



Popular

Latest

Newsletters

The Atlantic

Saved Stories

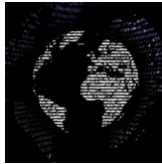
My Account

Give a Gift

MORE FROM ARTIFICIAL INTELLIGENCE

We Don't Actually Know If AI Is Taking Over Everything

KAREN HAO



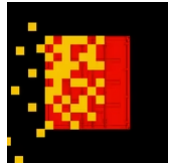
Computers Are Learning to Smell

MATTEO WONG



The New AI Panic

KAREN HAO



TECHNOLOGY




We Don't Actually Know If AI Is Taking Over Everything

A test of AI transparency gave every major company an F.

By Karen Hao



Illustration by The Atlantic. Source: Getty. OCTOBER 19, 2023

SHARE & GIFT  SAVED STORIES  SAVE 

Since the release of ChatGPT last year, I've heard some version of the same thing over and over again: *What is going on?* The rush of chatbots and endless

“AI-powered” apps has made starkly clear that this technology is poised to upend everything—or, at least, something. Yet even the AI experts are struggling with a dizzying feeling that for all the talk of its transformative potential, so much about this technology is veiled in secrecy.

It isn't just a feeling. More and more of this technology, once developed through open research, has become almost completely hidden within corporations that are opaque about what their AI models are capable of and how they are made. Transparency isn't legally required, and the secrecy is causing problems: Earlier this year, *The Atlantic* revealed that Meta and others had used nearly 200,000 books to train their AI models without the compensation or consent of the authors.

Now we have a way to measure just how bad AI's secrecy problem actually is. Yesterday, Stanford University's Center for Research on Foundation Models launched a new index that tracks the transparency of 10 major AI companies, including OpenAI, Google, and Anthropic. The researchers graded each company's flagship model based on whether its developers publicly disclosed 100 different pieces of information—such as what data it was trained on, the wages paid to the data and content-moderation workers who were involved in its development, and when the model should *not* be used. One point was awarded for each disclosure. Among the 10 companies, the highest-scoring

barely got more than 50 out of the 100 possible points; the average is 37. Every company, in other words, gets a resounding F.

Foundation Model Transparency Index Total Scores, 2023

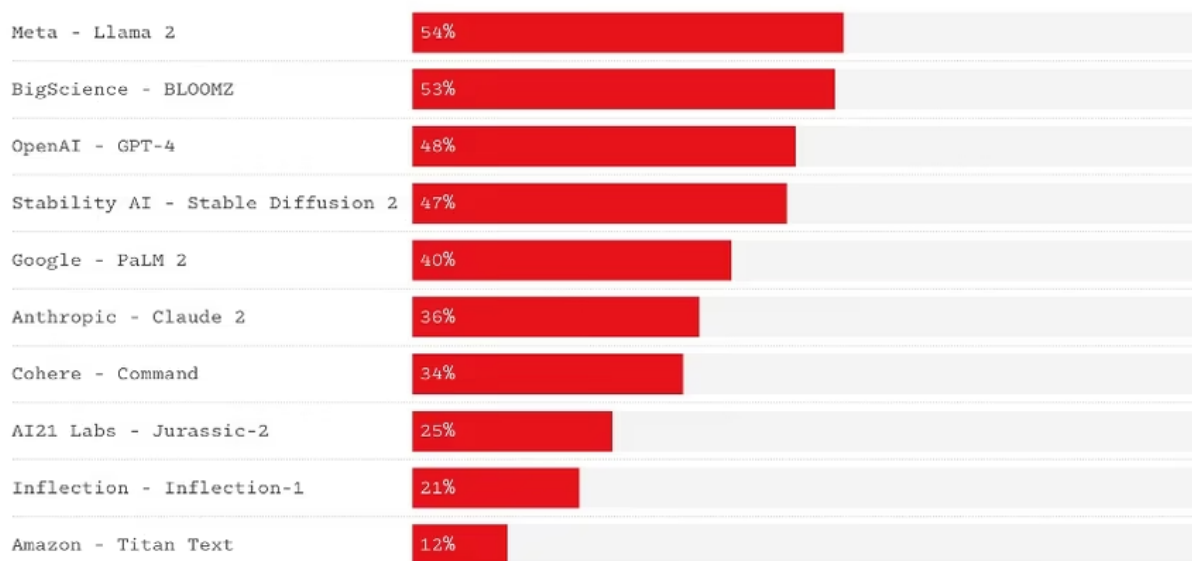


Chart: The Atlantic • Source: Foundation Model Transparency Index / Stanford University



Take OpenAI, which was named to indicate a commitment to transparency. Its flagship model, GPT-4, scored a 48—losing significant points for not revealing information such as the data that were fed into it, how it treated personally identifiable information that may have been captured in said

scraped data, and how much energy was used to produce the model. Even Meta, which has prided itself on openness by allowing people to download and adapt its model, scored only 54 points. “A way to think about it is: You are getting baked cake, and you can add decorations or layers to that cake,” says Deborah Raji, an AI accountability researcher at UC Berkeley who wasn’t involved in the research. “But you don’t get the recipe book for what’s actually in the cake.”

