Full Text | Other Source

This is where the data to build AI comes from

Heikkilä, Melissa; Arnett, Stephanie.

MIT Technology Review.com; Cambridge : Technology Review, Inc. (Dec 18, 2024)

Full text

Listen

AI is all about data. Reams and reams of data are needed to train algorithms to do what we want, and what goes into the AI models determines what comes out. But here's the problem: AI developers and researchers don't really know much about the sources of the data they are using. AI's data collection practices are immature compared with the sophistication of AI model development. Massive data sets often lack clear information about what is in them and where it came from.

The Data Provenance Initiative, a group of over 50 researchers from both academia and industry, wanted to fix that. They wanted to know, very simply: Where does the data to build AI come from? They audited nearly 4,000 public data sets spanning over 600 languages, 67 countries, and three decades. The data came from 800 unique sources and nearly 700 organizations.

Their findings, shared exclusively with *MIT Technology Review*, show a worrying trend: AI's data practices risk concentrating power overwhelmingly in the hands of a few dominant technology companies.

In the early 2010s, data sets came from a variety of sources, says Shayne Longpre, a researcher at MIT who is part of the project.

It came not just from encyclopedias and the web, but also from sources such as parliamentary transcripts, earning calls, and weather reports. Back then, AI data sets were specifically curated and collected from different sources to suit individual tasks, Longpre says.

Then transformers, the architecture underpinning language models, were invented in 2017, and the AI sector started seeing performance get better the bigger the models and data sets were. Today, most AI data sets are built by indiscriminately hoovering material from the internet. Since 2018, the web has been the dominant source for data sets used in all media, such as audio, images, and video, and a gap between scraped data and more curated data sets has emerged and widened.

"In foundation model development, nothing seems to matter more for the capabilities than the scale and heterogeneity of the data and the web," says Longpre. The need for scale has also boosted the use of synthetic data massively.

The past few years have also seen the rise of multimodal generative AI models, which can generate videos and images. Like large language models, they need as much data as possible, and the best source for that has become YouTube.

For video models, as you can see in this chart, over 70% of data for both speech and image data sets comes from one source.

This could be a boon for Alphabet, Google's parent company, which owns YouTube. Whereas text is distributed across the web and controlled by many different websites and platforms, video data is extremely concentrated in one platform.

"It gives a huge concentration of power over a lot of the most important data on the web to one company," says Longpre.

And because Google is also developing its own AI models, its massive advantage also raises questions about how the company will make this data available for competitors, says Sarah Myers West, the co–executive director at the AI Now Institute.

"It's important to think about data not as though it's sort of this naturally occurring resource, but it's something that is created through particular processes," says Myers West.

"If the data sets on which most of the AI that we're interacting with reflect the intentions and the design of big, profit-motivated corporations—that's reshaping the infrastructures of our world in ways that reflect the interests of those big corporations," she says.

This monoculture also raises questions about how accurately the human experience is portrayed in the data set and what kinds of models we are building, says Sara Hooker, the vice president of research at the technology company Cohere, who is also part of the Data Provenance Initiative.

People upload videos to YouTube with a particular audience in mind, and the way people act in those videos is often intended for very specific effect. “Does [the data] capture all the nuances of humanity and all the ways that we exist?” says Hooker.

Hidden restrictions

AI companies don’t usually share what data they used to train their models. One reason is that they want to protect their competitive edge. The other is that because of the complicated and opaque way data sets are bundled, packaged, and distributed, they likely don’t even know where all the data came from.

They also probably don’t have complete information about any constraints on how that data is supposed to be used or shared. The researchers at the Data Provenance Initiative found that data sets often have restrictive licenses or terms attached to them, which should limit their use for commercial purposes, for example.

“This lack of consistency across the data lineage makes it very hard for developers to make the right choice about what data to use,” says Hooker.

It also makes it almost impossible to be completely certain you haven’t trained your model on copyrighted data, adds Longpre.

More recently, companies such as OpenAI and Google have struck exclusive data-sharing deals with publishers, major forums such as Reddit, and social media platforms on the web. But this becomes another way for them to concentrate their power.

“These exclusive contracts can partition the internet into various zones of who can get access to it and who can’t,” says Longpre.

The trend benefits the biggest AI players, who can afford such deals, at the expense of researchers, nonprofits, and smaller companies, who will struggle to get access. The largest companies also have the best resources for crawling data sets.

“This is a new wave of asymmetric access that we haven’t seen to this extent on the open web,” Longpre says.

The West vs. the rest

The data that is used to train AI models is also heavily skewed to the Western world. Over 90% of the data sets that the researchers analyzed came from Europe and North America, and fewer than 4% came from Africa.

“These data sets are reflecting one part of our world and our culture, but completely omitting others,” says Hooker.

The dominance of the English language in training data is partly explained by the fact that the internet is still over 90% in English, and there are still a lot of places on Earth where there’s really poor internet connection or none at all, says Giada Pistilli, principal ethicist at Hugging Face, who was not part of the research team. But another reason is convenience, she adds: Putting together data sets in other languages and taking other cultures into account requires conscious intention and a lot of work.

The Western focus of these data sets becomes particularly clear with multimodal models. When an AI model is prompted for the sights and sounds of a wedding, for example, it might only be able to represent Western weddings, because that's all that it has been trained on, Hooker says.

This reinforces biases and could lead to AI models that push a certain US-centric worldview, erasing other languages and cultures.

"We are using these models all over the world, and there's a massive discrepancy between the world we're seeing and what's invisible to these models," Hooker says.

Suggested sources

Looking for more sources on the same subject? Try these suggestions:

Deep learning driven interpretable and informed decision making model for

Scientific Reports (Nature Publisher Group); London Vol. 15, Iss 1, (2025): 19223.

Considerations for taking forward the UAE-Belém

Lamhauge, Nicolina; Duluk, Margot. **OECD/IEA Climate Change Expert Group Papers; Paris**, 46 pp.Jun 5, 2025.

Reinforcing Chilean policies for the inclusion

OECD Education Policy Perspectives; Paris, 71 pp.Jul 25, 2025.

OECD support towards STI statistical

OECD Science, Technology and Industry Policy Papers; Paris, 32 pp.Jun 17, 2025.

Show more suggestions...