COMMENTS OF THE KNOWING MACHINES RESEARCH PROJECT
Melodi Dincer,[1] Jason Schultz,[2] & Christo Buschek[3]
to the
United States Copyright Office
Notice of Inquiry on Artificial Intelligence and Copyright
88 Fed Reg. 59,942
October 30, 2023

.

The Knowing Machines Research Project (Knowing Machines) submits these comments in response to the United States Copyright Office's (USCO) Notice of Inquiry and Request for Comments on Artificial Intelligence and Copyright (NOI) published August 30, 2023.[4] This NOI builds on USCO's comprehensive Artificial Intelligence (AI) Initiative and seeks responses relating to "the use of copyrighted works to train AI models," among other issues.[5] The purpose of the NOI is to provide USCO with information and perspectives on the myriad law and policy issues raised by advanced AI systems, especially recent generative AI systems (GenAI).[6]

We appreciate the opportunity to contribute to USCO's inquiry. Knowing Machines is an interdisciplinary research project tracing the histories, practices, and politics of how automated systems are trained to interpret the world from vast, nebulous datasets. Our research targets the assumptions underlying emerging machine-learning technologies with the hope that greater transparency will encourage meaningful interventions. [7] We are a team of lawyers, researchers, science and technology studies (STS) professors, artists, and data scientists who have published extensively on the construction of training datasets, their central role in AI development, and the social values embedded within them.[8]

[1] Legal Research Fellow, Knowing Machines Research Project; Supervising Attorney, NYU's Technology Law & Policy Clinic, and Fellow, Engelberg Center on Innovation Law & Policy, NYU School of Law.
[2] Co-Principal Investigator, Knowing Machines Research Project; Professor of Clinical Law, Director of NYU's Technology Law & Policy Clinic, and Co-Director of the Engelberg Center on Innovation Law & Policy, NYU School of Law.
[3] Data Investigator & Knowing Machines Fellow, Engelberg Center on Innovation Law & Policy, NYU School of Law.
[4] 88 Fed Reg. 59942, *available at* https://www.federalregister.gov/documents/2023/08/30/2023-18624/artificial-intelligence-and-copyright.
[5] *Id.* at 59,945.
[6] *See id.*
[7] For more information, *see* https://knowingmachines.org.
[8] See, e.g., KATE CRAWFORD, ATLAS OF AI: POWER, POLITICS, AND THE PLANETARY COSTS OF ARTIFICIAL INTELLIGENCE 89–121 (2021) [hereinafter, ATLAS OF AI]; Sarah Ciston, Knowing Machines Res. Project, *A Critical Field Guide for Working with Machine Learning Datasets* (2023), https://knowingmachines.org/critical-field-guide [hereinafter, *Datasets Critical Field Guide*]; Kate Crawford, Knowing Machines Res. Project, *9 Ways to See a Dataset: What's at*

In response to NOI Questions 6.1.–3.,[9] 7.4.,[10] 8.3,[11] 15.1.,[12] and 15.2.,[13] Knowing Machines recommends that USCO rely on research-based, empirical findings to inform its regulatory agenda and any recommendations to Congress on this topic. Specifically, USCO should advocate for support and funding to develop data investigatory tools to inform its assessment of training datasets for GenAI and their potential impact on the copyright system as a whole. In this response, we briefly discuss Knowing Machines' experience building a data investigatory tool for training datasets as one example of such tools and to demonstrate some of the ways in which data investigations may provide empirical findings to support evidence-based policymaking. For example, using our tool, we have shown that determining the copyright status of a given image among the millions or billions of images included can be challenging and contextual for dataset users.[14] We have also shown that licensing metadata is often missing, incomplete, or misapplied.[15] Each of these findings suggests that blanket assumptions as to the ease of copyright licensing for large datasets may be inaccurate and that further investigation of datasets is needed to fully account for their copyright implications.

Stake in Examining Datasets? (2023), https://knowingmachines.org/9-ways-to-see/9-ways-to-see-a-dataset; Christo Buschek, Knowing Machines Res. Project, *9 Ways to See a Dataset: Investigating Datasets* (2023), https://knowingmachines.org/9-ways-to-see/investigating-datasets; Will Orr and Kate Crawford, *The Social Construction of Datasets: On the Practices, Processes and Challenges of Dataset Creation for Machine Learning* (forthcoming).

