A Privacy Dustup at Microsoft Exposes Major Problems for A.I.

Dave Gershgorn

The most important datasets are rife with sexism and racism

Credit: Andrej Karpathy/Stanford

The results you get when you search for an image on Google have something in common with Siri's ability to listen to your commands. And they each share DNA with the facial recognition projects rolling out to the world's airports and beyond.

All of these features are fed by enormous piles of data. These datasets might contain thousands of pictures of faces or gigabytes of audio logs of human speech. They are the raw material used by nearly everyone who wants to work with A.I., and they don't come cheap. It takes expertise and investment to build them. That's why corporate and academic institutions are constructing their own datasets — and only sometimes share them with the world, creating what are known as open datasets.

But open doesn't automatically mean good — or ethical. Last week, Microsoft took down its MS Celeb facial recognition dataset following a report from MegaPixels, a watchdog project dedicated to uncovering how open datasets are used to build surveillance infrastructure across the world. MegaPixels showed that Microsoft's data included photographs not just of public people like celebrities, but of private citizens and journalists as well — and that it had been downloaded by private U.S. researchers and state-backed surveillance operations in China.

"There's clearly a large misalignment between what researchers and the general public think is acceptable," Adam Harvey, the creator of MegaPixels, tells *OneZero.*

MS Celeb was created for a competition in 2016. A.I. researchers used the dataset, which included 10 million images of celebrities collected from around the internet, to train their facial recognition algorithms, and then compete for the highest accuracy on a standardized set of face images. Following the competition, MS Celeb was made freely available online for anybody to download and use to train their own facial recognition algorithms. But no one realized that the dataset included images of private people — none of whom knew they were included in the data — until MegaPixels pointed it out.

This isn't just Microsoft's problem, however. While MS Celeb is under intense scrutiny now, other large datasets are also ripe for misuse. For example, a hiring algorithm trained on data that associates men with leadership positions because it's been given more data about men in senior roles will reinforce that unconscious societal bias. A facial recognition algorithm trained on faces with lighter skin will be significantly worse at identifying faces with darker skin, making it an unreliable and even dangerous tool for law enforcement and security.

Microsoft's MS Celeb isn't even the most important dataset out there. Perhaps the most widely-used image recognition dataset is ImageNet, which was created by Fei-Fei Li, a Stanford professor and former chief scientist of Google Cloud. In 2007, when the ImageNet project first started, the prevailing theory among computer scientists was that there was an undiscovered algorithm that would allow A.I. to learn like a human. Li had a different strategy. Rather than trying to perfect a core algorithm, Li focused on the data instead, giving algorithms more examples from which to extract patterns. She created ImageNet to feed millions of images to computer vision programs, and then started a competition to drive researchers to compete on image recognition accuracy. In 2012, a team from Toronto led by the renowned computer scientist Geoffrey Hinton blew away rivals using a fringe idea he had worked on for decades: an artificial neural network, which demonstrated that given enough data, A.I. could learn to identify the complex patterns of pixels that we use to represent objects in images.

Other datasets are tailored to specific use cases. In facial recognition, the most widely used dataset is called Labeled Faces in the Wild, created by the University of Massachusetts-Amherst to hone facial recognition's ability to identify people at different angles and in different lighting situations. Much like Microsoft's MS Celeb, it is mainly populated by photographs of actors, celebrities, and other public

figures.

The reason why these datasets exist in public in the first place is a vestige of A.I.'s origins in academia, and an oddity in the increasingly cutthroat A.I. development being carried out by big tech companies in the private sector. A.I. research — despite billions in corporate investment from companies like Google, Facebook, Microsoft, and Amazon — is still rooted in universities. Academics in the field have long tried to make computer science and the data that fuels it more accessible, last year boycotting a closed-access journal created by *Nature* because it put research behind a paywall. As a result, there are a number of free and publicly available datasets for any A.I. researcher to use.

But despite their welcome openness in an increasingly closed field, these datasets are problematic, just as MS Celeb was. ImageNet itself is built on a language dataset called WordNet, which was created by a group led by psychologist George Miller in the late 1980s as a way to organize words and ideas by arranging them in a hierarchy. For example, the word "chair" is categorized within the word "furniture," which is categorized within the category "artifact."

> "Many people have downloaded WordNet and made it their own. We cannot control what they do with it"

ImageNet uses these classifications to describe the pictures contained within its dataset. But WordNet, and thus ImageNet as well, harbor outdated racial language and stereotypes. The category for black people of African descent, which contains 1,404 images, includes words like "Black, Black person, blackamoor, Negro, Negroid." Some words inside the dataset are outright racial slurs. Similar language is found in nearly every racial or sexual classification.

"WordNet was created starting some 30 years ago, and I'm afraid this entry made its way into the database and was never edited," says Christiane Fellbaum, a professor of computer science at Princeton who now maintains WordNet, in an email. "I must add that many people have downloaded WordNet and made it their own. We cannot control what they do with it, so this unfortunate entry may live on in other dictionaries (though I hope this isn't the case)."

