JCOTS
JOINT COMMISSION ON TECHNOLOGY & SCIENCE

Brief: AI Training Data Transparency

Summary

Training data is computer-readable information (datasets) used to teach artificial intelligence (AI) models to identify patterns, make predictions, or generate new content. Datasets can be composed of many kinds of data, from text to numerical data to images or videos. (For more information on how AI models "learn" from training data, please see the JCOTS 2024 report, Artificial Intelligence: Policy and Practice.) AI development relies on extremely large amounts of training data, so collecting, curating, and procuring high-quality, high-volume training datasets is an integral—and sometimes commercially competitive and sensitive—part of the AI industry. The race to develop powerful AI models has created a parallel race to collect and leverage huge amounts of training data in the absence of agreed standards about sourcing, tracking, cleaning, or combining data. AI training data influence a model's ultimate behavior and performance, and mismatches between datasets (e.g. poor quality or irrelevant content) and the real-world setting where they are deployed can result in negative outcomes.

Defining training data

Training data can originate from various sources, and few explicit rules govern how data can be collected and used for AI training in the United States. The choice of training data also sometimes depends on the intended use of the AI model it will be used to train. For example, a chatbot intended for use by clinicians might be trained on a specialized dataset of peer-reviewed medical journal articles. However, foundation models and generative models intended for general use across a wide range of domains and tasks are often trained on vast quantities of diverse data collected from the internet. These generalist models can be later fine-tuned to perform more specific tasks by adjusting the model's parameters, or weights, or introducing additional, specialized datasets.

Several free and public training datasets are widely used in the industry, such as ImageNet (a collection of labeled images used to train machine vision models), Common Crawl (a repository of data scraped from the internet), and DataComp CommonPool (a massive collection of image-text pairs used to train image generation models). Aggregators like the UC Irvine Machine Learning Repository or the Hugging Face Datasets library collate publicly available training datasets. But companies also curate their own datasets, which might come from their own data collections or from datasets purchased from intermediaries. People produce huge amounts of data in their everyday lives—from information generated by smart home devices to social media activity to credit card purchases and much more—so, the market for data is large and lucrative, and these datasets can wind up in AI training data.

Defining the public harm

As generative AI has become a popular consumer product (e.g. ChatGPT, Claude, Gemini), the data used to train these models has received greater attention. Public training datasets that can be more easily examined by journalists and researchers have been shown to contain a large amount of unfiltered toxic content, copyrighted material that creators did not consent to using in AI training, personally identifiable information (such as photographs of people's identification documents), and skewed or unrepresentative data that disproportionately excludes certain racial, ethic, and linguistic groups. These findings have raised questions about how to mitigate public harm, such as **copyright infringement**, **exposure of minors to harmful content**, and **discrimination** when models trained on unrepresentative data are used in the general population, among other possible adverse effects. In addition, little is known about the data companies may be using to train models, as companies are not required to document their training process. Many of these issues relate to **digital consent**, as people often do not know or have not explicitly consented to the use of their data (whether as creators or consumers) in AI training.

In June, two separate cases involving Meta and Anthropic resulted in rulings that signal the use of copyrighted material in AI training data constitutes "fair use" if the material was obtained legally. However, the companies may still be liable for using pirated material, and journalists have uncovered cases of major companies' training datasets containing large quantities of pirated and paywalled content.

Training data transparency as a policy solution

As the raw material that powers AI, training data is an essential component of the pipeline that produces AI outputs, including decisions, predictions, and actions that impact people's everyday lives. Training data transparency generally refers to practices that would (1) render training data scrutable to the public, researchers, and/or oversight and auditing bodies; and (2) provide the public with clear information about how their data could be used in AI training alongside options for opting in or out. There is currently no accepted or leading standard for training data transparency, but scholars have put forward several proposals with many similarities, which are summarized in the table below.