[9] 6.1. "How or where do developers of AI models acquire the materials or datasets that their models are trained on? To what extent is training material first collected by third-party entities (such as academic researchers or private companies)?"
6.2. "To what extent are copyrighted works licensed from copyright owners for use as training materials? To your knowledge, what licensing models are currently being offered and used?"
6.3. "To what extent is non-copyrighted material … used for AI training? Alternatively, to what extent is training material created or commissioned by developers of AI models?"

[10] "Absent access to the underlying dataset, is it possible to identify whether an AI model was trained on a particular piece of training material?"

[11] "The use of copyrighted materials in a training dataset or to train generative AI models may be done for noncommercial or research purposes. How should the fair use analysis apply if AI models or datasets are later adapted for use of a commercial nature? Does it make a difference if funding for these noncommercial or research uses is provided by for-profit developers of AI systems?"

[12] "What level of specificity [for training material records] should be required?"

[13] "To whom should disclosure be made?"

[14] See Jason Schultz, Knowing Machines Res. Project, *9 Ways to See a Dataset: What Can LAION Teach Us About Copyright Law* (2023), https://knowingmachines.org/9-ways-to-see/LAION-copyright.

[15] Results on file with authors and available upon request. See also Shayne Lonpre et al., *The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI* 7–8 (2023), https://arxiv.org/pdf/2310.16787.pdf (finding a "crisis in misattribution" stemming in part from frequent miscategorization of licenses, license omissions, and high rates of attribution error on popular dataset hosting websites).

A portion of the NOI also concerns transparency and recordkeeping for dataset creators and users. Knowing Machines recommends that USCO further study and advocate for funding to support best practices for training dataset creation and maintenance. This includes dataset transparency and recording requirements to facilitate keeping track of the vast, internet-scale datasets that power GenAI, as well as guidelines for preventing the use of depreciated datasets to train these systems.

We understand the difficulties of gaining a deep understanding of these training datasets firsthand. We need new investigatory methods to uncover the hidden problems inscribed in machine learning processes. Because dataset creators and AI developers lack standardized *ex ante* dataset transparency and recordkeeping requirements, we now rely almost exclusively on *ex post* data investigative tools for research, often unable to identify all the necessary information we need to understand datasets, especially in a copyright context. Although it is challenging, we urge the USCO to support evidence-based research concerning the nature of training datasets and their role in GenAI outputs, minimizing the influence of conjecture in the policy-making process.

I. AI Researchers Are Studying the Contents and Centrality of Vast, Internet-Sized Training Datasets, Despite Challenges Posed by Unprecedented Scale.[16]

The NOI asks several important questions regarding training AI models.[17] To adequately answer these questions, however, it helps to have a basic understanding of training datasets and the values that motivate their creators. This is Knowing Machines' core mission. Our research unpacks what training datasets contain, where they come from, and how they are used to train AI systems, including recent GenAI. We combine STS research methodologies with technical data investigations to render these vast datasets more legible. Our expertise comes from years of research into the contents of these datasets, their influence on dominant AI systems, and the motivations of their creators.

A dataset can be any set of collected, curated, and interrelated data. The term "dataset" often refers to large collections of data used in computation, especially in machine learning.[18]

[16] This Section responds primarily to Questions 6.1.–3.

[17] See NOI Questions 6.–14.

[18] *Datasets Critical Field Guide*, *supra* note 8.

But a dataset could have numerous versions all up for use in training AI models, and this is especially so for earlier datasets that have become influential through decades of developers relying on them to train their models. For example, ImageNet is one of the most influential datasets that, since its release in 2009, has been used to "train and evaluate nearly every AI model in the object recognition task."[19] But despite its popularity in the field, "no single 'ImageNet' dataset exists."[20] There are nine versions of the dataset, some designed for specific tasks while others resulted from filtering and moderation meant to improve the perceived fairness of the original dataset.[21] When developers rely on influential datasets like ImageNet that have been used to train numerous AI systems, they are not always specific about which version(s) they used to train their models and why. This is in part because there are no consistent legal requirements for dataset recordkeeping and transparency.