Following *OneZero*'s query, Fellbaum said she would update the dataset to remove these words. She added that WordNet was assembled by multiple people, and relied on many dictionaries, though one used was the American Heritage Dictionary. Searches of the online American Heritage Dictionary show that many of the words in WordNet, like "jezebel" or "negroid," do indeed exist within the dictionary.

You can actually see which ImageNet category a computer vision algorithm would place you in by using an online tool called ImageNet Roulette that matches a portrait or other image to the closest image in the ImageNet dataset. It's less than perfect. Sarah Myers West, a postdoctoral researcher at the A.I. Now Institute, says that when she submitted her Twitter bio photo to the ImageNet Roulette, it returned "hussy" and "jezebel."

(My pictures returned "beard," described here as a woman who camouflages a homosexual man's sexuality, and "kisser," described as someone who, yes, kisses. My male editor was categorized as a "sister," described as a nun.)

Looking at the latest version of Labeled Faces in the Wild through a similar lens, the dataset shows itself to be overwhelmingly white and male. A 2014 study which attempted to automate the categorization of faces into a gender binary to test a gender recognition algorithm found 11,590 males and 4,109 females within the dataset. The breakdown of race into three arbitrary categories — “white,” “darker skin,” and “Asian” — was even starker, with 12,373 images of white faces, 1,145 images of faces with darker skin, and 2,166 images of Asian faces. While the authors of the dataset haven’t explained how the specific images were chosen, they have said that the images were taken from news articles.

In a report published by the A.I. Now Institute, however, Myers noted that the makeup of the Labeled Faces in the Wild dataset reflected social values that were themselves often biased. “The news media at the time predominantly featured white men in positions of celebrity and power,” she wrote. “Drawing from this source, [Labeled Faces in the Wild’s] representation of ‘human faces’ can be understood as a reflection of early 2000s social hierarchy, as reproduced through visual media.”

As A.I. spreads throughout society, the makeup of these datasets will become increasingly important. Artificial intelligence algorithms have been shown to directly reflect the human biases encoded in their data. Some job recommendation products, for example, use algorithms that might suggest women should be selected for certain jobs, like nurses, while men should be selected for other jobs, like doctors or managers. Those algorithms have been trained on historical social data, effectively locking in outdated gender stereotypes.

And unbalanced datasets have already been shown to result in biased and error-prone facial recognition algorithms. Such algorithms trained by large tech companies were a third more accurate on white male faces than darker skinned female faces, according to Gender Shades, a project from MIT

Media Lab researcher Joy Buolamwini.

Once these datasets are released to the public, it's difficult — if not impossible — to fix such problems. Adam Harvey, the researcher who created MegaPixels, says that even though Microsoft deleted the dataset from their own website, it hasn't disappeared from the internet. The entire dataset is still available online as an unauthorized torrent, where it's been downloaded more than 50 times over the last week.

These datasets are often the lowest common denominator of data science, scraping the internet for as much data as possible and then packaging it as a tool for machine learning. Because they're so large — MS Celeb is 250 gigabytes — the people using them may not know every piece of data they contain.

At the same time tech companies that can afford to pay more rarely rely on open and free sources. Many maintain their own internal datasets, like Google's JFT-300M, which consists of 300 million images and is not available to the public. This dataset could be harnessed for anything Google uses computer vision algorithms on, like Google Photos' image labeling system, which in 2015 categorized two black Google Photos users as gorillas. The company's fix was to remove the "gorilla" classification from Google Photos altogether. There's a good chance these more recent datasets are also biased, but because they're locked away from researchers, no one can say for sure.

Garbage in, garbage out has been a core truth of computer programming since the field's earliest days, and experts are realizing that steps need to be taken to reverse the corrosive effect of cheap datasets. The EU is making large, accessible datasets a major aim of its A.I. policy, and is building "data spaces" to create better datasets that can be harnessed by researchers and companies while still respecting user privacy. The state of Illinois has specifically targeting facial recognition and biometric data with its Biometric Information Privacy Act, which was signed into law in 2008. The law requires that private companies obtain consent from Illinois citizens before adding them to any kind of biometrics database. Facebook and Google have both been sued under this law, which was also recently challenged by Six Flags after the amusement park company allegedly fingerprinted a minor without parental consent.

Wilma Bainbridge, co-creator of the 10k US Adult Faces Database, has taken a different approach to try to ensure that her dataset won't be used maliciously or in a way that would violate someone's privacy. With the advice of a lawyer, she manually approves or denies applications to download the dataset, injecting an element of human decision-making into a process that often unthinkingly automated.

We need to figure this out because datasets and their value are only growing. They are commonly accepted as test beds benefiting both tech companies and independent researchers; they serve as commonly shared benchmarks that allow computer scientists to understand how far technology has progressed.

But datasets based on random pictures from the internet are past their prime. The value extracted from computer vision algorithms by companies like Microsoft is immense while providing no return to those people whose faces have trained its algorithms, even without their knowledge. The backlash over MS Celeb might signal that people are finally learning what their faces might be worth to the trillion-dollar companies using that data to advance the state of the art in A.I..