Position: Data Authenticity, Consent, & Provenance for AI are all broken: what will it take to fix them?

Shayne Longpre [1] **Robert Mahari** [1] **Naana Obeng-Marnu** [1 2] **William Brannon** [1 2] **Tobin South** [1] **Katy Gero** [3] **Sandy Pentland** [1] **Jad Kabbara** [1 2]

Abstract

New capabilities in foundation models are owed in large part to massive, widely-sourced, and under-documented training data collections. Existing practices in data collection have led to challenges in tracing authenticity, verifying consent, preserving privacy, addressing representation and bias, respecting copyright, and overall developing ethical and trustworthy foundation models. In response, regulation is emphasizing the need for training data transparency to understand foundation models' limitations. Based on a large-scale analysis of the foundation model training data landscape and existing solutions, we identify the missing infrastructure to facilitate responsible foundation model development practices. We examine the current shortcomings of common tools for tracing data authenticity, consent, and documentation, and outline how policymakers, developers, and data creators can facilitate responsible foundation model development by adopting universal data provenance standards.

1. The Need for Data Provenance

In the last decade, data from across the web, such as news, social media, and encyclopedias, has become a vital resource in general-purpose, generative AI consumer technologies[1] like GPT-4, Midjourney, and Whisper. These technologies, some with over 100 million weekly users (Malik, 2023), have already begun to catalyze innovation (Brynjolfsson et al., 2018) and scientific inquiry while starting to affect wide swaths of the economy and many everyday consumers (Bommasani et al., 2021). These models are trained on diverse compilations of text, image, and audio data scraped from the web, synthetically generated, or hand-curated. The resulting race to scrape, secure, and mass-produce massive collections of loosely structured data has come with consequences.

Current practices include widely sourcing and bundling data without tracking or vetting their original sources (Longpre et al., 2023a), creator intentions (Chayka, 2023), copyright and licensing status (Bandy & Vincent, 2021), or even basic composition and properties (Dodge et al., 2021). The lack of transparency on this metadata and public infrastructure available to track it has led developers into ethical and legal challenges.

Data that is used for training without significant due diligence has resulted in numerous recent, real-world problems. For instance, the LAION-5B dataset was among the most widely used text-to-image datasets on HuggingFace, before being removed after thousands of Child Sexual Abuse Material (CSAM) images were reported (Internet Watch Foundation, 2023; David, 2023a). The use of certain data sources has triggered intellectual property disputes, culminating in lawsuits against Stability AI, Midjourney, and OpenAI (see e.g., Tremblay v. OpenAI and The New York Time v. Microsoft Corporation). Additionally, there is evidence that foundation models can leak personally identifiable information (PII) (Carlini et al., 2021; Nasr et al., 2023), generate non-consensual intimate imagery (NCII), create misinformation or deepfakes (Rawte et al., 2023; Westerlund, 2019), and proliferate biases or discrimination (Buolamwini & Gebru, 2018). Methods to retract, or 'unlearn', data from a model after training is complete are currently of limited reliability (Villaronga et al., 2018). Machine unlearning methods often fail to fully remove the intended information, or can harm other aspects of a production model, disincentivizing their use (Xu et al., 2023; Shaik et al., 2024). As a result, early choices around training data have long-term

[1]Media Lab, Massachusetts Institute of Technology, Cambridge, USA [2]MIT Center for Constructive Communication, Cambridge, USA [3]Harvard University, Cambridge, USA. Correspondence to: Shayne Longpre <slongpre@media.mit.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]This work is scoped specifically to AI technologies often called foundation models that are (a) generative, meaning they can create potentially inauthentic text, images or other content, and (b) are typically general-purpose, and therefore tend to scrape vast, diverse distribution of web data, that are under-documented, curated, and checked for consenting use.

consequences, creating a pressing need for resources that allow developers to find and fully understand the benefits and risks of different training datasets.

These issues have motivated new data infrastructure and frameworks (which we survey in Section 5) to overcome challenges in sourcing training data responsibly. We isolate several tools in the existing ecosystem for foundation model data management that we identify as commonly used or cited in the AI literature. However, while these tools offer promising solutions, we find they neglect key aspects of the problems, lack interoperability with parallel standards/tools, or haven't yet succeeded in reaching widespread adoption (Section 5.5 and Section 6).

This position paper argues that no complete system for data provenance exists despite a multitude of solutions to different elements of the problem. **A unified framework dedicated to the structured documentation of data properties is needed.** This requires action from multiple stakeholders:

- Data Creators: Adopt and advocate for standardized annotation practices; implement tagging and licensing for content.
- AI Developers: Commit to providing structured data provenance documentation; actively contribute to and utilize dataset libraries in model training.
- Regulators: Set and enforce minimum standards for data provenance for major developers and providers; provide necessary support and funding for the development of data libraries.
- Researchers: Foster norms around data provenance in academic research; collaborate with other stakeholders in developing universal data provenance standards.

To respond to this urgent need, we outline the importance of data provenance for different stakeholders (Section 2), recognition of the problem (Section 3), and legal and regulatory considerations (Section 4). Moreover, we investigate the shortcomings in existing data infrastructure and highlight how the current set of solutions to trace data authenticity, consent, and provenance have limitations and trade-offs (Section 5). Finally, we outline recommendations to facilitate a standardized library for tracing critical data features, emphasize the need for interoperable data documentation standards, and address challenges facing informed and responsible AI (Section 6 and Section 7).

2. Who Needs Data Provenance?

2.1. Creators: Protecting Rights and Avoiding Harms

Artists and data creators emphasize data provenance in their pursuit of "the three 'C"'s of creative rights: compensation, control, and credit (Chayka, 2023). Use of creators' work in computational contexts prior to the current AI trends, such as in academic research or for transformative works, followed social norms which developed over many years of practice (Fiesler & Bruckman, 2019; Klassen & Fiesler, 2022). Current AI technologies have side-stepped such social norms, leaving creators' dissatisfied and unempowered. Creators' work is frequently used to train commercial AI models which can be used to generate competing works. As plaintiffs in lawsuits against major AI companies (e.g. NY Times v. Microsoft and Sancton v. OpenAI) and as part of the "Writer's Strike" (Writers Guild of America, 2023), creators have raised concerns about the legality and ethics of using their data as well as the resulting effects on the creative economy (Epstein et al., 2023). A data transparency and provenance framework can help address these problems: it would afford creators valuable insight into how their work is used in AI, giving them an opportunity to provide consent for their data to be used, verify proper credit, and seek fair compensation in applicable cases where their data is used.

The unsettled legal status of AI training data (Congressional Research Service, 2023) has led to some initial compensation proposals (Cheng, 2023), over which creators have little control. Creators are also rarely credited as having contributed training data, despite the occasional artist signature or watermark that slips through into generated content (Vincent, 2023a). This lack of data transparency and legal clarity has led to several lawsuits against leading AI companies (Vincent, 2022; Small, 2023; Brittain, 2023; Vincent, 2023b) and to calls from creators and publishers for strong transparency requirements around AI data (Cooke, 2023; David, 2023b; Bommasani et al., 2023b; 2024).