Like the data they contain, datasets are often constructed by people and will always reflect the contexts of their creation, including technical constraints of the tools and methods used to construct the dataset, the intentions and perceptions of their creators, and the ability to afford data training, storage, and transmission equipment.[22] A dataset can contain specific types of data or a mixture of different types; for example, computer vision datasets include thousands of image or video files, natural language processing datasets contain millions of bytes of text, and increasingly more datasets contain multimodal data.[23] Training data refers specifically to a portion of the full dataset used to create a specific machine learning model. In supervised machine learning processes, where algorithms are trained to produce desired outcomes based on labeled data, training data includes results that are like the desired results the model will generate.[24] Similarly with GenAI models, training involves presenting data in an iterative process and frequently adjusting its parameters until generated outputs match the

[19] Sasha Luccioni, Knowing Machines Res. Project, *9 Ways to See a Dataset: Investigating Imagenet* (2023), https://knowingmachines.org/9-ways-to-see/ImageNet.
[20] *Id.*
[21] *Id. See also* Kate Crawford and Trevor Paglen, *Excavating AI: The Politics of Training Sets for Machine Learning* (Sept. 19, 2019), https://excavating.ai/ (describing problematic image labels in ImageNet).
[22] *See Datasets Critical Field Guide*, *supra* note 8. *See also* Bernard Koch et al., *Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research* (2021), https://openreview.net/forum?id=zNQBIBKJRkd (finding that the majority of widely-used benchmark datasets come from just a handful of elite institutions, including Stanford, Microsoft, Princeton, Google, CUHK, AT&T, NYU, Georgia Tech, Berkeley, and Facebook).
[23] *Datasets Critical Field Guide*, *supra* note 8.
[24] *Id.*

desired outputs. Training datasets are foundational to contemporary machine learning systems, as they shape "the epistemic boundaries governing how AI systems operate."[25]

As artist and technology researcher Mimi Onuoha states, "[d]atasets are the results of their means of collection."[26] Creating a dataset involves two main steps: collecting data to create a set of candidates (i.e., data that *could* be included in the dataset), and curating the set of candidates to produce the final dataset. Creating candidates is an important-yet-understudied element in determining the outer edges of a training dataset. From exploring open-access datasets, we know that AI model developers tend to mostly defer to third-party entities, like the nonprofit Common Crawl, to provide a list of candidates. To build out its massive repository of data, Common Crawl automatically scraped over 250 billion web pages spanning 16 years, adding 3–5 billion new pages each month.[27]

Indiscriminate scraping of the internet—without permission or consent from data subjects—is a common technique among dataset creators.[28] The training datasets that constitute "ground truth" for AI developers are less reflections of reality and more "jumble[s] of [data] scraped from whatever various online resources were available."[29] To our knowledge, the technical methods used to scrape data to build such vast, internet-sized datasets do not include mechanisms to determine the copyright status of a piece of data before including it as a candidate. Currently, most dataset creators appear to operate under the default assumption that everything on the public internet is fair game unless told otherwise. How copyright owners or data subjects "tell" dataset creators not to include or to train on their data is also an area of law that needs clarification. Until more recently, many in the machine learning community operated under a "train first, ask questions later" approach advocated by the creators of influential datasets in the field.[30] As more rules and responsibilities around datasets emerge,

[25] Crawford and Paglen, *supra* note 21.
[26] Mimi Onuoha, *The Point of Collection*, MEDIUM (Feb. 10, 2016), https://medium.com/datasociety-points/the-point-of-collection-8ee44ad7c2fa.
[27] *See* Common Crawl, https://commoncrawl.org/ (last visited Oct. 30, 2023).
[28] NOI Question 6.1. *See also* ATLAS OF AI, *supra* note 8, at 96 ("Yet now it's common practice for the first steps of creating a computer vision system to scrape thousands—or even millions—of images from the internet, create and order them into a series of classifications, and use this as a foundation for how the system will perceive observable reality"); Luccioni, *supra* note 19 ("The process for creating ImageNet—large-scale web scraping followed by manual validation by Mechanical Turk workers—has become the de facto standard for creating AI datasets.").
[29] ATLAS OF AI, *supra* note 8, at 96.
[30] Luccioni, *supra* note 19.

law must play a stronger role in ensuring that data-based rights and responsibilities are respected.

