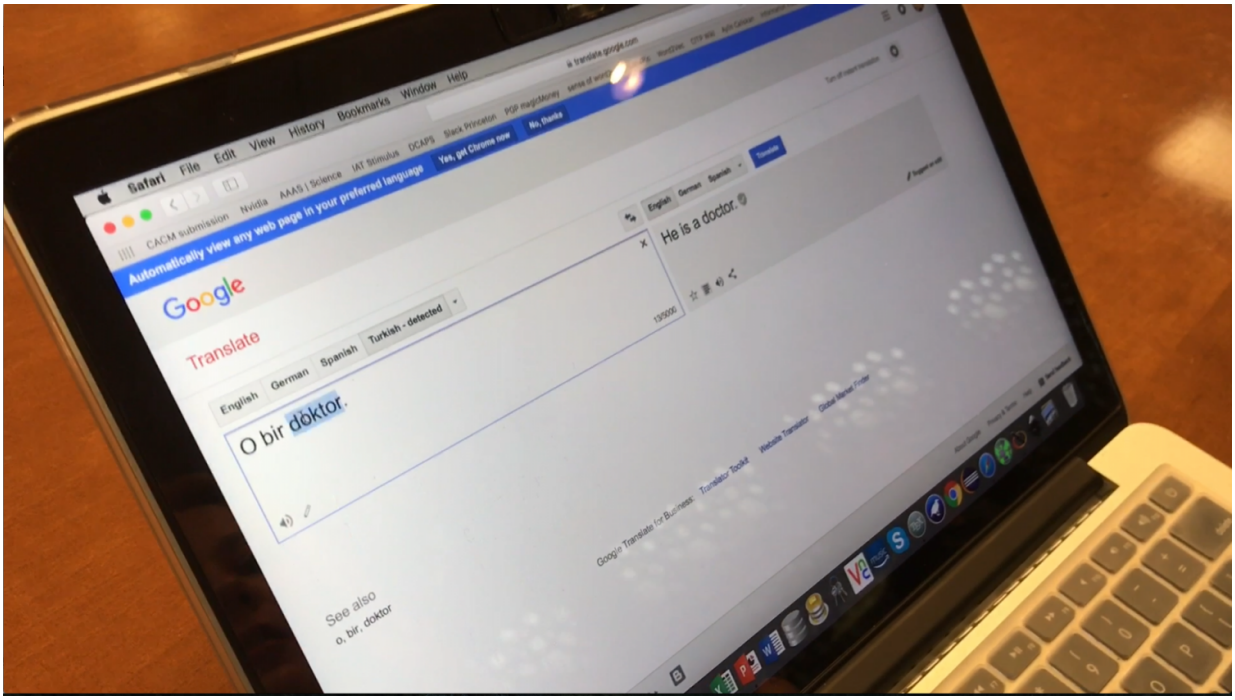
Biased bots: Human prejudices sneak into artificial intelligence systems

April 13 2017



Researchers found that certain search terms revealed AI bias. Credit: Princeton University