The prospective benefits to creators, however, are not only about reducing harms. Greater data transparency can help users—including creators—know which model is best suited to their needs. The disclosure of whether and to what extent a model has been trained on a particular language, literary genre, or style of visual art, for example, can help artists in these media find the right AI tool for their work. Such disclosures also assist creators who see benefits in the use of AI technology in artistic practice (Smee, 2023) by highlighting likely gaps in model coverage and abilities and flagging potential misuse. As subject matter experts, creators could also be effective in recommending or providing new data sources.

2.2. Developers: Informed Data Use

AI developers have an acute interest in data and its provenance for model performance, model behavior, and for anticipating limitations and risks. While model performance generally improves with more data, the quality and diversity of data are also critical factors for reliable performance (Ka-

plan et al., 2020; Hoffmann et al., 2022). Model behavior tends to emulate the structure and composition of data at both the pretraining and finetuning stages (Chung et al., 2022; Longpre et al., 2023b). For these reasons, developers curate specialized pretraining corpora for scientific writing (Lewkowycz et al., 2022), code (Li et al., 2022), biomedical content (Singhal et al., 2023) and legal works (Henderson et al., 2022). As such, information about data sources and their properties informs AI model training.

Data documentation, provenance and analysis tooling have proven particularly important for helping practitioners understand very large datasets (Bender & Friedman, 2018; Gebru et al., 2021). Existing datasets suffer from discoverability of issues related to biased or sexual content (David, 2023a)), private data (Nasr et al., 2023), copyright infringing data (Min et al., 2023), or non-commercially restricted data (Heath, 2023). Lastly, reproducibility and scientific progress more broadly are accelerated by data transparency and structured documentation (Kapoor & Narayanan, 2022).

Tools to trace data origins (Longpre et al., 2023a), and provide large corpus analysis tools (Gao et al., 2020; Dodge et al., 2021; Elazar et al., 2023) are increasingly relevant for more informed and responsible model development. The Hugging Face platform has structured documentation for models and datasets (Wolf et al., 2019; Lhoest et al., 2021)[2], while the Data Provenance Initiative (Longpre et al., 2023a) has closely traced data properties, permissions, and lineage. Model developers recognize that detailed systematic coverage of training data would enable more informed modeling, deeper analysis, and accessibility, and increase the usage of useful datasets that remain underutilized for lack of documentation. At the same time, data transparency provides model developers with information needed to avoid unintentional leakage of synthetic data into the training set (Shumailov et al., 2023).

2.3. Science & Scholarship: Enabling Research

Standardizing the data ecosystem promises to enhance scientific enquiry into AI and unlock new opportunities in adjacent scientific fields. Indeed, there is already considerable research interest in the economic impacts of generative AI (McElheran et al., 2023; Brynjolfsson et al., 2023), from productivity improvements (Dell'Acqua et al., 2023; Peng et al., 2023) to labor exploitation (Hao & Seetharaman, 2023) to market concentration (Vipra & Korinek, 2023), and beyond. The AI industry itself is also a subject of social science research, and its inner workings are of interest to several social science disciplines (Lee et al., 2023; Ziems et al., 2023). Training data, as an important part of this industry's supply chain (Bommasani et al., 2023c; Cen et al., 2023), is central and can only be properly studied if easily accessible. Models themselves are also increasingly used in and for research, from the use of generative AI to annotate data (Ding et al., 2023; Gilardi et al., 2023) — sometimes covertly used by crowdworkers (Veselovsky et al., 2023) — to the rise of LLM-based agents (Xi et al., 2023) in applications as complex as simulations of entire societies (Horton, 2023). There are also studies of phenomena like public opinion (Chu et al., 2023) based on training diets.

In all of these cases, researchers have less and less understanding of how their data is generated: annotation decisions are more opaque, pretrained models are not always fully open, and model-based agents are neither as controllable as a human confederate nor as realistic as a study participant. A better understanding of what goes into these data-generating models, along dimensions like language, country of origin and tasks present in model training data, can inform us about their strengths, weaknesses, and social biases. Such an understanding begins with systematic data collection, which a data provenance standard can enable, simplify, and make accessible to a significantly larger pool of scholars.

Scholars and researchers across disciplines thus have a strong interest in ensuring model training data is annotated for provenance in a consistent, standardized way — before biases in it can influence future research and policy.

2.4. Broader Impacts: Reducing Risk and Fighting Bias

Societal risks from AI run the gamut from privacy violations (Maiberg, 2023) and exposure of personally identifying information, to systemic economic impacts and job displacement (Autor et al., 2022; Gruetzemacher et al., 2020), to bias and discriminatory behavior (Kurita et al., 2019; Kapoor et al., 2024). These risks are fundamentally tied to the data that models are trained on (Longpre et al., 2023b), which, together with the context and affordances of their applications, dictate their behavior.

Perhaps the most pressing concerns, and the ones most tightly coupled to data, relate to social bias and inequitable behavior. There are already prominent examples of AI systems acquiring and perpetuating the biases present in their training data (Steed & Caliskan, 2021; Kurita et al., 2019; David, 2023a), especially in facial recognition systems (Buolamwini & Gebru, 2018; Raji et al., 2020). Social pressure can only be applied to companies' data choices if they are broadly visible and documented. Post hoc attempts to mitigate or train out biases (Zhang et al., 2018), or to retroactively remove contentious data sources, are per se *reactive*. A more proactive approach could be beneficial by providing deeper insights into model training data. Recent work on data humanism (Lupi, 2017) and data feminism (D'Ignazio & Klein, 2023) illuminates key considera-

[2] https://huggingface.co/datasets/librarian-bots/dataset_cards_with_metadata

tions and frameworks to make data accessible through visualization and transparent data management practices, crucial steps in giving affected communities' agency. A community may, for example, wish not to be represented in or (supposedly) served by an application like facial recognition-based surveillance.

3. Growing Interest in Data Provenance

Existing norms for tracing AI data provenance have major and increasingly widely acknowledged deficits (Bommasani et al., 2023b; Longpre et al., 2023a). Popular AI systems do not disclose even basic information about their training data (e.g., ChatGPT, Bard, Llama 2). The pace of innovation has prompted community calls for more systematic (Gebru et al., 2021; Bender & Friedman, 2018; Mitchell et al., 2019) and extensive (Sambasivan et al., 2021; Rogers, 2021; Bandy & Vincent, 2021; Dodge et al., 2021; Longpre et al., 2023a) data documentation. However, these calls have resulted in uneven adoption and adherence. Documentation issues remain particularly acute for so-called "datasets-of-datasets": massive collections of hundreds of datasets where the original provenance information is often neglected or lost due to the lack of standard structures. Meanwhile, practitioners have called for greater data transparency (Bommasani et al., 2023b; Narayanan & Kapoor, 2023), for data supply chain and ecosystem monitoring (Bommasani et al., 2023a; Competition and Markets Authority, 2023; Khan, 2021), for content authenticity verification (Rosenthol, 2022), for detailed provenance tracing on behalf of reproducible, explainable, and trustworthy AI systems (Kale et al., 2023; Werder et al., 2022), and specifically for a standardized database to document trustworthy data (Lohr, 2023; Data Nutrition Team, 2021; Longpre et al., 2023a).