The next step in creating training datasets is curating the candidates to determine whether a piece of data will be included or not. In open-access datasets, this is mostly an automated process modeled after a statistical process. In the case of the LAION-5B dataset, which was used to train popular text-to-image GenAI like Stable Diffusion,[31] an automated process examines each candidate image to determine if it is "similar" enough to its caption text to be included, based on an algorithmically-set threshold of similarity. By contrast, there is no published process for algorithmically determining the copyright status of a given image. Thus, while image similarity has been both automated and benchmarked to scale across billions of images, image copyright status has not. Even watermark detection processes, while helpful in flagging some copyright-protected images, have not been developed or benchmarked sufficiently to serve as a reliable proxy for copyright status itself. Whether copyright status processes can be automated and included as part of dataset creation is a key question for policymakers to consider and a key area for researchers to continue investigating.[32]

Dataset creators wield immense downstream influence in this initial step of creating and curating training datasets by defining the parameters of what will be exposed to AI models in training. For example, the nonprofit that created the Common Crawl dataset determined the parameters of the data included, setting up downstream developers to reinscribe those choices when they use Common Crawl's dataset to train their models without significant filtering. 60% of the training data used to train OpenAI's GPT-3 language model was "derived from raw Common Crawl with only quality-based filtering."[33] Of course, we only know this about open-access datasets because they are publicly available. We do not know how creators of proprietary datasets select a set of candidates or curate which data to include. This is a major reason to support calls for dataset transparency, especially for researchers and investigators.

[31] *See* Andy Biao, *Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator*, WAXY.ORG (Aug. 30, 2022), https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/.
[32] NOI Question 6.2.
[33] Tom Brown et al., *Language Models Are Few-Shot Learners* 14 (2020), https://arxiv.org/pdf/2005.14165.pdf.

1. Dataset Provenance Is an Open Question that Investigatory Tools Can Help Answer.

The NOI also asks questions about the provenance of training data,[34] but these questions are often difficult to answer as AI developers generally choose to withhold information about which datasets they use to train their proprietary models.[35] Their secrecy makes it difficult to know how and where they acquire training datasets,[36] as well as the specific types of data they access to train their models.[37] While large technology companies like Google, Amazon, and Microsoft are perceived to have mass amounts of internal data to use to develop their AI products, dataset creators interviewed by Knowing Machines noted that these corporate players often adopt open-access datasets that are widely available, like Common Crawl, LAION, and the Yahoo Flickr Creative Commons 100m (YFCC100M) datasets.[38] Training data thus comes from a mixture of corporate datasets and internet-scale, publicly-scraped datasets. Some consider this to be "data laundering," where private corporations attempt to avoid copyright and data protection issues by training on nonprofit and/or academic datasets, which presumably are more likely exempt from liability.[39] Thus it is important to increase transparency in the supply chain of AI systems and to understand the exact role that each actor within the chain plays, what their commercial interest is (if any), and what their specific organizational policies are regarding dataset use for training. Such facts are essential to any copyright analysis, including whether fair use should apply to a given actor's conduct.[40]

In this mishmash of datasets, issues of provenance become murkier and murkier, often confusing AI researchers as to what legal constraints, if any, apply.[41] For example, our data investigations have shown that the only licensing information the creators of the LAION-5B

[34] NOI Questions 6.1.–3. (where AI developers acquire training data, to what extent copyrighted works are licensed for use as training data, and to what extent non-copyrighted materials are used to train AI, respectively).
[35] *See* Tiernan Ray, *With GPT-4, OpenAI Opts for Secrecy Versus Disclosure*, ZDNET (Mar. 16, 2023), https://www.zdnet.com/article/with-gpt-4-openai-opts-for-secrecy-versus-disclosure/.
[36] NOI Question 6.1.
[37] NOI Questions 6.2.–3.
[38] Orr and Crawford, *supra* note 8. In their article, the authors use the term "dataset creator" to refer to the individuals "who had substantive first-hand experience producing datasets." *Id.*
[39] Andy Baio, *AI Data Laundering: How Academic and Nonprofit Researchers Shield Tech Companies from Accountability*, WAXY.ORG (Sept. 30, 2022), https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield-tech-companies-from-accountability/.
[40] NOI Question 8.3
[41] NOI Question 6.3.