Proposal	Description	Reference
Data sheets	Data sheets are based loosely on the standard practice of providing detailed information on all manufactured components in the electronics industry. Data sheets would require training data creators to keep records of the process of developing these datasets, focusing on the key stages of the dataset lifecycle: motivation, composition, collection process, pre- processing/cleaning/labeling, uses, distribution, and maintenance.	Gebru et al., https://doi.org/10.1145/3458723.
Data statements	Data statements are closely related to data sheets, but specifically focused on natural language processing (NLP) models. They aim to mitigate issues with exclusion and bias in these models by providing context on the datasets used to train them. Documentation should include categories like curation rationale, language variety, and annotator demographics, among others.	Bender & Friedman, https://aclanthology.org/Q18-1041/
Data cards	Data cards are structured summaries of essential facts about machine learning datasets across the dataset's lifecycle, using 31 themes that	Pushkarna et al., https://dl.acm.org/

	could generally describe any dataset and meet four transparency objectives: consistency, comprehensiveness, intelligibility and concision, and explainability and uncertainty.	doi/10.1145/3531146.3533231
Data nutrition labels	Based on the concept of the Nutrition Facts Label, required by the federal Nutrition Labeling and Education Act (1990), a dataset nutrition label is a modular reporting template that includes categories such as metadata, provenance, variables, and statistics. Authors envision the label to be both generated and viewed by web-based applications, requiring dedicated applications for label-making and -reading.	Holland et al., http://arxiv.org/abs/1805.03677
Data provenance standard	A set of data transparency standards established by a working group of technical and industry experts from across 15 industries. The standard stipulates metadata fields to report under three categories: source, provenance, and use.	Data and Trust Alliance, https://dataandtrustalliance.org/work/data-provenance-standards
Data provenance libraries	Some libraries of training datasets (e.g. Common Crawl or Hugging Face) provide information about data origins. However, these reporting conventions are voluntary, and some independent evaluations have found substantial errors and gaps in information provided.	Various; see, for example: Longpre et al., https://arxiv.org/abs/2310.16787

Note: Data transparency and data provenance are often used interchangeably. Provenance has also been used to refer to the origin of AI-generated content, rather than the underlying model(s) or the datasets used to train the model(s).

Critiques of training data transparency

Training data transparency proposes to mitigate the harms outlined above through disclosure—publishing information that would otherwise be hard to ascertain. Guha et al. (2023) identify several considerations and drawbacks to disclosure as a policy remedy. To be effective, disclosures must be understandable, actionable, and verifiable. To address understandability, disclosures need to follow a standard format and apply across the board, but this can lead to "disclosure fatigue," where the disclosure becomes an inconvenience that is ultimately ignored (e.g. cookie warnings). Actionability hinges on providing meaningful choices to different actors (e.g. AI model developers choosing which datasets to use or consumers deciding whether to use an AI product based on its training record). But it is not clear that providing detailed information about training data will provide different actors with information that enhances their agency vis a vis AI systems. Finally, disclosures about training data may be useless if they are not independently verified or audited by a designated body or third party—a pitfall the authors note regarding food nutrition labels.

Key Policy Questions

1. Defining transparency: What standards are needed to ensure consistent and comparable reporting of training data? How should these standards be developed and mandated? How can reporting standards balance disclosure in the public interest and the commercial preservation of trade secrets?

2. Consumer protection: What information should be communicated to consumers about how their data is/can be used to train AI models? What options/recourse should consumers have to assert their data rights in relation to training data?
3. Enforcement: Who should oversee the enforcement and accountability of transparency standards? (i.e. What technical expertise is needed? Who has the right of action?)
4. Exceptions*:* For example, how can transparency requirements avoid unfairly burdening smaller AI developers or requiring infeasible technical interventions/assessments?

Policy Landscape - Virginia

HB2250 - Artificial Intelligence Training Data Transparency Act. Failed in the 2025 regular session. Patroned by Delegate Michelle Lopes Maldonado, the law would have required a generative AI developers to post certain descriptive information about training data on their website and to provide a mechanism for members of the public to opt out of providing data for AI training and make requests to have training data deleted.

Policy Landscape – Other Jurisdictions

AB 2013 (California) - Generative artificial intelligence: training data transparency. Passed and signed by Governor Newsome in 2024. The law requires developers of generative AI systems to publish on their websites a summary of the datasets used to train their models, to include 12 pieces of information, such as sources or owners of the datasets and whether the datasets were purchased or licensed.

Artificial Intelligence Act (European Union) - This act came into force in August 2025 and requires providers of general purpose AI to "draw up and keep up-to-date the technical documentation of the model, including its training and testing process and the results of its evaluation." In July, the European Commission published its "Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models," which details the information providers must supply on training data.

How this brief was developed

- Conducted a search using Bill Track 50 (July 2025). Search parameters: "artificial intelligence training data"
- General web search on "AI training data transparency legislation" and "AI training data transparency"
- Google Scholar search on "AI training data transparency"
- Literature review of sources generated in the web and Google Scholar searches