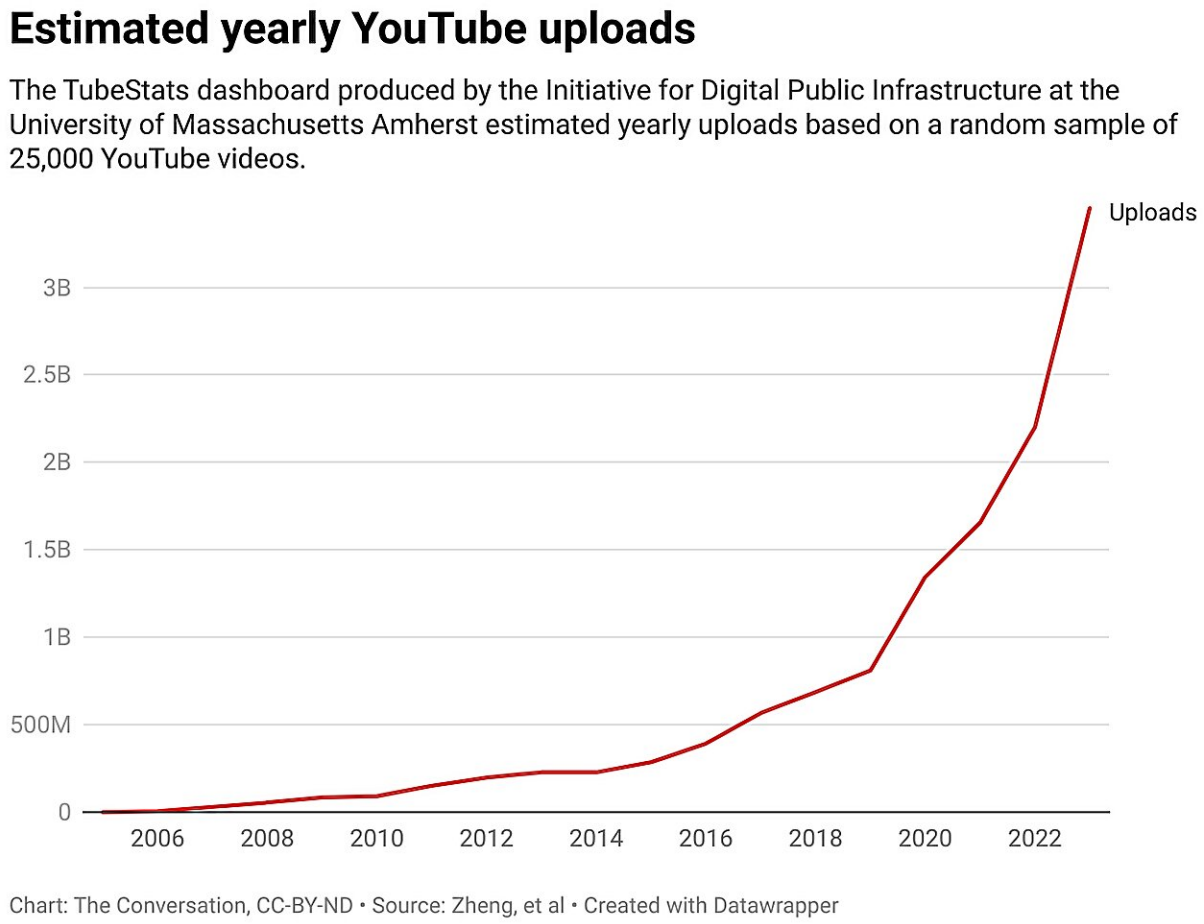
AI companies train language models on YouTube's archive—making family-and-friends videos a privacy risk

June 27 2024, by Ryan McGrady and Ethan Zuckerman



Credit: The Conversation

The promised artificial intelligence revolution requires data. Lots and lots of data. OpenAI and Google have begun using YouTube videos to train their text-based AI models. But what does the YouTube archive actually include?

Our team of digital media researchers at the University of Massachusetts Amherst collected and analyzed random samples of YouTube videos to learn more about that archive. We published an 85-page paper about that dataset and set up a website called TubeStats for researchers and journalists who need basic information about YouTube.

Now, we're taking a closer look at some of our more surprising findings to better understand how these obscure videos might become part of powerful AI systems. We've found that many YouTube videos are meant for personal use or for small groups of people, and a significant proportion were created by children who appear to be under 13.

Bulk of the YouTube iceberg

Most people's experience of YouTube is algorithmically curated: Up to 70% of the videos users watch are recommended by the site's algorithms. Recommended videos are typically popular content such as influencer stunts, news clips, explainer videos, travel vlogs and video game reviews, while content that is not recommended languishes in obscurity.