dataset record in their "license field" appears to be whether a piece of data has a Creative Commons (CC) license and, if so, which type.[42] As a result, while LAION-5B includes a "license" field for each of its 5.85 billion data points,[43] this field lacks any licensing information for 99.7% of the data. And even for the 0.3% of data which are supposedly CC licensed, the results are not verified.[44] One recent investigation traced over 1,800 text datasets used in supervised natural language processing systems by developing tools and standards to trace their provenance.[45] It found rampant miscategorizations in self-reported license information by popular open-access dataset repositories, with license omission rates of over 72% and error rates above 50%.[46]

It is challenging to answer the NOI's training data questions because of the immense scale of datasets today. Dataset creators have been motivated by a prodigious pursuit for scale that involves metrics beyond most measures used in copyright policy. In the past two decades, training datasets have metamorphosed into "massive, indiscriminate dragnets of the internet with little to no curation at all."[47] In 2003, the early Caltech 101 dataset had under 10,000 images; by 2010, the ImageNet database approached 14 million images. That is a 140,000% increase in scale in seven years. Twelve years later in 2022, LAION-5B had over 5 billion web-scraped image-text pairs, and by April 2023, CommonPool launched with 12.8 billion web-scraped image-text pairs.[48] In just under a year, we have witnessed a 256% increase in scale with a starting point already in the billions. The difference between datasets built just a decade ago and now is notable. It would take a person four and a half years to look at each image in ImageNet for ten seconds; it would take the same person 4,000 years to glance at each data point in CommonPool—about fifty lifetimes.[49] And this is just at the level of training data. The

[42] Results on file with authors and available upon request.
[43] *See* Romain Beaumont, *LAION-5B: A New Era of Open Large-Scale Multi-Modal Datasets*, LAION.AI BLOG (Mar. 31, 202), https://laion.ai/blog/laion-5b/.
[44] In further response to Question 6.3., we posit that, while AI developers can create or commission synthetic data to train their models theoretically, this has little influence on the development of AI practically and even less so on large language models like ChatGPT. It is uncertain whether it is possible to build datasets of synthetic training data that can match the sheer volume of data used to train today's AI models.
[45] *See* Lonpre et al., *supra* note 15.
[46] *Id.* at 2.
[47] Crawford, *supra* note 8.
[48] *Id.*
[49] *Id.*

behavior of AI models is increasingly opaque and unexplainable, and sometimes the only piece of the equation AI researchers can access is the data that informs model outputs.[50]

In this complex data landscape, AI researchers are at an impasse. On the one hand, training datasets are more important than ever for understanding the boundaries of what AI systems can do and how. On the other hand, the vast datasets that dominate AI model training today are uniquely challenging to understand. While close engagements with training datasets have been rare in the AI field so far, it is becoming increasingly necessary to understand them to be able to regulate these technologies.[51]

II. To Regulate Generative AI Systems Effectively, Researchers, Regulators, and Policymakers Must Be Able to Investigate the Training Datasets That Power Them.[52]

The internet-scale datasets that power GenAI require new data investigation tools and research methods to inform regulatory choices. The unfathomable scale of more recent training datasets should motivate researchers, regulators, and policymakers alike to develop novel investigative methods that make these invisible building blocks of GenAI more visible. Such tools must offer new ways of seeing data by enabling investigators to focus on specific slices of training data, search and probe the data and metadata directly, and draw broader conclusions through statistical analyses. These investigations should go beyond itemizing each piece of data in a dataset, uncovering whose specific worldviews training datasets encode, when that encoding occurs during training, and to what ends.[53] To develop targeted and effective policy recommendations, researchers, regulators, and policymakers must be able to explore the

[50] In response to NOI Question 7.4., given that AI model behavior is largely unexplainable, looking at the underlying training dataset(s) is often one of the few ways to develop an understanding of the chain of interactions that lead to a particular output. We emphasize the importance of data investigations because they allow us to understand the implications of an AI model even when the model itself is unexplainable, or it is impossible to identify whether an AI model was trained on a particular piece of training material.

[51] *See* Crawford, *supra* note 8.

[52] This Section responds to NOI Question 34.