[Read: These 183,000 books are fueling the biggest fight in publishing and tech](#)

Many companies, including OpenAI and Anthropic, have held that they keep such information secret for competitive reasons or to prevent risky proliferation of their technology, or both. I reached out to the 10 companies indexed by the Stanford researchers. An Amazon spokesperson said the company looks forward to carefully reviewing the index. Margaret Mitchell, a researcher and chief ethics scientist at Hugging Face, said the index misrepresented BLOOMZ as the firm’s model; it was in fact produced by an international research collaboration called the BigScience project that was co-organized by the company. (The Stanford researchers acknowledge this in the body of the report. For this reason, I marked BLOOMZ as a BigScience

model, not a Hugging Face one, on the chart above.) OpenAI and Cohere declined a request for comment. None of the other companies responded.

The Stanford researchers selected the 100 criteria based on years of existing AI research and policy work, focusing on inputs into each model, facts about the model itself, and the final product's downstream impacts. For example, the index references scholarly and journalistic investigations into the poor pay for data workers who help perfect AI models to explain its determination that the companies should specify whether they directly employ the workers and any labor protections they put in place. The lead creators of the index, Rishi Bommasani and Kevin Klyman, told me they tried to keep in mind the kinds of disclosures that would be most helpful to a range of different groups: scientists conducting independent research about these models, policy makers designing AI regulation, and consumers deciding whether to use a model in a particular situation.

In addition to insights about specific models, the index reveals industry-wide gaps of information. Not a single model the researchers assessed provides information about whether the data it was trained on had copyright protections or other rules restricting their use. Nor do any models disclose sufficient information about the authors, artists, and others whose works were scraped and used for training. Most companies are also tight-lipped about the

shortcomings of their models, whether their embedded biases or how often they make things up.

That every company performs so poorly is an indictment on the industry as a whole. In fact, Amba Kak, the executive director of the AI Now Institute, told me that in her view, the index was not high *enough* of a standard. The opacity within the industry is so pervasive and ingrained, she told me, that even 100 criteria don't fully reveal the problems. And transparency is not an esoteric concern: Without full disclosures from companies, Raji told me, "it's a one-sided narrative. And it is almost always the optimistic narrative."

[Read: AI's present matters more than its imagined future](#)

In 2019, Raji co-authored a paper showing that several facial-recognition products, including ones being sold to the police, worked poorly on women and people of color. The research shed light on the risk of law enforcement using faulty technology. As of August, there have been six reported cases of police falsely accusing people of a crime in the U.S. based on flawed facial recognition; all of the accused are Black. These latest AI models pose similar risks, Raji said. Without giving policy makers or independent researchers the evidence they need to audit and back up corporate claims, AI companies can easily inflate their capabilities in ways that lead consumers or third-party app

developers to use faulty or inadequate technology in crucial contexts such as criminal justice and health care.

There have been rare exceptions to the industry-wide opacity. One model not included in the index is BLOOM, which was similarly produced by the BigScience project (but is different from BLOOMZ). The researchers for BLOOM conducted one of the few analyses available to date of the broader environmental impacts of large-scale AI models and also documented information about data creators, copyright, personally identifiable information, and source licenses for the training data. It shows that such transparency is *possible*. But changing industry norms will require regulatory mandates, Kak told me. “We cannot rely on researchers and the public to be piecing together this map” of information, she said.

Perhaps the biggest clincher is that across the board, the tracker finds that all of the companies have particularly abysmal disclosures in “impact” criteria, which includes the number of users who use their product, the applications being built on top of the technology, and the geographic distribution of where these technologies are being deployed. This makes it far more difficult for regulators to track each firm’s sphere of control and influence, and to hold them accountable. It’s much harder for consumers as well: If OpenAI technology is helping your kid’s teacher, assisting your family doctor, and

powering your office productivity tools, you may not even know. In other words, we know *so* little about these technologies we're coming to rely on that we can't even say how much we rely on them.

Secrecy, of course, is nothing new in Silicon Valley. Nearly a decade ago, the tech and legal scholar Frank Pasquale coined the phrase *black-box society* to refer to the way tech platforms were growing ever more opaque as they solidified their dominance in people's lives. "Secrecy is approaching critical mass, and we are in the dark about crucial decisions," he wrote. And yet, despite the litany of cautionary tales from other AI technologies and social media, many people have grown comfortable with black boxes. Silicon Valley spent years establishing a new and opaque norm; now it's just accepted as a part of life.