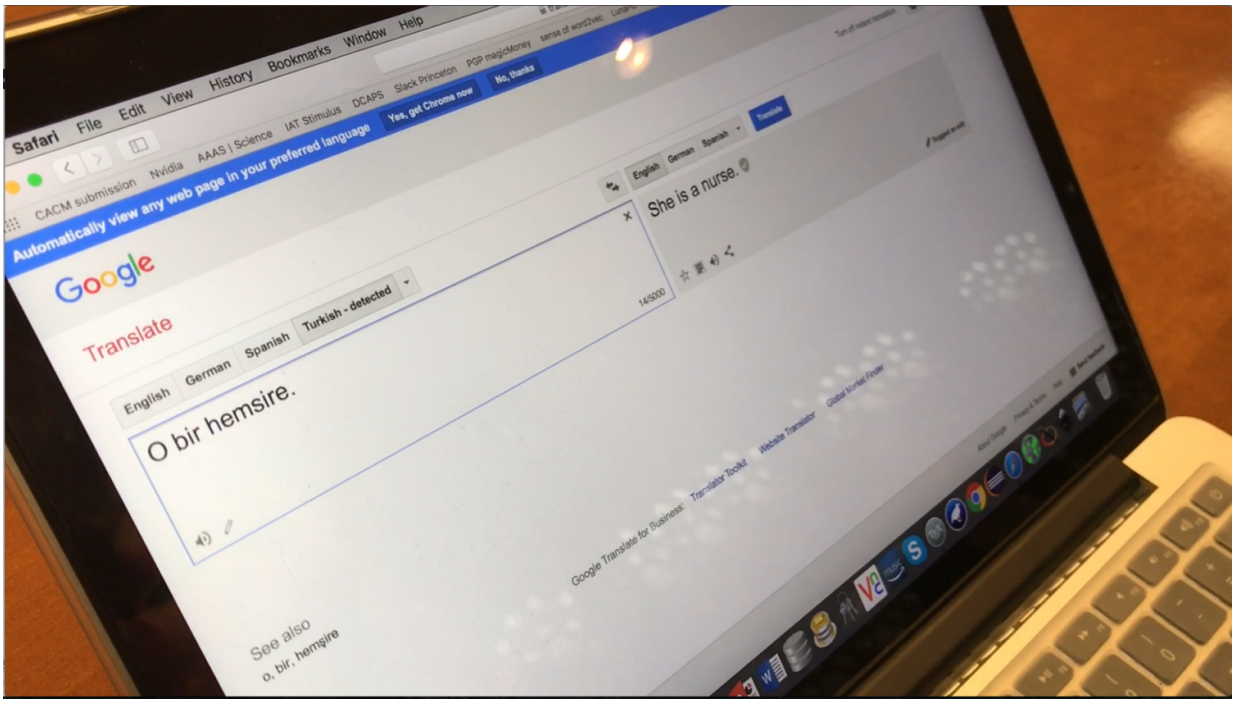
In debates over the future of artificial intelligence, many experts think of the new systems as coldly logical and objectively rational. But in a new study, researchers have demonstrated how machines can be reflections of us, their creators, in potentially problematic ways. Common machine learning programs, when trained with ordinary human language available

online, can acquire cultural biases embedded in the patterns of wording, the researchers found. These biases range from the morally neutral, like a preference for flowers over insects, to the objectionable views of race and gender.

Identifying and addressing possible bias in machine learning will be critically important as we increasingly turn to computers for processing the natural language humans use to communicate, for instance in doing online text searches, image categorization and automated translations.

"Questions about fairness and bias in machine learning are tremendously important for our society," said researcher Arvind Narayanan, an assistant professor of computer science and an affiliated faculty member at the Center for Information Technology Policy (CITP) at Princeton University, as well as an affiliate scholar at Stanford Law School's Center for Internet and Society. "We have a situation where these artificial intelligence systems may be perpetuating historical patterns of bias that we might find socially unacceptable and which we might be trying to move away from."

The paper, "Semantics derived automatically from language corpora contain human-like biases," published April 14 in *Science*. Its lead author is Aylin Caliskan, a postdoctoral research associate and a CITP fellow at Princeton; Joanna Bryson, a reader at University of Bath, and CITP affiliate, is a coauthor.



Researchers found that searches revealed hidden AI bias. Credit: Princeton University

