# HOW TO FIX DATA AUTHENTICITY, DATA CONSENT & DATA PROVENANCE FOR AI

By Shayne Longpre, Robert Mahari, Naana Obeng-Marnu, William Brannon, Tobin South, Jad Kabbara, Alex Pentland

## IN THIS BRIEF


- *New artificial intelligence (AI) capabilities rely on massive, widely sourced and undocumented collections of training data. Dubious collection practices have spurred crises in data transparency, data authenticity, data privacy, copyright infringement and more.*
- *In response, new global regulations are being formulated to promote greater data transparency. The authors maintain that these regulations are needed to understand AI models' limitations.*
- *The authors identify the missing infrastructure to ease responsible AI development practices. They also explain why existing tools for data authenticity, consent and documentation are individually insufficient to solve this problem on a global scale.*
- *Looking ahead, the authors propose how policymakers, developers, data creators and researchers can pave the way for responsible AI development by creating and applying universal standards for data provenance.*


Online data from across the web, news sites, social media and encyclopedias are a vital resource for Generative Artificial Intelligence (GenAI) technologies such as ChatGPT and Midjourney, a generator of images based on natural language prompts. These models are trained on diverse compilations of text, image and audio data, typically scraped from the web, generated by other sources, or manually collected.

An arms race to collect this loosely structured data has come with troubling consequences. GenAI systems often bundle scraped data without vetting the original sources, creator intentions, copyright and licensing status, or even basic composition and properties (Longpre et al., 2023; Chayka, 2023; Bandy & Vincent, 2021). This data can lead to, or accentuate, many real-world problems such as leaking personal information, generating intimate images or child sexual-abuse materials, creating misinformation and "deepfakes," proliferating biases and discrimination, and triggering intellectual-property disputes.

Also, once a GenAI system has been trained—itself an expensive and time-consuming process—AI developers have no reliable way of retracting data from a model. As a result, early choices made when training large-scale machine learning systems can lead to serious long-term consequences.

## INTRODUCING DATA PROVIDENCE

We believe that one important tool in correcting this situation is known as data provenance. It involves discovering a piece of data's origin, source and history of ownership.

Data provenance can help data creators, AI developers and society as a whole. Data creators benefit by knowing how their work is used in AI, giving them an opportunity to provide consent, verify proper credit and seek fair compensation when appropriate. AI developers benefit from data provenance because it leads to greater data transparency. This, in turn, can help developers avoid harmful pitfalls such as using biased data, infringing copyrights and exposing private information. Finally, society as a whole can benefit because data provenance helps to limit social bias and inequitable behavior such as discrimination and the exposure of private information.

> ***Data provenance can help data creators, AI developers and society as a whole.***

The need for AI data provenance is becoming well understood and accepted, leading to community calls for more systematic and extensive data documentation (Gebru et al., 2021; Bender & Friedman, 2018; Mitchell et al., 2019; Sambasivan et al., 2021).

Regulators and lawmakers in many countries have shown interest, too. For example, the U.S. Congress has proposed a regulatory framework (S.3312) for AI that promotes greater data transparency. Similarly, United Nations groups have recommended the adoption of international regulations for data transparency (United Nations, 2023). However, these calls have to date received uneven adoption and adherence.

## A STANDARD FRAMEWORK

What might a standard framework for GenAI look like? First and foremost, it would need to serve the various needs of model developers, data creators and the public, each of whom require structured transparency into the data, but for different reasons. Where existing solutions tend to address these diverse needs in isolation, a standard data provenance framework would address them all.

We maintain that the current patchwork of existing standards (many listed below) can be unified to address the current challenges. This unified data provenance framework would be:

- **Modality- and source-agnostic**: The framework would not be limited to either modalities—such as text, images and video—or sources.
- **Verifiable:** The metadata could be verified, and its reliability assessed. To counter inevitable errors, a combination of editing systems and provenance confirmations would provide greater transparency and consistency.
- **Structured:** Information would be searchable, filterable and composable, allowing automated tools to navigate the data. This would also let developers infer the qualities of combined datasets by merging their structured properties, such as license types.
- **Extensible and adaptable:** The framework would adapt to new types of metadata. It would also adapt to jurisdictional requirements for transparency.
- **Symbolically attributable:** Relevant data sources would be attributable, even after datasets are repackaged and compiled.

