Elor, Noah Snavely, and Helen Nissenbaum. 2020. An Ethical Highlighter for People-Centric Dataset Creation. (nov 2020). arXiv:2011.13583 http://arxiv.org/abs/2011.13583

[41] Adam Harvey and Jules LaPlace. 2021. Exposing.ai. https://exposing.ai/

[42] Christine Hauser. 2018. $6.4 Million Judgment in Revenge Porn Case Is Among Largest Ever. https://www.nytimes.com/2018/04/11/us/revenge-porn-california.html

[43] Melissa Heikkilä. 2021. European Parliament calls for a ban on facial recognition – POLITICO. *POLITICO* (2021). https://www.politico.eu/article/european-parliament-ban-facial-recognition-brussels/

[44] R. A. Hibbin, G. Samuel, and G. E. Derrick. 2018. From "a Fair Game" to "a Form of Covert Research": Research Ethics Committee Members' Differing Notions of Consent and Potential Risk to Participants Within Social Media Research. *Journal of Empirical Research on Human Research Ethics* 13, 2 (apr 2018), 149–159. https://doi.org/10.1177/1556264617751510

[45] Kashmir Hill and Aaron Krolik. 2019. How Photos of Your Kids Are Powering Surveillance Technology. https://www.nytimes.com/interactive/2019/10/11/technology/flickr-facial-recognition.html

[46] Anna Lauren Hoffmann. 2020. Terms of inclusion: Data, discourse, violence. *New Media & Society* 23, 12 (sep 2020), 3539–3556. https://doi.org/10.1177/1461444820958725

[47] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. (may 2018). arXiv:1805.03677 http://arxiv.org/abs/1805.03677

[48] James Vincent. 2017. Transgender YouTubers had their videos grabbed to train facial recognition software.

[49] Nivash Jeevanandam. 2021. The Conundrum Of User Data Deletion From ML Models. *Analytics India Mag* (2021). https://analyticsindiamag.com/data-deletion-from-ml-models/

[50] Sarah Jeong. 2014. Reddit As A Government. https://www.forbes.com/sites/sarahjeong/2014/09/08/reddit-as-a-government/?sh=10ebeb57856d

[51] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 306–316. https://doi.org/10.1145/3351095.3372829 arXiv:1912.10389

[52] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (nov 2018), 1–22. https://doi.org/10.1145/3274357

[53] Zaid Khan and Yun Fu. 2021. One Label, One Billion Faces: Usage and Consistency of Racial Categories in Computer Vision. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (feb 2021), 587–597. https://doi.org/10.1145/3442188.3445920 arXiv:2102.02320v1

[54] Harris Kornstein. 2019. Under her eye: Digital drag as obfuscation and countersurveillance. *Surveillance and Society* 17, 5 (dec 2019), 681–698. https://doi.org/10.24908/ss.v17i5.12957

[55] Amparo Lasén and Edgar Gómez-Cruz. 2009. Digital Photography and Picture Sharing: Redefining the Public/Private Divide. *Knowledge, Technology & Policy* 22, 3 (2009), 205–215. https://doi.org/10.1007/s12130-009-9086-8

[56] Mark A. Lemley and Bryan Casey. 2021. Fair Learning. *Texas Law Review* 99, 4 (jan 2021), 744–785. https://doi.org/10.2139/ssrn.3528447

[57] Amanda Levendowski. [n.d.]. Resisting Face Surveillance with Copyright Law. https://papers.ssrn.com/abstract=3924647

[58] Yangzi Li. 2021. Does black-box machine learning shift the US fair use doctrine? *Journal of Intellectual Property Law & Practice* 16, 11 (dec 2021), 1175–1185. https://doi.org/10.1093/jiplp/jpab118

[59] Louise Matsakis. 2020. Scraping the Web Is a Powerful Tool. Clearview AI Abused It. *WIRED* (2020). https://www.wired.com/story/clearview-ai-scraping-web/

[60] Matthew S. Mayernik, Tim DiLauro, Ruth Duerr, Elliot Metsger, Anne E. Thessen, and G. Sayeed Choudhury. 2013. Data conservancy provenance, context, and lineage services: Key components for data preservation and curation. *Data Science Journal* 12, 0 (nov 2013), 158–171. https://doi.org/10.2481/dsj.12-039

[61] Matthew B.A. McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Luca Foschini, and Marzyeh Ghassemi. 2021. Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine* 13, 586 (mar 2021). https://doi.org/10.1126/SCITRANSLMED.ABB1655/SUPPL_FILE/ABB1655_SM.PDF

[62] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. , 23 pages. https://doi.org/10.1145/3359174

[63] Ruth McNally, Adrian Mackenzie, Allison Hui, and Jennifer Tomomitsu. 2012. Understanding the 'Intensive' in 'Data Intensive Research': Data Flows in Next Generation Sequencing and Environmental Networked Sensors. *International Journal of Digital Curation* 7, 1 (mar 2012), 81–94. https://doi.org/10.2218/ijdc.v7i1.216