Some YouTube content emulates popular creators or fits into established genres, but much of it is personal: family celebrations, selfies set to music, homework assignments, video game clips without context and kids dancing. The obscure side of YouTube—the vast majority of the estimated 14.8 billion videos created and uploaded to the platform—is poorly understood.

Illuminating this aspect of YouTube—and social media generally—is difficult because big tech companies have become increasingly hostile to researchers.

We've found that many videos on YouTube were never meant to be shared widely. We documented thousands of short, personal videos that have few views but high engagement—likes and comments—implying a small but highly engaged audience. These were clearly meant for a small audience of friends and family. Such social uses of YouTube contrast with videos that try to maximize their audience, suggesting another way to use YouTube: as a video-centered social network for small groups.

Other videos seem intended for a different kind of small, fixed audience: recorded classes from pandemic-era virtual instruction, school board meetings and work meetings. While not what most people think of as social uses, they likewise imply that their creators have a different expectation about the audience for the videos than creators of the kind of content people see in their recommendations.

Fuel for the AI machine

It was with this broader understanding that we read The New York Times exposé on how OpenAI and Google turned to YouTube in a race to find new troves of data to train their large language models. An archive of YouTube transcripts makes an extraordinary dataset for text-based models.

There is also speculation, fueled in part by an evasive answer from OpenAI's chief technology officer Mira Murati, that the videos themselves could be used to train AI text-to-video models such as OpenAI's Sora.

The New York Times story raised concerns about YouTube's terms of

service and, of course, the copyright issues that pervade much of the debate about AI. But there's another problem: How could anyone know what an archive of more than 14 billion videos, uploaded by people all over the world, actually contains? It's not entirely clear that Google knows or even could know if it wanted to.

Kids as content creators

We were surprised to find an unsettling number of videos featuring kids or apparently created by them. YouTube requires uploaders to be at least 13 years old, but we frequently saw children who appeared to be much younger than that, typically dancing, singing or playing video games.

In our preliminary research, our coders determined nearly a fifth of random videos with at least one person's face visible likely included someone under 13. We didn't take into account videos that were clearly shot with the consent of a parent or guardian.

Our current sample size of 250 is relatively small—we are working on coding a much larger sample—but the findings thus far are consistent with what we've seen in the past. We're not aiming to scold Google. Age validation on the internet is infamously difficult and fraught, and we have no way of determining whether these videos were uploaded with the consent of a parent or guardian. But we want to underscore what is being ingested by these large companies' AI models.

Small reach, big influence

It's tempting to assume OpenAI is using highly produced influencer videos or TV newscasts posted to the platform to train its models, but previous research on large language model training data shows that the most popular content is not always the most influential in training AI

models. A virtually unwatched conversation between three friends could have much more linguistic value in training a chatbot language model than a music video with millions of views.

Unfortunately, OpenAI and other AI companies are quite opaque about their training materials: They don't specify what goes in and what doesn't. Most of the time, researchers can infer problems with training data through biases in AI systems' output. But when we do get a glimpse at training data, there's often cause for concern. For example, Human Rights Watch released a report on June 10, 2024, that showed that a popular training dataset includes many photos of identifiable kids.

The history of big tech self-regulation is filled with moving goal posts. OpenAI in particular is notorious for asking for forgiveness rather than permission and has faced increasing criticism for putting profit over safety.

Concerns over the use of user-generated content for training AI models typically center on intellectual property, but there are also privacy issues. YouTube is a vast, unwieldy archive, impossible to fully review.

Models trained on a subset of professionally produced videos could conceivably be an AI company's first training corpus. But without strong policies in place, any company that ingests more than the popular tip of the iceberg is likely including content that violates the Federal Trade Commission's Children's Online Privacy Protection Rule, which prevents companies from collecting data from children under 13 without notice.

With last year's executive order on AI and at least one promising proposal on the table for comprehensive privacy legislation, there are signs that legal protections for user data in the U.S. might become more robust.

Have you unwittingly helped train ChatGPT?

The intentions of a YouTube uploader simply aren't as consistent or predictable as those of someone publishing a book, writing an article for a magazine or displaying a painting in a gallery. But even if YouTube's algorithm ignores your upload and it never gets more than a couple of views, it may be used to train models like ChatGPT and Gemini.

As far as AI is concerned, your family reunion video may be just as important as those uploaded by influencer giant Mr. Beast or CNN.

This article is republished from The Conversation under a Creative Commons license. Read the original article.

Provided by The Conversation

Citation: AI companies train language models on YouTube's archive—making family-and-friends videos a privacy risk (2024, June 27) retrieved 30 June 2024 from https://techxplore.com/news/2024-06-ai-companies-language-youtube-archive.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.