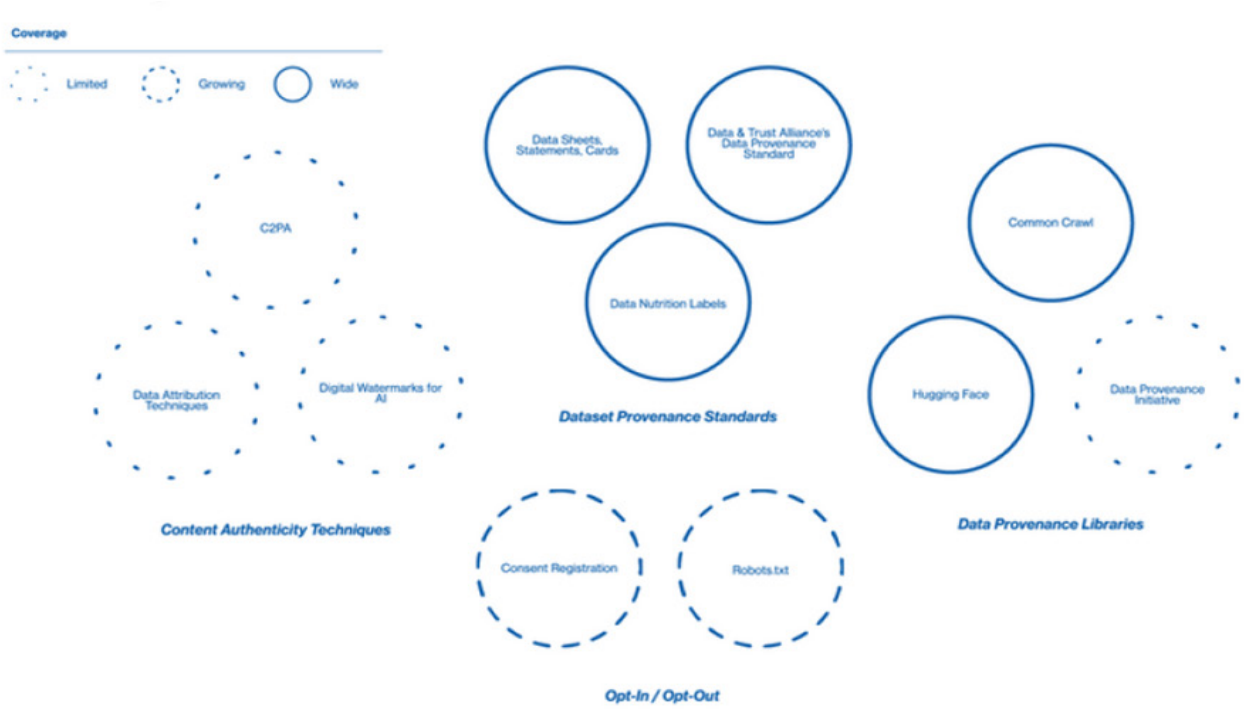
## EXISTING SOLUTIONS—AND THEIR TRADE-OFFS

While no complete system for data provenance currently exists, four broad categories of solutions are now widely available:

- **Content authenticity techniques:** These methods embed provenance information directly alongside or into data, allowing a downstream user to ascertain the data's source and authenticity. Examples include the

**Coalition for Content Provenance and Authenticity (C2PA),** a partnership of Adobe, Microsoft and several other corporations; and digital watermarks for AI, which are embedded into machine-generated content (now including text).

*Figure 1. Comparison of Existing Data Provenance Interventions*




- **Opt-in and opt-out tools:** These let content creators register how their content should be used. One example is the **Robots Exclusion Standard**, which uses a robots.txt file in a website's directory to indicate to crawlers which parts of the site the webmaster would like to include and exclude from search indexing. Similarly, organizations that include **Spawning** are building infrastructure for the consent layer of AI data. Building in this way involves sourcing opt-in and opt-out information directly from the content's creators, not third parties the content's creator, not third parties.

- **Data provenance standards:** These let dataset creators document information about their datasets, thereby heading off challenges such as data privacy, sensitive content and licenses. Several approaches have been tried; these include datasheets, data statements and so-called "data nutrition" labels. Another effort is the **Data & Trust Alliance's** data provenance standard, a joint documentation effort by nearly 20 corporations, among them IBM, Pfizer and Walmart

- **Data provenance libraries:** These libraries aggregate information on datasets and their content, providing a necessary step once content authenticity is embedded in the data. The libraries also allow for searching, filtering and machine navigation. Early examples include **Common Crawl**, a free library of crawled and compiled web data; **Hugging Face Datasets,** a data library with integrated data cards; and the **Data Provenance Initiative**, a joint effort by AI and legal experts to add more comprehensive and structured information around the most popular datasets in AI.

Because each of these four categories target different problems, they come with trade-offs in their benefits and limitations. None offers a complete solution to the challenge of data provenance (*Figure 1*).

Clearly, authenticity techniques, data-consent mechanisms and data-provenance standards are complementary; each conveys distinct and important information to AI developers. For example, while content-authenticity techniques offer built-in and verifiable provenance, they authenticate only the data's source or veracity, overlooking other important metadata such as copyright, privacy and potential bias. Similarly, opt-in and opt-out approaches aim to facilitate creator consent, yet each AI company requires custom code for their own scrapers, and many AI developers may ignore these guidelines.