Regulators and lawmakers in many countries have also shown interest in these objectives. The U.S. and EU have taken significant steps toward data transparency and both the EU AI Act and President Biden's recent executive order on AI[3] include provisions related to transparency, provenance, and the need to thoroughly understand AI models' inputs (See Section 4.2). The EU Act, in particular, spells out specific requirements for providers of foundation models related to training data provenance. In addition, several recent U.S. congressional bills[4][5] propose regulatory regimes for AI that would require data transparency. A voluntary code of conduct put forward by Canadian authorities (Government of Canada, 2023) calls on model developers to publish training data descriptions while UN bodies have recommended international regulations on data rights that enshrine transparency (United Nations, 2023).

This clear interest from both the researchers and lawmakers motivates our work on unified frameworks for data provenance and transparency. While such standards do not address AI risks directly, they are a key prerequisite to assessing risks and helping foster more responsible AI development.

4. The Legal Dimension of Data Provenance

Before exploring the details of a Data Provenance Standard (Section 6), we explore relevant legal considerations. We also outline how lawmakers can pave the way for a standard that achieves important regulatory objectives.

4.1. Provenance and copyright

There are two general ways in which an AI model may violate copyright interests. First, training a model can infringe on the copyrights of those whose works are in the training data or on the copyrights of those who created the training data corpora.[6] Second, specific outputs of an AI model may infringe on the copyrights associated with individual works in the training data.

AI models sometimes produce outputs that closely resemble items in the pretraining data and may thus infringe on the rights of the creators of these works (who rarely consent to their content being used). It is important to underscore that although the use of pretraining data may be protected by fair use,[7] this does not mean that specific output will not create copyright violations. Meanwhile, instruction tuning, finetuning and alignment datasets are frequently used in ways not permitted by their license agreements (Longpre et al., 2023a). These datasets contain expressive elements created for the sole purpose of training machine learning models and thus their use for this purpose is unlikely to be covered by fair use (Mahari et al., 2023).

A robust data provenance standard could help address many important issues related to the use of copyrighted material for AI training. For both pretraining and finetuning, a standard data provenance framework can help mitigate legal risks and aid in the enforcement of copyright interests. Copyright infringement hinges on access to protected works[8] and thus knowing what datasets were used to train

[3] https://www.federalregister.gov/d/2023-24283.

[4] https://www.congress.gov/bill/118th-congress/senate-bill/3312.

[5] https://www.blumenthal.senate.gov/imo/media/doc/09072023bipartisanaiframework.pdf

[6] The creation of these datasets raises its own copyright issues since data is generally *copied* to create them (Lemley & Casey, 2020).

[7] The U.S. doctrine of fair use allows copyrighted material to be used for purposes that are (among other requirements) distant form the original purpose and in ways that have minimal impact on the market for the original work (17 U.S.C. § 107).

[8] *Sid & Marty Krofft Television Prods., Inc. v. McDonald's*

a model and what works are contained in these datasets is critical to assessing copyright issues. As discussed, training data is often mixed and repackaged, which complicates the task of precisely identifying what data was used to train a particular model. A robust framework not only helps creators assert their rights when generated outputs infringe on their copyright, but also helps model developers tune their models to avoid this infringement in the first place. Meanwhile, for finetuning and other curated datasets, a data provenance standard can ensure that model developers have access to accurate license information, making it easier to comply with relevant restrictions. We note that this analysis focuses on the U.S. fair use principle and that other jurisdictions have different approaches to copyright (e.g. the EU's Text and Data Mining Exception and Singapore's Copyright Act 2021).

4.2. AI Regulation

Both the EU AI Act and President Biden's recent Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence directly and indirectly highlight the need for transparency for AI systems. Both texts require clear communication of the limitations of AI systems to consumers. The AI Act requires the disclosure of relevant information about training, validation, and testing datasets for high-risk AI systems and a summary of copyright protected training data used in foundation models. The technical specifications in the Act include specific data provenance information such as how the data was obtained, labeled, and processed. Meanwhile, the Executive Order encourages regulatory agencies to emphasize transparency requirements for AI models to protect consumers. We note that privacy regulations like the EU's GDPR are generally less relevant in the context of finetuning data which does not generally fall under the definition of personal data per Art. 4(1) of GDPR.

4.3. The role of lawmakers in encouraging responsible AI practices

This paper is a call to action for dataset creators, model developers, researchers, and lawmakers. By understanding the nature of AI ecosystems, lawmakers can create incentives that encourage better documentation of new datasets and audits of existing data. While the term "transparency" is often ill-defined in AI regulation, regulators could leverage transparency obligations to encourage model developers to record information about the datasets that have been used to train models. In addition, policymakers could provide funding for research related to data provenance. Today, there are perverse incentives that prevent many companies from disclosing information about their datasets as doing so may increase the probability of legal action. Legal authorities could consider providing safe harbor to organizations that provide necessary information about their datasets to regulators and the public.

Corp., 562 F.2d 1157, 1172 (9th Cir. 1977); *Art Attacks Ink, LLC v. MGA Entm't Inc.*, 581 F.3d 1138, 1143 (9th Cir. 2009)

5. Existing Solutions

No complete system for data provenance exists. Instead, there is a patchwork of solutions to different elements of the problem. We identify four broad categories of these: *content authenticity techniques* that attach themselves to data as it spreads, *opt-in & opt-out tools* that allow content creators to register how content should be used, *data provenance standards* that allow dataset creators to document information about datasets, and *data provenance libraries* that aggregate information on datasets. These interventions imbue unstructured data with more attribution, authenticity, and navigability in machine-interpretable formats, but none is a complete solution to the challenge of data provenance.

5.1. Content authenticity techniques

The growing concern for manipulated media, disinformation, and deepfakes has spurred methods that attempt to embed provenance information directly alongside or into data. In this way, a downstream user can ascertain the data's source and authenticity to avoid copyright issues, ensure academic integrity, or for journalism and fact-finding.

Coalition for Content Provenance and Authenticity (C2PA). A prominent example of content authenticity is the C2PA, a partnership between Adobe, Microsoft, and dozens of other corporations to design a specification that "addresses the prevalence of misleading information online through the development of technical standards for certifying the source and history (or provenance) of media content."[9] To this end, verifiable information may be cryptographically embedded into images, videos, audio, and some types of documents in a way that is difficult to remove and that makes tampering evident.