[53] *See* Crawford, *supra* note 8.

hidden folds of these massive datasets and ground a regulatory agenda in empirical, factual findings.

For Knowing Machines, training datasets have become the focus for developing data investigation tools. An investigation is a critical scrutiny of the investigative subject meant to uncover something hidden and tell a particular story about it. Investigatory tools are designed to enable this scrutiny and introspection. Investigating datasets specifically is a method as well as a critical practice. It is a powerful "antidote to suspending one's disbelief," the commonplace reaction to the recent deluge of GenAI.[54] When investigating datasets, researchers engage them from an outsider's perspective, taking an adversarial standpoint to unearth both the explicit and implicit qualities of datasets. Initially, all data is just information; how we engage with this information shapes our results. But while designing investigatory tools can be the first step in understanding datasets, focusing too heavily on the tools used can distract researchers from the real goal—gaining insights and facts about what these datasets contain and where those data come from. The tools must follow the research questions, not the other way around. When developing an investigatory tool for training datasets, the research question is relatively straightforward: How can we explore their contents and make them comparable to one another? For copyright researchers, regulators, and policymakers, the research questions might include several NOI questions concerning the copyright status of training data within dominant training datasets in the field. Datasets are built for machine consumption, so investigatory tools must render them human-readable.[55]

1. Knowing Machines' See:Set Tool Provides Useful Insights into Creating Investigatory Tools for Training Datasets.

Knowing Machines developed an investigatory tool to break up the patchwork of training datasets used in AI development into human-readable galleries. Our See:Set tool renders these datasets more knowable by giving users the ability to browse and search the data within. It engages with ten of the most popular datasets and 80 million data points within them. It allows investigators to inspect these datasets, and to search and compare data and metadata across multiple datasets. Today, datasets are multimodal, so See:Set can handle a wide variety

[54] Buschek, *supra* note 8.
[55] *Id.*

of media, including text, images, or audio data. We designed See:Set primarily for nontechnical individuals to use, so it displays a simple graphical user interface and a search bar which users can use to query one or multiple selected datasets. Because a plethora of tools are needed at different stages of an investigation, See:Set was designed to maximize compatibility with other tools, for instance through data exporting. It also has a workflow that uses spreadsheets as a second interface to further analyze data insights. One can easily export data from one or more datasets into a spreadsheet, then organize and analyze those dataset contents like any other type of numerical information.

When developing See:Set, our main research question was how to navigate datasets that contain millions or billions of individual data points to see those data points both independently and holistically in the context of their greater datasets. This data investigation tool needed to serve the broader investigatory mission of Knowing Machines to trace how AI systems are trained to interpret reality. From the beginning, a core design choice was to use a search function to parse through dataset contents. See:Set was designed to foster collaborative investigations among Knowing Machines' members, so it includes two features to maximize collaborations. The first feature enables an investigator to export data discovered within See:Set using an open data format (CSV in this case), which allows them to piece together insights from various tools, including See:Set, without requiring See:Set to implement a long list of features that other tools are better suited for. The second feature is having each piece of data individually addressable with a direct link. This link feature is crucial because it allows investigators to "name" a piece of data, and this naming gives collaborating investigators a shared language to discuss individual pieces of data. While data is often viewed as useful in the aggregate, individual data points can also contain useful insights for regulators (for example, if a data point contains racist labeling or clearly infringes on copyright-protected or licensed work). One year into See:Set's development, we added another big feature—curating collections of data. In doing so, we expanded on the collaboration-fostering features described above. But these collections also allowed us to break down datasets into digestible chunks and focus on representative slices of data. Instead of having to look at one million images containing watermarks, we could instead analyze a collection of a few thousand watermarked images and gain the same insights about their prevalence, sources, metadata, and other features.

See:Set is not meant to be a centralized, comprehensive tool. It helps build a "foundational information architecture to store and query data," which pushes researchers to readjust their assumptions and ask different questions based on the insights that their searches unearth.[56] We describe See:Set here to suggest that data investigations are a crucial element to building a fact-based regulatory approach to GenAI, because they help bolster or challenge our initial assumptions about the content and provenance of training data. The example of See:Set shows that it is possible to interrogate otherwise confounding datasets. More importantly, it can inspire the development of newer and better tools in the future.