Nevertheless, unifying these frameworks into a standardized data infrastructure layer holds tremendous promise. In fact, it's a precondition to meeting the challenges of ethical, legal and sustainable AI.

## CALL TO ACTION

The proliferation of AI models—along with their diverse training data sources and associated ethical, legal and transparency concerns—have culminated in the critical need for a comprehensive approach to data documentation.

What's needed is a unified data provenance framework that addresses the complex challenges of AI development. While several data provenance solutions exist, they largely function in isolation, addressing only limited aspects of a much broader issue.

The creation of a unified data provenance framework would help stakeholders establish an ecosystem in which data authenticity, consent, privacy, legality and relevance are all holistically considered and managed. Implementing this framework will require serious efforts by all AI participants, including creators, developers and policymakers.

What's more, solutions for AI transparency need to be interdependent. Without robust and accessible data provenance libraries, AI developers will find it challenging to locate and evaluate datasets. And without standardized documentation and metadata attachment to data, tracking and using data downstream will become unfeasible.

By working together, AI stakeholders can create a data ecosystem that is both sustainable and trustworthy.

## REPORT

Read the **full position paper**

## ABOUT THE AUTHORS

**Shayne Longpre** is a doctoral student and Research Assistant at the MIT Media Lab. His research focuses on training and evaluating large language models.

**Robert Mahari** is a Research Assistant at the MIT Media Lab. He's pursuing a joint J.D.-Ph.D. degree at MIT and Harvard Law School.

**Naana Obeng-Marnu** is a Research Assistant at the MIT Media Lab. She's also a writer, digital artist and designer working toward a master's degree at MIT as an Amazon Robotics Day One Fellow

**William Brannon** is a doctoral student at the MIT Center for Constructive Communication and a Research Assistant at MIT's Laboratory for Social Machines.

**Tobin South** is a Research Assistant at the MIT Media Lab and a doctoral student in the Lab's Human Dynamics group.

**Jad Kabbara** is a postdoctoral researcher working at the MIT Center for Constructive Communication and the MIT Media Lab.

**Alex "Sandy" Pentland** leads the MIT Initiative on the Digital Economy's Building a Distributed Economy research group. He also directs both MIT's Human Dynamics Laboratory and the MIT Media Lab's Entrepreneurship program.

## ACKNOWLEDGEMENTS

The authors wish to thank Katy Gero, Kevin Klyman, Yacine Jernite, Hanlin Zhang, Kristina Podnar and Saira Jesani for their generous advice and feedback.

## REFERENCES

Bandy, J.; Vincent, N. (2021). **Addressing "Documentation Debt" in Machine Learning Research: A Retrospective Datasheet for BookCorpus**. Working paper.

Bender, E. M.; Friedman, B. (2018). **Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science**. Transactions of the Association for Computational Linguistics, no. 6, pp. 587-605.

Chayka, K. (2023). **Is AI Art Stealing from Artists?** The New Yorker.

Gebru, T., et al. (2021). **Datasheets for Datasets.** Communications of the ACM 64, no. 12, pp. 86-92.

Longpre, S., et al. (2023). **The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI**. Working paper.

Mitchell, M., et al. (2019). **Model Cards for Model Reporting**. Proceedings of the Conference on Fairness, Accountability and Transparency, pp. 220-229, ACM, New York.

Sambasivan, N., et al. (2021). **"Everyone Wants to do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI**. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, no. 39, pp. 1-15, ACM, New York.

United Nations (2023). **A Global Digital Compact—An Open, Free and Secure Digital Future for All.** Our Common Agenda Policy Brief No. 5.

MIT INITIATIVE ON THE DIGITAL ECONOMY

**MIT Initiative on the Digital Economy**
MIT Sloan School of Management
245 First St, Room E94-1521
Cambridge, MA 02142-1347
**ide.mit.edu**

**Our Mission:** The MIT Initiative on the Digital Economy (IDE) is shaping a brighter digital future. We conduct groundbreaking research on the promise--and peril--of new digital techonologies including generative artificial intelligence (GenAI), quantum computing, data analytics, and distributed marketplaces. We also investigate the rise of fake news and misinformation and the development of a digital culture. Through research and the convening of leaders from academia, industry, and government, the IDE provides critical, actionable insight for people, businesses, and government to understand and benefit from new technologies and how they're repidly changing the ways we live, work, and communicate.

**Contact Us:** David Verrill, Executive Director,
MIT Initiative on the Digital Economy
617-452-3216
**dverrill@mit.edu**

**Become a Sponsor:** The generous support of individuals, foundations, and corporations help to fuel cutting-edge research by MIT faculty and graduate students. It also enables new faculty hiring, curriculum development, events, and fellowships.

**Additional Contact:** Albert Scerbo, Associate Director,
MIT Initiative on the Digital Economy
267-980-2616
**ascerbo@mit.edu**

**View all our sponsors**

**Connect with us:**