Companies gather massive databases of people’s images for facial recognition tools from OKCUPID

Cade Metz

a room with many items: In an undated handout image from the website Megapixels, a sample image from the Brainwash database, created by Stanford University researchers, which contains more than 10,000 images and nearly 82,000 annotated heads.
© Open Data Commons Public Domain Dedication and License/Megapixels via The New York Times In an undated handout image from the website Megapixels, a sample image from the Brainwash database, created by Stanford University researchers, which contains more than 10,000 images and nearly 82,000…

SAN FRANCISCO — Dozens of databases of people’s faces are being compiled without their knowledge by companies and researchers, with many of the images then being shared around the world, in what has become a vast ecosystem fueling the spread of facial recognition technology.

The databases are pulled together with images from social networks, photo websites, dating services like OkCupid, and cameras placed in restaurants and on college quads. Although

there is no precise count of the datasets, privacy activists have pinpointed repositories that were built by Microsoft, Stanford University, and others, with one holding more than 10 million images while another had more than 2 million.

The facial compilations are being driven by the race to create leading-edge facial recognition systems. This technology learns how to identify people by analyzing as many digital pictures as possible using “neural networks,” which are complex mathematical systems that require vast amounts of data to build pattern recognition.

Tech giants Facebook and Google have most likely amassed the largest face data sets, which they do not distribute, according to research papers. But other companies and universities have widely shared their image troves with researchers, governments, and private enterprises in Australia, China, India, Singapore, and Switzerland for training artificial intelligence, according to academics, activists, and public papers.

Companies and labs have gathered facial images for more than a decade, and the databases are merely one layer to building facial recognition technology. But people often have no idea that their faces ended up in them. And while names are typically not attached to the photos, individuals can be recognized because each face is unique to a person.

Questions about the datasets are rising because the technologies that they have enabled are being used in potentially invasive ways. Documents released last Sunday revealed that Immigration and Customs Enforcement officials employed facial recognition technology to scan motorists’ photos to identify unauthorized immigrants. The FBI also spent

more than a decade using such systems to compare driver's license and visa photos against the faces of suspected criminals, according to a Government Accountability Office report last month. On Wednesday, a congressional hearing tackled the government's use of the technology.

There is no oversight of the datasets. Activists and others said they were angered by the possibility that people's likenesses had been used to build ethically questionable technology and that the images could be misused. At least one facial database created in the United States was shared with a company in China that has been linked to ethnic profiling of the country's minority Uighur Muslims.

Over the past several weeks, some companies and universities, including Microsoft and Stanford, removed their facial datasets from the internet because of privacy concerns. But given that the images were already so well distributed, they are most likely still being used in the United States and elsewhere, researchers and activists said.

"You come to see that these practices are intrusive, and you realize that these companies are not respectful of privacy," said Liz O'Sullivan, who oversaw one of these databases at the artificial intelligence startup Clarifai. She said she left the New York-based company in January to protest such practices.

"The more ubiquitous facial recognition becomes, the more exposed we all are to being part of the process," she said.

Google, Facebook, and Microsoft declined to comment.

One database, which dates to 2014, was put together by researchers at Stanford. It was called Brainwash, after a San Francisco cafe of the same name, where the researchers tapped into a camera. Over three days, the camera took more than 10,000 images, which went into the database, the researchers wrote in a 2015 paper. The paper did not address whether cafe patrons knew their images were being taken and used for research. (The cafe has closed.)

The Stanford researchers then shared Brainwash. According to research papers, it was used in China by academics associated with the National University of Defense Technology and Megvii, an artificial intelligence company that The New York Times previously reported has provided surveillance technology for monitoring Uighurs.

The Brainwash dataset was removed from its original website last month after Adam Harvey, an activist in Germany who tracks the use of these repositories through a website called MegaPixels, drew attention to it. Links between Brainwash and papers describing work to build AI systems at the National University of Defense Technology in China have also been deleted, according to documentation from Harvey.

Stanford researchers who oversaw Brainwash did not respond to requests for comment. “As part of the research process, Stanford routinely makes research documentation and supporting materials available publicly,” a university official said. “Once research materials are made public, the university does not track their use nor did university officials.”

At Microsoft, researchers have claimed on the company’s website to have created one of the biggest facial datasets. The

collection, called MS Celeb, spanned over 10 million images of more than 100,000 people.

MS Celeb was ostensibly a database of celebrities, whose images are considered fair game because they are public figures. But MS Celeb also brought in photos of privacy and security activists, academics, and others, such as Shoshana Zuboff, the author of the book “The Age of Surveillance Capitalism,” according to documentation from Harvey of the MegaPixels project. MS Celeb was distributed internationally before being removed this spring after Harvey and others flagged it.

Matt Zeiler, founder and chief executive of Clarifai, the AI startup, said his company had built a facial database with images from OkCupid, a dating site. He said Clarifai had access to OkCupid’s photos because some of the dating site’s founders invested in his company.

He added that he had signed a deal with a large social media company — he declined to disclose which — to use its images in training facial recognition models. The social network’s terms of service allow for this kind of sharing, he said.

“There has to be some level of trust with tech companies like Clarifai to put powerful technology to good use and get comfortable with that,” he said.

An OkCupid spokeswoman said that Clarifai contacted the company in 2014 “about collaborating to determine if they could build unbiased AI and facial recognition technology” and that the dating site “did not enter into any commercial agreement then and have no relationship with them now.” She did not address

whether Clarifai had gained access to OkCupid's photos without its consent.

Clarifai used the images from OkCupid to build a service that could identify the age, sex, and race of detected faces, Zeiler said. The startup also began working on a tool to collect images from a website called Insecam — short for "insecure camera" — which taps into surveillance cameras in city centers and private spaces without authorization. Clarifai's project was shut down last year after some employees protested and before any images were gathered, he said.

Zeiler said Clarifai would sell its facial recognition technology to foreign governments, military operations, and police departments provided the circumstances were right. It did not make sense to place blanket restrictions on the sale of technology to entire countries, he added.

O'Sullivan, the former Clarifai technologist, has joined a civil rights and privacy group called the Surveillance Technology Oversight Project. She is now part of a team of researchers building a tool that will let people check whether their image is part of the openly shared facial databases.

"You are part of what made the system what it is," she said.