Digital Watermarks for AI. Digital watermarks have historically been used in visual and audio content. For applications of AI, these watermarks are embedded into machine-generated content (Hsu & Wu, 1999), and more recently into machine-generated text (Kirchenbauer et al., 2023). While these technologies hold promise, they are vulnerable to removal (Zhang et al., 2023), especially for text media (Sadasivan et al., 2023).

[9] https://c2pa.org/

NAME	MEDIA	COVERAGE	VERIFY	SEARCH	EXTEND	ATTR.	FOR
Content Authenticity	Limited	Growing	✔	✗	✗	✔	authenticity
Robots.txt Consent	Webpages	Growing	✔	✔	✗	✔	consent
Consent Registration	All	Growing	✔	✔	✗	✔	consent
Data Standards	All	Growing (uneven)	●	✔	✔	✔	data properties
Common Crawl	Webpages	Wide	✔	✔	✗	✗	data properties
Hugging Face	Limited	Growing (uneven)	✗	●	✔	●	data properties
Data Prov. Init.	Limited	Limited	✔	✔	✔	✔	data properties

Table 1. **Comparison of data provenance interventions.** A summary across interventions of the current scope of data coverage, the ability to verify origins of data or metadata claims, the ability to search for data using metadata, whether the standard is extensible, and the ability for symbolic attribution. We see that content authenticity techniques are usually embedded into or with the data itself, but have limited scope of where they can be applied, and cannot easily be extended or searched over. Consent opt-in has not organized around one standard yet, and only addresses the consent problem in isolation. Data Provenance Standards and their associated libraries can provide significantly more information, extensibility, and structure that is machine readable or searchable. The challenge arises in maintaining them, and ensuring the metadata is verifiable and accurate.

5.2. Opt-In & Opt-Out

Robots.txt The Robots Exclusion Standard uses a robots.txt file in a website's directory to indicate to crawlers (e.g., from search engines) which parts of the website the webmaster would like to include and exclude from search indexing. Though this protocol lacks enforcement mechanisms, major search engines have historically respected it (Keller & Warso, 2023). Recent proposals extend this idea to AI data, including learners.txt (Ippolito & Yu, 2023), ai.txt[10], or "noai" tag by artists (DeviantArt Team, 2023). In response, Google and OpenAI have instantiated their own versions, "User Agents" and "GPTBots", giving websites an avenue to implement opt-out standards.[11] So far, none of these has been widely adopted. While a robots.txt-type implementation signals a website's preferences as an opt-in/opt-out binary, it fails to provide a more nuanced spectrum of preferences (e.g., only non-commercial open-source models may train on my data), as well as other useful metadata.

Consent Registration. Organizations, such Spawning AI, are attempting to build infrastructure for the "consent layer" of AI data.[12] This involves sourcing opt-in and opt-out information from data creators, and compiling this into searchable databases (e.g., their Do Not Train Registry[13]). This approach obtains consent directly from creators rather than web hosts, but its granularity increases the burden associated with compiling these consent databases.

[10] https://site.spawning.ai/spawning-ai-txt

[11] https://developers.google.com/search/docs/crawling-indexing/overview-google-crawlers, https://platform.openai.com/docs/gptbot

[12] https://api.spawning.ai/spawning-api

[13] https://haveibeentrained.com/

5.3. Dataset Provenance Standards

Beyond data authenticity and consent, standards for broader data documentation have been proposed to resolve the many other challenges including data privacy, sensitive content, licenses, source and temporal metadata, as well as relevance for training.

Datasheets, Statements & Cards. Datasheets (Gebru et al., 2021), data statements (Bender & Friedman, 2018), and data cards (Pushkarna et al., 2022) propose documentation standards for AI datasets. Each of these standardizes documentation of AI dataset creators, annotators, language and content composition, innate biases, collection and curation processes, uses, distribution, and maintenance. Though unevenly adopted, these efforts are widely recognized for improving scientific reproducibility and responsible AI development (Boyd, 2021).

Data Nutrition Labels. Based on the FDA's Nutrition Facts label, a Data Nutrition Label[14] automates data documentation using a registration form with 55 mostly free-text question responses.

Data & Trust Alliance's Data Provenance Standard. The Data & Trust Alliance Data Provenance Standard[15] is the product of joint data documentation efforts from 19 corporations, including IBM, Nike, Mastercard, Walmart, Pfizer, and UPS. Motivated by the absence of an "industry-wide method to gauge trustworthiness of data based on how it was sourced", this standard assimilates a wide variety of industry documentation needs into a succinct structured format. The standard provides structured documentation analogous to Nutrition Labels, but also provides a way to trace data lineages.

[14] https://datanutrition.org/

[15] https://dataandtrustalliance.org/our-initiatives/data-provenance-standards

5.4. Data Provenance Libraries

While content authenticity is embedded in the data, data documentation and provenance standards need to be aggregated in libraries for searching, filtering, and machine navigation. Prior work has formalized and even operationalized data governance standards (Jernite et al., 2022; Pistilli et al., 2023; Desai et al., 2023; Porciuncula & De La Chapelle, 2022), but only a few efforts have gained traction in guiding AI development.

Common Crawl. For pretraining text data, models routinely rely on the Common Crawl.[16] For instance, sixty percent of GPT-3 training data is from Common Crawl (Brown et al., 2020). This resource provides URLs, scrape times, and the raw documents, following the robots.txt exclusion protocol, and even adheres to requests for removal (e.g., the New York Times requested their data removed from the repository (Barr & Hays, 2023)). This free library of crawled web data has wide adoption, is accurate, and has comprehensive web coverage, but it provides limited metadata.

Hugging Face Datasets. Hugging Face Datasets has become a widely adopted data library for AI (Lhoest et al., 2021). It has integrated data cards and Spawning's data consent information into datasets to encourage documentation and consent filtering. While its coverage of AI data is vast, the documentation is uneven and often incorrect, as this information is loosely crowdsourced. Hugging Face data cards also have limited structure and searchability in the current API.

The Data Provenance Initiative. The Data Provenance Initiative[17] is a joint effort by AI and legal experts to contribute comprehensive and accurate structured information around the most popular textual datasets in AI, including their lineage of sources, licenses, creators, and characteristics (Longpre et al., 2023a). This provides a richer and more accurate collection of annotations and provenance for searchable, filterable, and standardized datasets. However, this accuracy requires expert human labor, limiting its scale compared to Hugging Face or Common Crawl.

5.5. Discussion & Trade-Offs of Interventions

Each category of interventions above targets different problems and comes with clear trade-offs in their benefits and limitations, as illustrated in Table 1. For instance, content authenticity techniques offer built-in and verifiable provenance. However, they only authenticate the source or veracity of the data, without covering (or being easily extensible to) other important metadata for AI applications like copyright, privacy, bias considerations, characteristics, intended uses, or complex lineage information. See Table 2 for a more comprehensive list of the different metadata that has become important to AI development in recent years. Content authenticity techniques also apply primarily to atomic units of data, like individual images, recordings, or text files, rather than derivations or compilations, which are increasingly common for multimodal AI training.