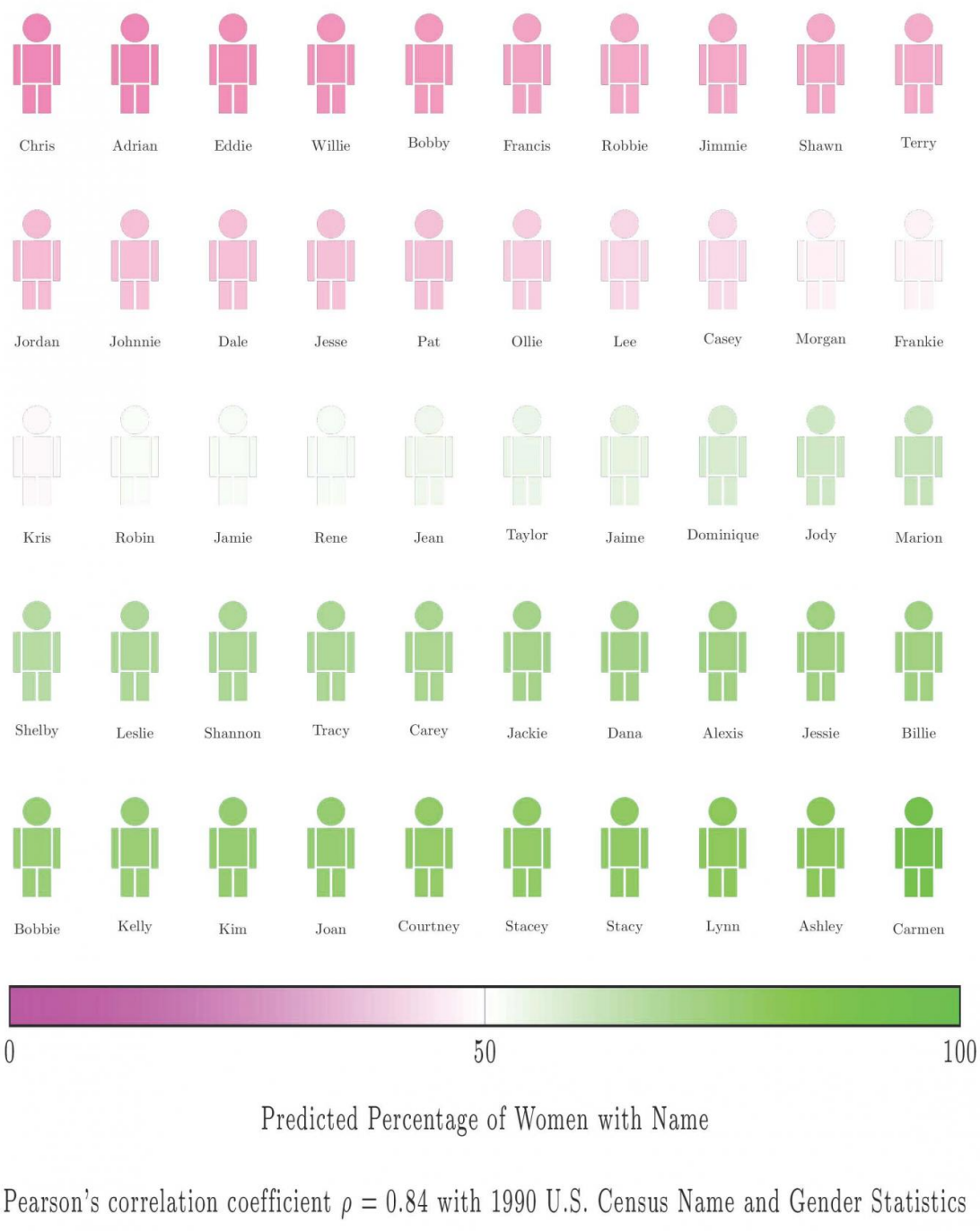
As a touchstone for documented human biases, the study turned to the Implicit Association Test, used in numerous social psychology studies since its development at the University of Washington in the late 1990s. The test measures response times (in milliseconds) by human subjects asked to pair word concepts displayed on a computer screen. Response times are far shorter, the Implicit Association Test has repeatedly shown, when subjects are asked to pair two concepts they find similar, versus two concepts they find dissimilar.

Take flower types, like "rose" and "daisy," and insects like "ant" and "moth." These words can be paired with pleasant concepts, like "caress" and "love," or unpleasant notions, like "filth" and "ugly." People more quickly associate the flower words with pleasant concepts, and the insect

terms with unpleasant ideas.

The Princeton team devised an experiment with a program where it essentially functioned like a machine learning version of the Implicit Association Test. Called GloVe, and developed by Stanford University researchers, the popular, open-source program is of the sort that a startup machine learning company might use at the heart of its product. The GloVe algorithm can represent the co-occurrence statistics of words in, say, a 10-word window of text. Words that often appear near one another have a stronger association than those words that seldom do.

The Stanford researchers turned GloVe loose on a huge trawl of contents from the World Wide Web, containing 840 billion words. Within this large sample of written human culture, Narayanan and colleagues then examined sets of so-called target words, like "programmer, engineer, scientist" and "nurse, teacher, librarian" alongside two sets of attribute words, such as "man, male" and "woman, female," looking for evidence of the kinds of biases humans can unwittingly possess.