III. USCO Should Recommend Further Study and Funding for Training Dataset Best Practices That Facilitate Recordkeeping and Transparency Throughout Dataset Lifecycles.

To aid USCO's consideration of these issues with an eye towards the ultimate report and further recommendations to Congress, Knowing Machines urges USCO to recommend further study and funding to develop specific best practices for training datasets. These best practices should include (1) developing data investigatory tools for use by researchers, regulators, and policymakers to inform regulatory agenda; (2) endorsing standardized, industry-wide dataset transparency and recordkeeping requirements to facilitate data investigations; and (3) adopting guidelines to prevent or minimize the use of depreciated datasets in training AI models.

First, USCO can recommend that Congress allocate a portion of AI research funding towards developing data investigation tools, like our See:Set tool, to enable researchers, regulators, and policymakers to interrogate datasets. While it is easy to operate under the assumption that most training datasets exploit copyright-protected materials without permission, regulating under this assumption may lead to flawed and overbroad policies that fail to distinguish approaches based on copyright status. Instead, USCO should strive to align its policies with factual evidence about what these datasets contain, where those contents come from, and how dataset creators navigate licensed or copyright-protected materials in

[56] Buschek, *supra* note 8.

constructing datasets. Because many dataset creators and AI model developers choose to keep those choices private, the only way to retroactively access these insights is through data investigations paired with transparency and recordkeeping requirements.

Second, USCO can endorse and encourage standardized dataset transparency and recordkeeping requirements that apply to both dataset creators and AI developers.[57] In its recent AI Risk Management Framework, the National Institute of Standards and Technology (NIST) notes that "[m]aintaining the provenance of training data and supporting attribution of the AI system's decision to subsets of training data can assist with both transparency and accountability."[58] Additionally, training data "may also be subject to copyright and should follow applicable intellectual property rights laws."[59] But at this point, keeping track of data provenance is a voluntary element of dataset creation. The "train first, ask questions later" norm in AI development runs against NIST's guidance, so dataset creators should be incentivized, or if necessary required, to adopt standardized data transparency and recordkeeping tools. These tools will then enable data subjects and copyright owners to investigate the status of any data included to determine what rights, if any, they have in the data and what options they have for opting in or out of inclusion in the dataset. USCO can support increased transparency by requiring records on three specific areas of dataset development: (1) the data capture (i.e., images, text, or audio that can be used to train a model); (2) the metadata, or data describing the actual data, such as license information; and (3) the process that compiles the list of candidates and curates the dataset (as described *supra* in Section I.). To understand key aspects related to the construction and outputs of AI systems, all three of these areas must be knowable.[60] Transparency mechanisms should also apply to AI developers so they actively disclose which datasets they used in training, which portions of those datasets were used, why those datasets were used, and whether AI developers consulted the datasets' transparency records before use.

In endorsing dataset transparency and recordkeeping requirements, USCO does not have to reinvent the wheel. Critical AI researchers have offered several transparency

[57] This responds to NOI Questions 15.1.–2.
[58] Nat'l Inst. Standards & Tech., Dep't of Commerce, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* 16 (2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.
[59] *Id.*
[60] NOI Question 15.1.

mechanisms that USCO can consult. For example, in *Datasheets for Datasets,* researcher Timnit Gebru and others argue that datasets should be accompanied by comprehensive datasheets that document their "motivation, composition, collection process, recommended uses," and more.[61] Datasheets are standard features in the electronics industry, where every electronic component has a concomitant datasheet describing its main characteristics, recommended usage, and other critical information.[62] In alignment with the European Union AI Act's Technical Documentation Requirements, requiring datasheets for training datasets would create a standardized, industry-wide process that could increase transparency and accountability for dataset creators.[63] It would also make it easier for USCO officials, AI researchers, and others to trace the contents and provenance of the massive training datasets that undergird today's powerful GenAI. With more documentation surrounding the construction and motivations of these datasets comes less need for retrospective data investigations that attempt to glean information that would otherwise be readily available in the datasheets. USCO can help develop the required fields of information for such datasheets, including the potential copyright or license status of each datapoint in training datasets.[64]

Transparency records should also be made available to the public to promote trust, responsibility, and accountability around AI systems. It is in the public interest that any curious individual can understand the composition of datasets and AI models at the level of data, metadata, and process. While everyone should have the ability to access this knowledge, researchers, journalists, and civil society organizations are often the aggregate stakeholders of this public interest and have direct interests in accessing this information.[65]

Finally, USCO should consider the prevalence of depreciated data in popular training datasets today. Datasets are frequently depreciated for many reasons, including practical

[61] Timnit Gebru et al., *Datasheets for Datasets* 2 (2018), https://arxiv.org/pdf/1803.09010.pdf.