[64] Ryan Merkley. 2019. Use and Fair Use: Statement on shared images in facial recognition AI. https://creativecommons.org/2019/03/13/statement-on-shared-images-in-facial-recognition-ai/

[65] Jacob Metcalf and Kate Crawford. 2016. Where are human subjects in Big Data research? The emerging ethics divide. *Big Data & Society* 3, 1 (2016), 2053951716650211. https://doi.org/10.1177/2053951716650211

[66] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power?. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 6. ACM PUB27 New York, NY, USA. https://doi.org/10.1145/3492853

[67] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (oct 2020), 25. https://doi.org/10.1145/3415186

[68] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting computer vision datasets: An invitation to reflexive data practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. ACM, New York, NY, USA, 161–172. https://doi.org/10.1145/3442188.3445880

[69] Madhumita Murgia. 2019. Microsoft quietly deletes largest public face recognition data set. *Financial Times* (2019). https://www.ft.com/content/7d3e0d6a-87a0-11e9-a028-86cea8523dc2

[70] Madhumita Murgia. 2019. Who's using your face? The ugly truth about facial recognition. *Financial Times* (2019). https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e

[71] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research.* Technical Report. .S. Department of Health and Human Services.

[72] Melika Abbasian Nik, Melika Abbasian Nik, Mohammad Mahdi Dehshibi, and Dr. Azam Bastanfard. 2007. Iranian Face Database and Evaluation with a New Detection Algorithm. (2007). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.418.771

[73] Michael O'Flanagan. 2018. Photography and the Law: Rights and Restrictions. *Photography and the Law* (oct 2018). https://doi.org/10.4324/9780429468391

[74] Olivia Solon. 2019. Facial recognition's 'dirty little secret': Millions of online photos scraped without consent. https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921

[75] Petter Olsen and Melania Borit. 2013. How to define traceability. , 142–150 pages. https://doi.org/10.1016/j.tifs.2012.10.003

[76] Nicolas Papernot, Ian Goodfellow, Martín Abadi, Kunal Talwar, and Úlfar Erlingsson. 2017. Semi-supervised knowledge transfer for deep learning from private training data. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR. arXiv:1610.05755 https://arxiv.org/abs/1610.05755v4

[77] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* (jan 2019), 39–48. https://doi.org/10.1145/3287560.3287567

[78] Jessica Pater, Casey Fiesler, and Michael Zimmer. 2022. No Humans Here: Ethical Speculation on Public Data, Unintended Consequences, and the Limits of Institutional Review. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 6. ACM PUB27 New York, NY, USA. https://doi.org/10.1145/3492857

[79] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2020. Data and its (dis)contents: A survey of dataset development and use in machine learning research. (dec 2020). https://doi.org/10.1016/j.patter.2021.100336 arXiv:2012.05345

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

[81] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers. (aug 2021). arXiv:2108.02922 https://arxiv.org/abs/2108.02922v1http://arxiv.org/abs/2108.02922

[82] Nicholas Proferes. 2017. Information Flow Solipsism in an Exploratory Study of Beliefs About Twitter. *Social Media and Society* 3, 1 (mar 2017). https://doi.org/10.1177/2056305117698493

[83] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media and Society* 7, 2 (may 2021). https://doi.org/10.1177/20563051211019004

[84] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving Face: Investigating the ethical concerns of facial recognition auditing. In *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 145–151. https://doi.org/10.1145/3375627.3375820 arXiv:2001.00964

[85] John Rooksby. 2014. Can plans and situated actions be replicated?. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. Association for Computing Machinery, 603–614. https://doi.org/10.1145/2531602.2531627

[86] Jake Satisky. 2019. A Duke study recorded thousands of students' faces. Now they're being used all over the world. https://www.dukechronicle.com/article/2019/06/duke-university-facial-recognition-data-set-study-surveillance-video-students-china-uyghur

[87] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (oct 2021). https://doi.org/10.1145/3476058

[88] Morgan Klaus Scheuerman, Madeleine Pape, and Alex Hanna. 2021. Auto-essentialization: Gender in automated facial analysis as extended colonial project:. *https://doi.org/10.1177/20539517211053712* 8, 2 (dec 2021), 205395172110537. https://doi.org/10.1177/20539517211053712

[89] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services. 144 (2019), 33. https://doi.org/10.1145/3359246

[90] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1 (2020).