Proposals to extend robots.txt or register data opt-in/out are designed to facilitate creator consent, but each AI company requires custom code for their own scrapers, and many AI developers may still ignore these guidelines.

On the other end of the spectrum of data richness are standards like Datasheets, Data Nutrition Labels, or The Data & Trust Alliance's Data Provenance Standard, which encode more fine-grained information, but at the expense of accuracy and adoption incentives. For instance, consider the three data libraries discussed in Section 5.4. These three libraries trade off a spectrum of (a) the coverage of data, (b) the depth of provenance documentation, and (c) the accuracy of collected metadata. Common Crawl is accurate with wide coverage, but not detailed. Hugging Face can be inaccurate with varying levels of detail, but it has extensive coverage. The Data Provenance Initiative is highly accurate and detailed but is currently limited in scope.

Clearly, authenticity techniques, data consent mechanisms, and data provenance standards are complementary, and each conveys distinct and important information to AI developers. Unifying these frameworks into a standardized data infrastructure layer holds tremendous promise and is a precondition to solving the many problems of ethical, legal, and sustainable AI.

6. Elements of a Data Provenance Standard

Model developers, data creators, and the public could all benefit from transparency into AI data. A standard data provenance framework could address these diverse needs, but existing solutions tend to address different transparency problems *in isolation*. Instead of proposing a new standard, we highlight how existing standards can be unified to effectively address the range of relevant challenges. *A unified data provenance framework should be:*

1. **Modality and source agnostic**: An effective standard for tracing wide-ranging data provenance should not be limited to certain modalities (e.g. text, images, video, or audio), or sources.
2. **Verifiable**: The metadata can be verified, or its reliability assessed. Although metadata will inevitably contain errors, editing systems (like the one used by Wikipedia) or provenance confirmations created by data creators help to provide transparency and consistency.

[16] https://commoncrawl.org/
[17] www.dataprovenance.org

METADATA	DEFINITION
Source Authenticity	The authenticity of the data (including digital watermarks and embedded authenticity). (Rawte et al., 2023; Westerlund, 2019)
Consent & Use Restrictions	What uses or audiences the creators consent to. (Chayka, 2023; Epstein et al., 2023)
Data Types	The data types (text, images, tabular, etc.) and their digital formats.
Source Lineage	Links to data sources from which this was assembled or derived. (Buolamwini & Gebru, 2018)
Generation Methods	How the data was created, by which people, organizations, and/or models. (Longpre et al., 2023a)
Temporal Information	When the data was created, collected, released, or edited. (Luu et al., 2022; Longpre et al., 2023b)
Private information	If personally identifiable or private information is present. (Carlini et al., 2021; Nasr et al., 2023)
Legal considerations	What legal information is attached to the data.
Intellectual property	Copyright, trademark, or patent information attached. (Saveri et al., 2023)
Terms	Associated terms of use or service accessing the data. (Heath, 2023)
Regulations	Relevant regulations, though they may be jurisdiction-dependent (e.g., the EU AI Act).
Characteristics	An extensible set of properties for the data, relevant to different applications.
Dimensions	Size of the data in measurable units.
Sensitive content	If offensive, toxic, or graphic content is present. (David, 2023a; Internet Watch Foundation, 2023)
Quality metrics	Metrics associated to quality measures of the data. (Longpre et al., 2023b)
Metadata Contributors	A version controlled log of edits and responsible contributors to this metadata.

Table 2. **Relevant data properties** for facilitating authenticity, consent, and informed use of AI data. Each row provides a citation motivating the significance of that data property. A unified data standard would be able to codify these information pieces.

3. **Structured**: Structured information should be searchable, filterable, and composable, so automated tools can navigate the data, and the qualities of combined datasets can easily be inferred from merging their structured properties (e.g. combining license types).
4. **Extensible and adaptable**: An extensible framework is adaptable to new types of metadata that emerge over time as well as to various jurisdiction-specific transparency requirements.
5. **Symbolically attributable**: Relevant data sources should be attributed, even as datasets are repackaged and compiled. Codifying the lineage of sources allows the resulting properties to be determined by traversing the web of data "ancestors".

7. Key Lessons for Data Provenance

Existing data provenance solutions are piece-meal. Without a robust, well-resourced data provenance framework, developers will struggle to accurately identify and evaluate the safety, copyright implications, and relevance of datasets from a dizzying array of possibilities. Data creators will similarly struggle to identify how and where their content is. Without dataset provenance standards and documentation, creating such a framework will become increasingly difficult and ultimately untenable. While each existing solution provides important insights into the data ecosystem, a robust framework to attach metadata to datasets is needed to track how datasets are mixed, complied and used. We advocate for a set of actions that different stakeholders can take to make data authenticity, consent, and provenance more robust to future challenges.

New work should seek to unify, or make interoperable, data infrastructure for authenticity and consent with other important documentation for privacy, legality, and relevance. Solutions to these problems are being developed in isolation, but trustworthy and responsible AI requires assessing these factors together. Among others, Pistilli et al. (2023) underscore the "necessity of joint consideration of the ethical, legal, and technical in AI ethics frameworks to be used on a larger scale to govern AI systems."

For Policymakers: Regulators play a pivotal role in shaping the future of AI through policy and guidelines. A data-centric approach to AI regulation can help identify and mitigate key risks. Policymakers can provide funding for research related to data provenance and a centralized effort to document and build provenance infrastructure. Currently, perverse legal incentives inhibit companies from disclosing information about their data. Regulators should consider legal or legislative incentives for organizations to provide necessary data transparency, as they have in the DSA for social media platforms and require standardized documentation as part of AI transparency obligations. These types of

incentives can foster universal and interoperable standards for data authenticity, consent, and provenance.

For AI Developers: AI developers are at the forefront of creating AI models and thus bear significant responsibility in ensuring ethical AI practices. It is crucial for developers to prioritize documentation responsibilities and to make public the provenance of their training data. When there are compelling business reasons for confidentiality, at the very least, they should publish aggregate statistics about the data provenance. This level of transparency is essential for building trust with users and the wider community and for fostering a responsible AI ecosystem.

For Data Creators & Compilers: Creators of training data play a critical role in the AI development process. It is imperative for these creators to meticulously document not only consent criteria, but the provenance of their data, including the sources and processing. Repositories and databases to register this information are already available. Such detailed documentation will significantly aid AI developers in respecting underlying rights and in understanding the nature of the data they use.

For the Research Community: The research community is uniquely positioned to set norms and standards around provenance disclosure. This could involve incorporating provenance disclosure as a requirement during research publication which would complement efforts like reproducibility checklists (Rogers et al., 2021) and ultimately help foster scientific progress.

