WIRED

The Viral App That Labels You Isn't Quite What You Think

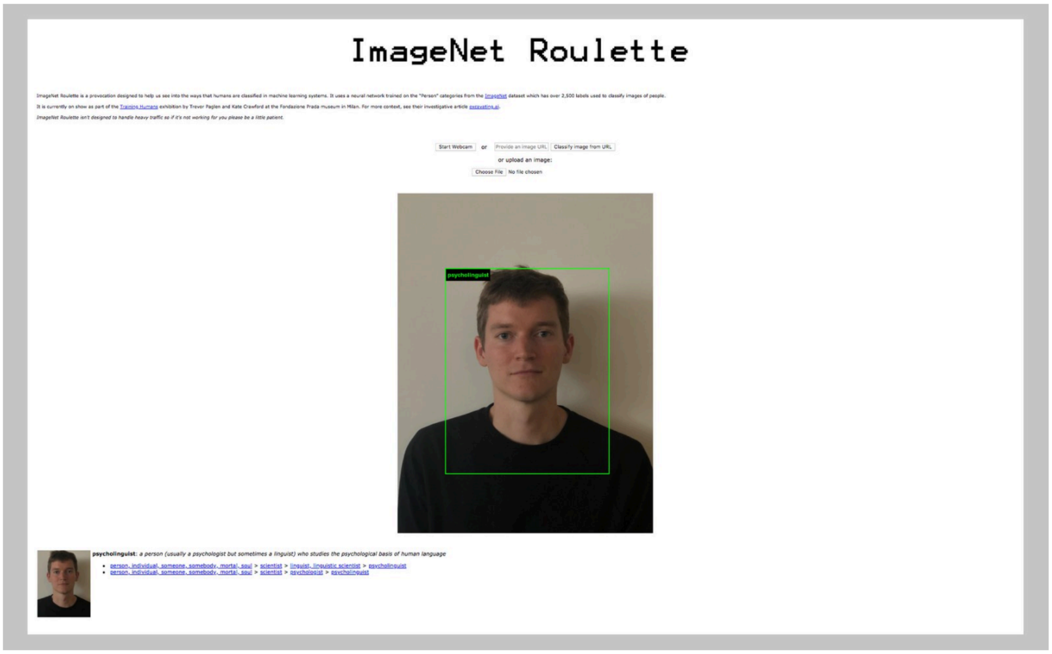
ImageNet Roulette reveals biases in artificial intelligence algorithms. But the vast majority of tags attached to people are rarely used.

GREGORY BARBER
Business
09.19.2019

BILLY H.C. KWOK/GETTY IMAGES

This week, the denizens of Twitter began posting photos of themselves with an odd array of labels. Some, like “face,” were confusingly benign, while others appeared to verify harder truths: Your humble writer was declared a cipher, a nobody, “a person of no influence.” Fair enough. But many of the labels were more troubling. There were rape suspects and debtors. A person would be labeled not just black, but “negro” and “negroid.”

The project, called ImageNet Roulette, is an effort by artist Trevor Paglen and researcher Kate Crawford to illustrate the dangers of feeding flawed data into artificial intelligence. It takes aim at one of the field’s seminal resources: ImageNet, the database of 14 million images that’s credited with unlocking the potential of deep learning, the technique used for everything from self-driving cars to facial recognition. The algorithm behind the Roulette tool is trained using images within ImageNet that label people across 2,395 categories, from “slatterns” to “Uzbeks.” “I wanted to crack ImageNet open and look at images that weren’t meant to be looked at,” says Paglen. The experiment, now viral, has plenty of people asking just how those labels got there in the first place, and why they remain.



ImageNet Roulette

ImageNet Roulette is a provocation designed to help us see into the ways that humans are classified in machine learning systems. It uses a neural network trained on the "Person" categories from the ImageNet dataset which has over 2,500 labels used to classify images of people.

It is currently on show as part of the Training Humans exhibition by Trevor Paglen and Kate Crawford at the Fondazione Prada museum in Milan. For more context, see their investigative article excavating.ai.

ImageNet Roulette isn't designed to handle heavy traffic so if it's not working for you please be a little patient.

psycholinguist: *a person (usually a psychologist but sometimes a linguist) who studies the psychological basis of human language*

- person, individual, someone, somebody, mortal, soul > scientist > linguist, linguistic scientist > psycholinguist
- person, individual, someone, somebody, mortal, soul > scientist > psychologist > psycholinguist

ImageNet labeled the author a "psycholinguist."

SCREENSHOT: GREGORY BARBER VIA IMAGENET ROULETTE

The answers stem from the swift evolution of AI from a juvenile science to everyday tool, and the burying of potential bias in reams of data. Recently the problem has started receiving attention from those within the field. That includes ImageNet's creators, who say they are well aware of the flaws in their database and have been working to fix problems within the "person" labels over the past year. They point out that the images of people are rarely used by researchers; all the same, the creators say they've been in the process of "debiasing" the data set.