[62] *Id.* at 1.

[63] *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, at Annex IV, COM (2021) 206 final (Apr. 21, 2021), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

[64] A more recent recommendation comes from Google Research members. They endorse "Data Cards" that contain "structured summaries of essential facts about various aspects of ML datasets needed by stakeholders across a dataset's lifecycle for responsible AI development." These summaries include "[e]xplanations of processes and rationales that shape the data and consequently the models—such as upstream sources, data collection and annotation methods; training and evaluation methods, intended use; or decisions affecting model performance." Mahima Pushkarna et al., *Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI* 1 (2022), https://dl.acm.org/doi/fullHtml/10.1145/3531146.3533231.

[65] NOI Question 15.2.

reasons like the creation of a new dataset version or the end of a funding grant.[66] But datasets are also depreciated due to issues around legality and ethics, often in response to valid criticism of or perceived injustices in the dataset's contents. For example, the Tiny Images dataset creators explain that they took the repository offline because it contained "biases, offensive and prejudicial images, and derogatory terminology."[67] Problematic datasets often continue to circulate long after they are depreciated, and the issues compound when they are cited in academic papers or used in training AI systems.[68] Some of the most popular datasets including MS-Celeb-1Mis, Duke MTMC, MegaFace, and Tiny Images have been depreciated but are still widely circulated and used in developing AI systems.[69] Continued use of such depreciated datasets stems in part from a lack of documentation about the reasons for why and how a datasets was removed, including inconsistent public notice and transparency describing removal, a lack of explicit instructions not to use depreciated datasets in AI training, a lack of central directory for depreciated datasets, and no systemic reevaluation of related datasets.[70]

If regulators and policymakers develop reporting mechanisms for dataset creators and AI developers, it will be relatively easy to track when various datasets have been updated due to legal issues. Although data depreciation is a necessary part of the dataset lifecycle and should be encouraged, USCO should consider mechanisms to prevent or minimize the use of datasets depreciated based on legal issues such as copyright concerns. Dataset creators should have clear guidelines to follow for providing public notice of the reason(s) for depreciation and clear instructions not to use the depreciated datasets in AI training. Additionally, USCO can maintain a centralized registry of datasets that have been depreciated due to legal concerns and should not be used anymore.[71]

IV. Conclusion

Knowing Machines supports USCO's comprehensive examination of copyright law and policy issues raised by machine learning technologies, including GenAI. USCO's findings can

[66] *See* Luccioni et al., *A Framework for Deprecating Datasets: Standardizing Documentation, Identification, and Communication* 1 (2022), https://dl.acm.org/doi/10.1145/3531146.3533086.
[67] *Id.* at 9 (internal quotations omitted).
[68] *Id.* at 2.
[69] *Id.*
[70] *Id*. at 5.
[71] *See id.* at 13.

influence congressional policy at a critical juncture in the race between hasty AI product rollouts and regulations preserving human rights. Accordingly, USCO should encourage direct funding to support the creation of data investigation tools to uncover problematic dataset practices, standardized dataset transparency and recordkeeping requirements, and a centralized registry for depreciated datasets that should be avoided in training future AI models. If USCO has any further questions, please reach out to Legal Research Fellow Melodi Dincer at melodi.dincer@law.nyu.edu, Co-Principal Investigator Jason Schultz at jason.schultz@exchange.law.nyu.edu, or Data Investigator Christo Buschek at christo.buschek@proton.me.

Respectfully Submitted,

Melodi Dincer
Legal Research Fellow
Knowing Machines Research Project

Jason Schultz
Professor, NYU School of Law
Co-Principal Investigator
Knowing Machines Research Project

Christo Buschek
Data Investigator & Knowing Machines Fellow
Engelberg Center on Innovation Law & Policy