In practice, it would be hard for individual stakeholder groups working in isolation to succeed in developing a data provenance framework. Instead, we believe that an overarching multi-stakeholder collaboration could play a key role in facilitating efforts and operations needed to develop such a standard. A consortium with representatives from each stakeholder group could coordinate the activities of each group and advocate for the adoption of a specific standard. This approach would mirror the W3 Consortium which played a central role in the development of standards for the World Wide Web, or The Data & Trust Alliance, a not-for-profit consortium that brings together businesses and institutions across multiple industries to develop and adopt responsible data and AI practices.

8. Conclusion

We underscore the urgent need for a standardized data provenance framework to address the complex challenges of AI development. The proliferation of AI models, their diverse training data sources and associated ethical, legal, and transparency concerns have culminated in a critical need for a comprehensive approach to data documentation.

Several data provenance solutions exist, such as content authenticity techniques, opt-in/opt-out mechanisms, and data documentation standards. However, each of these addresses specific aspects of a broader issue and functions in isolation. A unified data provenance framework is needed to establish an ecosystem where data authenticity, consent, privacy, legality, and relevance are holistically considered and managed.

The successful implementation of such a framework requires concerted efforts from all stakeholders in the AI field. This includes creators, who need to be empowered to tag and license their content; developers, who must adopt standards for data provenance and contribute to dataset libraries; and policymakers, who should establish transparency standards and fund the documentation and construction of data libraries.

The key takeaway is the interdependence of transparency solutions: without robust data provenance libraries, it is challenging for developers to find and evaluate datasets comprehensively. Conversely, without standardized documentation and metadata attachment to data as it travels across the web, tracking and downstream use become unfeasible. This effort requires the active participation and commitment of all involved parties to create a sustainable and trustworthy AI ecosystem.

Impact Statement

A standard data provenance framework could have far reaching positive impacts on responsible AI development, as outlined throughout. However, such a standard is challenging to design and adopt, as evidenced by the fact that it does not exist yet despite numerous calls from diverse stakeholders. It is important to underscore that while a well designed standard could further important social objectives, a poorly designed standard could further entrench problematic practices: an overly onerous framework could be adopted or worse impose excessive costs on under resourced researchers and developers, further benefiting large corporate AI developers. A standard that fails to capture important information about biases and limitations could also provide a false sense of security, and encourage the use of data that is not fit for purpose. Finally, a data standard cannot directly address most AI harms or risks and while it can help expose them it is not a substitute for well-informed AI policies.

Acknowledgements

We would like to thank Kevin Klyman, Yacine Jernite, Hanlin Zhang, Kristina Podnar, and Saira Jesani for their generous advice and feedback.

References

Autor, D., Chin, C., Salomons, A., and Seegmiller, B. New frontiers: The origins and content of new work, 1940–2018. NBER Working Paper 30389, National Bureau of Economic Research, August 2022. URL http://dx.doi.org/10.3386/w30389.

Bandy, J. and Vincent, N. Addressing "documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. *arXiv preprint arXiv:2105.05241*, 2021.

Barr, A. and Hays, K. The New York Times got its content removed from one of the biggest AI training datasets. here's how it did it. *Business Insider*, November 2023. URL https://www.businessinsider.com/new-york-times-content-removed-common-crawl-ai-training-dataset-2023-11. 2:00 AM PST.

Bender, E. M. and Friedman, B. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Bommasani, R., Kapoor, S., Zhang, D., Narayanan, A., and Liang, P. AI accountability policy request for comment. Department of Commerce, National Telecommunications and Information Administration Docket No. 230407-0093, RIN 0660-XC057, Stanford University; Princeton University, 06 2023a. Response to NTIA's AI Accountability Policy Request for Comment.

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., and Liang, P. The foundation model transparency index, 2023b.

Bommasani, R., Soylu, D., Liao, T., Creel, K. A., and Liang, P. Ecosystem graphs: The social footprint of foundation models. *ArXiv*, abs/2303.15772, 2023c. URL https://arxiv.org/abs/2303.15772.

Bommasani, R., Klyman, K., Longpre, S., Xiong, B., Kapoor, S., Maslej, N., Narayanan, A., and Liang, P. Foundation model transparency reports. *arXiv preprint arXiv:2402.16268*, 2024.

Boyd, K. L. Datasheets for datasets help ML engineers notice and understand ethical issues in training data. *Proceedings of the ACM on Human-Computer Interaction*, 5 (CSCW2):1–27, 2021.

Brittain, B. Lawsuits accuse AI content creators of misusing copyrighted work, 2023. URL https://www.reuters.com/legal/transactional/lawsuits-accuse-ai-content-creators-misusing-copyrighted-work-2023-01-17/.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Brynjolfsson, E., Rock, D., and Syverson, C. The productivity j-curve: How intangibles complement general purpose technologies. NBER Working Paper 25148, National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138, October 2018. URL https://www.nber.org/system/files/working_papers/w25148/w25148.pdf.

Brynjolfsson, E., Li, D., and Raymond, L. R. Generative AI at work. *National Bureau of Economic Research*, Apr 2023. doi: 10.3386/w31161. URL https://www.nber.org/papers/w31161. Working Paper No. 31161.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C. (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Cen, S. H., Hopkins, A., Ilyas, A., Madry, A., Struckman, I., and Videgaray, L. AI supply chains (and why they matter), April 2023. URL https://aipolicy.substack.com/p/supply-chains-2. The second post in our series On AI Deployment.

Chayka, K. Is A.I. Art Stealing from Artists? *The New Yorker*, February 2023. ISSN 0028-792X. URL https://www.newyorker.com/culture/infinite-scroll/is-ai-art-stealing-from-artists. Section: infinite scroll.

Cheng, M. How should creators be compensated for their work training AI models? *Quartz*, October 2023. URL https://qz.com/how-should-creators-be-compensated-for-their-work-train-1850932454.

Chu, E., Andreas, J., Ansolabehere, S., and Roy, D. Language models trained on media diets can predict public opinion, 2023.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Competition and Markets Authority. AI foundation models: Initial report. Government report, Competition and Markets Authority, 09 2023. A report following the CMA's review into AI Foundation Models, and their impact on competition and consumer protection.

Congressional Research Service. Generative artificial intelligence and copyright law. CRS Report No. LSB10922, 2023. URL https://crsreports.congress.gov/product/pdf/LSB/LSB10922.

Cooke, C. Creator groups call for EU AI Act to retain strong transparency obligations. *Complete Music Update*, November 2023. URL https://completemusicupdate.com/creator-groups-ai-act/.

Data Nutrition Team. The Data Nutrition Project. https://datanutrition.org/, 2021. An initiative aimed at providing tools and resources for evaluating the quality of datasets used in AI.

David, E. AI image training dataset found to include child sexual abuse imagery. *The Verge*, December 2023a. URL https://www.theverge.com/2023/12/20/24009418/generative-ai-image-laion-csam-google-stability-stanford. 7:57 AM PST.

David, E. News outlets demand new rules for AI training data. *The Verge*, August 2023b. URL https://www.theverge.com/2023/8/10/23827316/news-transparency-copyright-generative-ai.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. SSRN Scholarly Paper No. 4573321, 2023. Available at SSRN: https://ssrn.com/abstract=4573321.