That effort included removing most of the 14 million images from Stanford servers in January while the team reviewed categories deemed offensive and how to make the distribution of images more diverse. The team also plans to eliminate categories they consider "nonvisual," because how else does an algorithm identify someone as a

“Bahamian” or a “debtor” if not by some kind of contextual cheat or built-in bias? They submitted a publication describing their methods for peer review in August.

Still, the problems with ImageNet illustrate how bias can propagate from mostly forgotten sources. In this case, the source starts in the mid-1980s, with a project at Princeton called WordNet. WordNet was an effort by psychologists and linguists to provide a “conceptual dictionary,” where words were organized into hierarchies of related meaning. You might travel from animals to vertebrates to dogs to huskies, for example, or perhaps branch off along the way into cats and tabbies. The database goes beyond the pages of Merriam-Webster, including everything from obscure desserts to outdated slang. “A lot of of the terms that were considered socially appropriate then are totally inappropriate now,” says Alexander Wong, a professor of computer science at the University of Waterloo.

The latest on artificial intelligence, from machine learning to computer vision and more

In 2009, the creators of ImageNet, who include Fei-Fei Li and Kai Li, set out to create a similar hierarchy for images, believing it could be a useful tool for teaching AI how to identify and categorize objects. Their ambitions were grand: to create a visual library of nouns, using WordNet as a handy template. But annotating images was time-consuming and expensive, especially when it involved paying Princeton undergrads to do it. The process eventually scaled up with the help of annotators crowdsourced on Amazon’s Mechanical Turk, who would identify objects in images and remove bad matches.

The ImageNet researchers attribute the inclusion of offensive and insensitive categories to the overall size of the task, which ultimately involved 50,000 workers who evaluated 160 million candidate images. They also point out that only a fraction of the “person” images were actually used in practice. That’s because references to ImageNet typically mean a smaller version of the data set used in the ImageNet Challenge, a competition among research teams to build AI that detects and classifies objects in the images. Out of the 20,000 or so classes of objects, the competition was limited to 1,000, representing just over a million images. Only three “person” categories—scuba diver, groom, and baseball player—were included. The best models trained using that limited version are typically the ones used in other research and real-world applications.

Paglen says the debiasing effort is a positive step, but he finds it revealing that the data apparently went unexamined for 10 years. “The people building these data sets seem to have had no idea what’s in them,” he says. (The ImageNet team says the debiasing project is part of an “ongoing” effort to make machine learning more fair.)

Wong, the Waterloo professor, who has studied biases within ImageNet, says the inattention was likely in part because, at the time the database was made, researchers were focused on the basics of getting their object detection algorithms to work. The enormous success of deep learning took the field by surprise. “We’re now getting to a point where AI is usable, and now people are looking at the social ramifications,” he says.

The ImageNet creators acknowledge that their initial attempts at quality control were ineffective. The full data set persisted online until January, when the researchers removed all but the ImageNet Challenge images. The new release will include fewer than half of the original person images. It will also allow users to flag additional images and categories as offensive, an acknowledgement that “offensiveness is subject and also constantly evolving,” the ImageNet team writes.

“The people building these data sets seem to have had no idea what’s in them.”

TREVOR PAGLEN

The removal of images has itself proved controversial. “I was surprised a large chunk of the data just disappeared in January without anybody saying anything,” Paglen says. “This is a historically important database.” He points out that the data is likely still in the wild, downloaded on various servers and home computers; removing the data from an accessible home only makes biases more difficult to reproduce and study, he says.

Even researchers were surprised to find out that the data was removed as part of a debiasing project. Chris Dulhanty, one of Wong’s graduate students, says he had reached out to the ImageNet team to request data earlier this year but didn’t hear back. He assumed removal had to do with technical issues on the aging ImageNet site. (The ImageNet team did not respond to questions about the decision to remove the data but said they would discuss with other researchers the possibility of making it available again.)

In a paper accompanying ImageNet Roulette, Paglen and Crawford liken the removal of images from ImageNet to similar moves by other institutions. In June, for example, Microsoft removed its “MS-Celeb” database after a *Financial Times* investigation.

The ImageNet debiasing effort is a good start, Wong says. But he hopes the team will make good on plans to look at bias beyond the person categories. About 15 percent of the “nonperson” images do, in fact, contain people somewhere in the frame, he notes. That could lead to inadvertent associations—say, between black people and the label “basketball,” as one research team noted, or between objects related to computers and people who are young, white, and male. Those biases are more likely embedded in widely used models than in any of those contained in the “person” labels.

Paglen says that attempts to debias may be futile. “There’s no such thing as a neutral way of organizing information,” he says. He and Crawford point to other more recent data sets that have attempted a more nuanced approach to sensitive labels. He points to an

IBM attempt to bring more “diversity” to faces by measuring facial dimensions. The authors hope it’s an improvement over human judgments but note it raises new questions. Is skin tone a better measure? The answers will reflect evolving social values. “Any system of classification is going to be of its moment in time,” he says. Paglen is opening an exhibition in London next week that intends to illustrate AI’s blind ignorance in that area. It begins with a Magritte painting of an apple, labeled “Ceci n’est pas une pomme.” Good luck convincing an AI algorithm of that.