[91] Sovantharith Seng, Mahdi Nasrullah Al-Ameen, and Matthew Wright. 2021. A first look into users' perceptions of facial recognition in the physical world. *Computers and Security* 105 (jun 2021), 102227. https://doi.org/10.1016/j.cose.2021.102227

[92] Katie Shilton, Emanuel Moss, Sarah A. Gilbert, Matthew J. Bietz, Casey Fiesler, Jacob Metcalf, Jessica Vitak, and Michael Zimmer. 2021. Excavating awareness and power in data science: A manifesto for trustworthy pervasive data research. *Big Data and Society* 8, 2 (sep 2021). https://doi.org/10.1177/20539517211040759

[93] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *Proceedings - IEEE Symposium on Security and Privacy*. Institute of Electrical and Electronics Engineers Inc., 3–18. https://doi.org/10.1109/SP.2017.41 arXiv:1610.05820

[94] G. C. Smith, J. D. Tatum, K. E. Belk, J. A. Scanga, T. Grandin, and J. N. Sofos. 2005. Traceability from a US perspective. In *Meat Science*, Vol. 71. Elsevier, 174–193. https://doi.org/10.1016/j.meatsci.2005.04.002

[95] Julie Carr Smyth. 2021. States push back against use of facial recognition by police. https://abcnews.go.com/Politics/wireStory/states-push-back-facial-recognition-police-77510175https://apnews.com/article/race-and-ethnicity-health-coronavirus-pandemic-business-technology-e4266250f7e2d691d4d664735c2c6bc0

[96] Benjamin Sobel. 2020. A Taxonomy of Training Data: Disentangling the Mismatched Rights, Remedies, and Rationales for Restricting Machine Learning. *SSRN Electronic Journal* (aug 2020). https://doi.org/10.2139/ssrn.3677548

[97] Nikki Stevens and Os Keyes. 2021. Seeing infrastructure: race, facial recognition and the politics of data. *Cultural Studies* (mar 2021), 1–21. https://doi.org/10.1080/09502386.2021.1895252

[98] Harini Suresh and John V Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. (2019). arXiv:1901.10002 www.aaai.orghttp://arxiv.org/abs/1901.10002

[99] Hamed Taherdoost. 2018. Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research. *SSRN Electronic Journal* (apr 2018). https://doi.org/10.2139/ssrn.3205035

[100] Carlo Tomasi. 2019. Letter: Video analysis research at Duke. https://www.dukechronicle.com/article/2019/06/duke-university-video-analysis-research-at-duke-carlo-tomasi

[101] Antonio Torralba, Rob Fergus, and Bill Freeman. 2020. 80 Million Tiny Images. https://groups.csail.mit.edu/vision/TinyImages/

[102] Jorge Valero. 2020. Vestager: Facial recognition tech breaches EU data protection rules. https://www.euractiv.com/section/digital/news/vestager-facial-recognition-tech-breaches-eu-data-protection-rules/

[103] Emiel van Miltenburg. 2016. Stereotyping and Bias in the Flickr30K Dataset. (may 2016), 24. arXiv:1605.06083 https://arxiv.org/abs/1605.06083v1

[104] Jessica Vitak, Nicholas Proferes, Katie Shilton, and Zahra Ashktorab. 2017. Ethics Regulation in Social Computing Research: Examining the Role of Institutional Review Boards. *Journal of Empirical Research on Human Research Ethics* 12, 5 (dec 2017), 372–382. https://doi.org/10.1177/1556264617725200

[105] Jessica Vitak, Katie Shilton, and Zahra Ashktorab. 2016. Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, Vol. 27. Association for Computing Machinery, 941–953. https://doi.org/10.1145/2818048.2820078

[106] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

[107] Marilyn Domas White and Emily E Marsh. 2006. Content analysis: A flexible methodology. *Library Trends* 55, 1 (2006), 22–45. https://doi.org/10.1353/lib.2006.0053

[108] Zack Whittaker. 2021. Supreme Court revives LinkedIn case to protect user data from web scrapers. *TechCrunch* (2021). https://techcrunch.com/2021/06/14/supreme-court-revives-linkedin-bid-to-protect-user-data-from-web-scrapers/

[109] Max L Wilson. [n.d.]. RepliCHI-The Workshop II. *CHI '14 Extended Abstracts on Human Factors in Computing Systems* ([n. d.]). https://doi.org/10.1145/2559206

[110] Matthew W. Wilson. 2011. Data matter(s): Legitimacy, coding, and qualifications-of-life. *Environment and Planning D: Society and Space* 29, 5 (jan 2011), 857–872. https://doi.org/10.1068/d7910

[111] Stefan Wojcik, Emma Remy, and Chris Baronavski. 2019. How does a computer 'see' gender? Pew Research Center. https://www.pewresearch.org/interactives/how-does-a-computer-see-gender/

[112] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. 2979–2989. https://doi.org/10.18653/v1/d17-1323 arXiv:1707.09457

[113] Michael Zimmer. 2010. "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology* 12, 4 (dec 2010), 313–325. https://doi.org/10.1007/s10676-010-9227-5

[114] Michael Zimmer and Sarah Logan. 2021. Privacy concerns with using public data for suicide risk prediction algorithms: a public opinion survey of contextual appropriateness. *Journal of Information, Communication and Ethics in Society* ahead-of-p, ahead-of-print (dec 2021). https://doi.org/10.1108/jices-08-2021-0086

[115] Jonathan Zong and J. Nathan Matias. 2022. Bartleby: Procedural and Substantive Ethics in the Design of Research Ethics Systems. *Social Media and Society* 8, 1 (feb 2022). https://doi.org/10.1177/20563051221077021

Received January 2022; revised July 2022; accepted November 2022