Desai, M., Jacobs, A., and Card, D. An archival perspective on pretraining data. In *Socially Responsible Language Modelling Research*, 2023.

DeviantArt Team. UPDATE all deviations are opted out of AI datasets. https://www.deviantart.com/team/journal/Tell-AI-Datasets-If-They-Can-t-Use-Your-Content-934500371, November 2023. 3 min read.

D'Ignazio, C. and Klein, L. *Data Feminism*. MIT Press, 2023. ISBN 9780262547185. URL https://mitpress.mit.edu/9780262547185/data-feminism/.

Ding, B., Qin, C., Liu, L., Chia, Y. K., Li, B., Joty, S., and Bing, L. Is GPT-3 a good data annotator? In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11173–11195, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.626. URL https://aclanthology.org/2023.acl-long.626.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., and Gardner, M. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1286–1305, 2021.

Elazar, Y., Bhagia, A., Magnusson, I., Ravichander, A., Schwenk, D., Suhr, A., Walsh, P., Groeneveld, D., Soldaini, L., Singh, S., Hajishirzi, H., Smith, N. A., and Dodge, J. What's in my big data?, 2023.

Epstein, Z., Hertzmann, A., Herman, L., Mahari, R., Frank, M. R., Groh, M., Schroeder, H., Smith, A., Akten, M., Fjeld, J., et al. Art and the science of generative AI. *Science*, 380(6650):1110–1111, 2023.

Fiesler, C. and Bruckman, A. S. Creativity, copyright, and close-knit communities: a case study of social norm formation and enforcement. *Proceedings of the ACM on Human-Computer Interaction*, 3(GROUP):1–24, 2019.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Gilardi, F., Alizadeh, M., and Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023. doi: 10.1073/pnas.2305016120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2305016120.

Government of Canada. Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems. https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems, 2023.

Gruetzemacher, R., Paradice, D., and Lee, K. B. Forecasting extreme labor displacement: A survey of AI practitioners. *Technological Forecasting and Social Change*, 161: 120323, December 2020. ISSN 0040-1625. doi: 10.1016/j.techfore.2020.120323. URL http://dx.doi.org/10.1016/j.techfore.2020.120323.

Hao, K. and Seetharaman, D. Cleaning up ChatGPT takes heavy toll on human workers. *The Wall Street Journal*, Jul 2023. URL https://www.wsj.com/articles/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483.

Heath, A. ByteDance is secretly using OpenAI's tech to build a competitor. *The Verge*, December 2023. URL https://www.theverge.com/2023/12/15/24003151/bytedance-china-openai-microsoft-competitor-llm. 12:21 PM PST.

Henderson, P., Krass, M., Zheng, L., Guha, N., Manning, C. D., Jurafsky, D., and Ho, D. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems*, 35:29217–29234, 2022.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Horton, J. J. Large language models as simulated economic agents: What can we learn from homo silicus?, 2023.

Hsu, C.-T. and Wu, J.-L. Hidden digital watermarks in images. *IEEE Transactions on Image Processing*, 8(1): 58–68, 1999. doi: 10.1109/83.736686.

Internet Watch Foundation. How AI is being abused to create child sexual abuse imagery. https://www.iwf.org.uk/about-us/why-we-exist/our-research/how-ai-is-being-abused-to-create-child-sexual-abuse-imagery/, October 2023.

Ippolito, D. and Yu, Y. W. Donottrain: A metadata standard for indicating consent for machine learning. In *Workshop on Generative AI and Law, ICML 2023*, 2023.

Jernite, Y., Nguyen, H., Biderman, S., Rogers, A., Masoud, M., Danchev, V., Tan, S., Luccioni, A. S., Subramani, N., Johnson, I., et al. Data governance in the age of large-scale data-driven language technology. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2206–2222, 2022.

Kale, A., Nguyen, T., Jr., F. C. H., Li, C., Zhang, J., and Ma, X. Provenance documentation to enable explainable and trustworthy AI: A literature review. *Data Intelligence*, 5(1):139–162, 2023. doi: 10.1162/dint_a_00119. URL TheURLoftherarticle,ifavailable. Received October 9, 2021; Revised January 20, 2022; Accepted February 4, 2022.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models, 2020.

Kapoor, S. and Narayanan, A. Leakage and the reproducibility crisis in ML-based science. *arXiv preprint arXiv:2207.07048*, 2022.

Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., Hopkins, A., Bankston, K., Biderman, S., Bogen, M., Chowdhury, R., Engler, A., Henderson, P., Jernite, Y., Lazar, S., Maffulli, S., Nelson, A., Pineau, J., Skowron, A., Song, D., Storchan, V., Zhang, D., Ho, D. E., Liang, P., and Narayanan, A. On the societal impact of open foundation models, 2024.

Keller, P. and Warso, Z. Defining best practices for opting out of ml training. https://openfuture.eu/publication/defining-best-practices-for-opting-out-of-ml-training/, September 2023.

Khan, M. A new AI lexicon: Open. https://ainowinstitute.org/publication/a-new-ai-lexicon-open, 07 2021. Contextualizing 'Open Data' and AI.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. A watermark for large language models. *arXiv preprint arXiv:2301.10226*, 2023.

Klassen, S. and Fiesler, C. "this isn't your data, friend": Black Twitter as a case study on research ethics for public data. *Social Media+ Society*, 8(4):20563051221144317, 2022.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., and Tsvetkov, Y. Measuring bias in contextualized word representations. ArXiv, 2019. URL https://arxiv.org/abs/1906.07337.

Lee, K., Cooper, A. F., and Grimmelmann, J. Talkin"bout ai generation: Copyright and the generative-AI supply chain. *arXiv preprint arXiv:2309.08133*, 2023.

Lemley, M. A. and Casey, B. Fair learning. *Texas Law Review*, 99:743, 2020.

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

Lhoest, Q., del Moral, A. V., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., et al. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 175–184, 2021.

Li, Y., Choi, D., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Eccles, T., Keeling, J., Gimeno, F., Dal Lago, A., et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

Lohr, S. Big companies find a way to identify A.I. data they can trust. *The New York Times*, 11 2023. URL https://www.nytimes.com/2023/11/30/business/ai-data-standards.html.

Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., et al. The Data Provenance Initiative: A large scale audit of dataset licensing & attribution in ai. *arXiv preprint arXiv:2310.16787*, 2023a.

Longpre, S., Yauney, G., Reif, E., Lee, K., Roberts, A., Zoph, B., Zhou, D., Wei, J., Robinson, K., Mimno, D., and Ippolito, D. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity, 2023b.

Lupi, G. We've reached peak infographics. are you ready for what comes next? *Print Mag*, 161, January 2017. URL https://www.printmag.com/article/data-humanism-future-of-data-visualization/.

Luu, K., Khashabi, D., Gururangan, S., Mandyam, K., and Smith, N. A. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5944–5958, 2022.

Mahari, R., Longpre, S., Donewald, L., Polozov, A., Lipsitz, A., and Pentland, S. Comments regarding artificial intelligence and copyright. Submitted to U.S. Copyright Office, 2023.

Maiberg, E. 404 Media generative AI market analysis: People love to cum. *404 Media*, Sep 2023. URL https://www.404media.co/404-media-generative-ai-sector-analysis-people-love-to-cum/.

Malik, A. OpenAI's ChatGPT now has 100 million weekly active users. *TechCrunch*, Nov 2023. URL https://techcrunch.com/2023/11/06/openais-chatgpt-now-has-100-million-weekly-active-users/.

McElheran, K., Li, J. F., Brynjolfsson, E., Kroff, Z., Dinlersoz, E., Foster, L., and Zolas, N. AI adoption in America: Who, what, and where. Working Paper 31788, National Bureau of Economic Research, October 2023. URL http://dx.doi.org/10.3386/w31788.

Min, S., Gururangan, S., Wallace, E., Hajishirzi, H., Smith, N. A., and Zettlemoyer, L. Silo language models: Isolating legal risk in a nonparametric datastore. *ArXiv*, abs/2308.04430, 2023. URL https://arxiv.org/abs/2308.04430.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

Narayanan, A. and Kapoor, S. Generative AI companies must publish transparency reports. https://knightcolumbia.org/blog/generative-ai-companies-must-publish-transparency-reports, 06 2023. The debate about AI harms is happening in a data vacuum.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

Peng, S., Kalliamvakou, E., Cihon, P., and Demirer, M. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot, 2023.

Pistilli, G., Muñoz Ferrandis, C., Jernite, Y., and Mitchell, M. Stronger together: on the articulation of ethical charters, legal tools, and technical documentation in ML. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 343–354, 2023.

Porciuncula, L. and De La Chapelle, B. Hello datasphere — towards a systems approach to data governance. https://www.thedatasphere.org/news/hello-datasphere-towards-a-systems-approach-to-data-governance/, February 2022.

Pushkarna, M., Zaldivar, A., and Kjartansson, O. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1776–1826, 2022.

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., and Denton, E. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20. ACM, February 2020. doi: 10.1145/3375627.3375820. URL http://dx.doi.org/10.1145/3375627.3375820.

Rawte, V., Sheth, A., and Das, A. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.

Rogers, A. Changing the world by changing the data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2182–2194, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.170. URL https://aclanthology.org/2021.acl-long.170.

Rogers, A., Baldwin, T., and Leins, K. Just what do you think you're doing, Dave?'a checklist for responsible data use in NLP. *arXiv preprint arXiv:2109.06598*, 2021.

Rosenthol, L. C2PA: the world's first industry standard for content provenance (conference presentation). In Tescher, A. G. and Ebrahimi, T. (eds.), *Applications of Digital Image Processing XLV*. SPIE, October 2022. doi: 10.1117/12.2632021. URL http://dx.doi.org/10.1117/12.2632021.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. M. "everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445518. URL https://doi.org/10.1145/3411764.3445518.

Saveri, J. R., Zirpoli, C., Young, C. K., and McMahon, K. J. Paul Tremblay, Mona Awad vs. OpenAI, Inc., et al., 2023. URL https://storage.courtlistener.com/recap/gov.uscourts.cand.414822/gov.uscourts.cand.414822.1.0_1.pdf. Case 3:23-cv-03223-AMO Document 1 Filed 06/28/23, UNITED STATES DISTRICT COURT, NORTHERN DISTRICT OF CALIFORNIA, SAN FRANCISCO DIVISION.

Shaik, T., Tao, X., Xie, H., Li, L., Zhu, X., and Li, Q. Exploring the landscape of machine unlearning: A comprehensive survey and taxonomy, 2024.

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., and Anderson, R. The curse of recursion: Training on generated data makes models forget, 2023.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

Small, Z. Sarah Silverman sues OpenAI and Meta over copyright infringement, 2023. URL https://www.nytimes.com/2023/07/10/arts/sarah-silverman-lawsuit-openai-meta.html.

Smee, S. AI is no threat to traditional artists. but it is thrilling. *The Washington Post*, February 2023. URL https://www.washingtonpost.com/arts-entertainment/2023/02/15/ai-in-art/.

Steed, R. and Caliskan, A. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21. ACM, March 2021. doi: 10.1145/3442188.3445932. URL http://dx.doi.org/10.1145/3442188.3445932.

United Nations. A global digital compact — an open, free and secure digital future for all. Publisher or Organization, May 2023. URL https://indonesia.un.org/sites/default/files/2023-07/our-common-agenda-policy-brief-gobal-digi-compact-en.pdf. Our Common Agenda Policy Brief #5.

Veselovsky, V., Ribeiro, M. H., and West, R. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks, 2023.

Villaronga, E. F., Kieseberg, P., and Li, T. Humans forget, machines remember: Artificial intelligence and the right to be forgotten. *Computer Law & Security Review*, 34(2): 304–313, 2018.

Vincent, J. The lawsuit that could rewrite the rules of AI copyright, 2022. URL https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data.

Vincent, J. Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement — theverge.com. https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion, 2023a. [Accessed 18-12-2023].

Vincent, J. Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content. https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit, 2023b. URL https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit.

Vipra, J. and Korinek, A. Market concentration implications of foundation models: The invisible hand of ChatGPT. *Brookings*, Sep 2023. URL https://www.brookings.edu/articles/market-concentration-implications-of-foundation-models-the-invisible-hand-of-chatgpt/.

Werder, K., Ramesh, B., and Zhang, R. S. Establishing data provenance for responsible artificial intelligence systems. *ACM Transactions on Management Information Systems*, 13(2):Article 22, 03 2022. doi: 10.1145/3503488. URL https://dl.acm.org/doi/pdf/10.1145/3503488.

Westerlund, M. The emergence of deepfake technology: A review. *Technology innovation management review*, 9 (11), 2019.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Writers Guild of America. WGA negotiations—status as of may 1, 2023, May 2023. URL https://www.wga.org/uploadedfiles/members/member_info/contract-2023/WGA_proposals.pdf.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huang, X., and Gui, T. The rise and potential of large language model based agents: A survey, 2023.

Xu, H., Zhu, T., Zhang, L., Zhou, W., and Yu, P. S. Machine unlearning: A survey, 2023.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18. ACM, December 2018. doi: 10.1145/3278721.3278779. URL http://dx.doi.org/10.1145/3278721.3278779.

Zhang, H., Edelman, B. L., Francati, D., Venturi, D., Ateniese, G., and Barak, B. Watermarks in the sand: Impossibility of strong watermarking for generative models. *arXiv preprint arXiv:2311.04378*, 2023.

Ziems, C., Shaikh, O., Zhang, Z., Held, W., Chen, J., and Yang, D. Can large language models transform computational social science? *Computational Linguistics*, pp. 1–53, 2023.