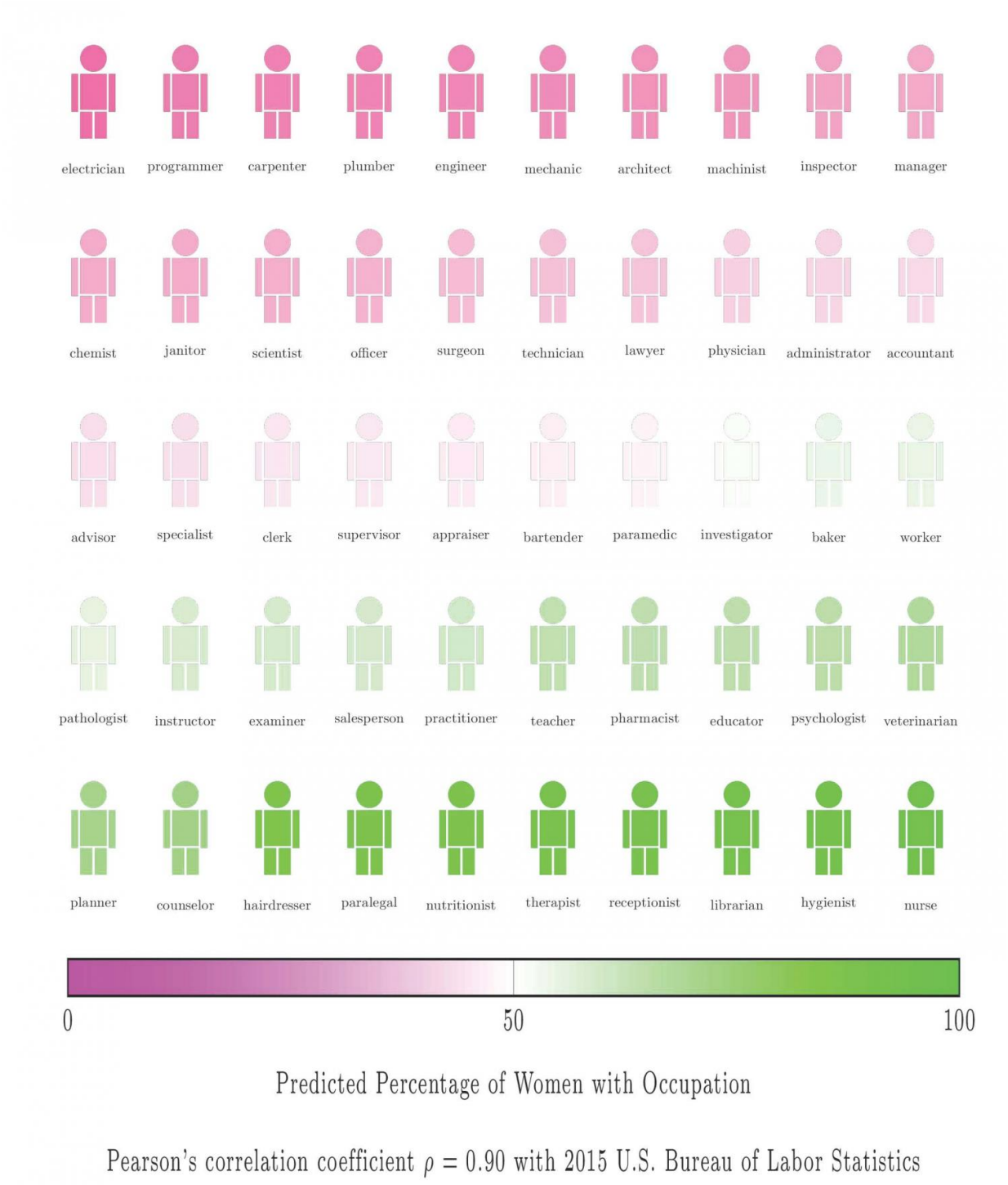
Predicted percentage of women with a certain name. Credit: Aylin Caliskan

In the results, innocent, inoffensive biases, like for flowers over bugs, showed up, but so did examples along lines of gender and race. As it turned out, the Princeton machine learning experiment managed to replicate the broad substantiations of bias found in select Implicit Association Test studies over the years that have relied on live, human subjects.

For instance, the machine learning program associated female names more with familial attribute words, like "parents" and "wedding," than male names. In turn, male names had stronger associations with career attributes, like "professional" and "salary." Of course, results such as these are often just objective reflections of the true, unequal distributions of occupation types with respect to gender—like how 77 percent of computer programmers are male, according to the U.S. Bureau of Labor Statistics.

Yet this correctly distinguished bias about occupations can end up having pernicious, sexist effects. An example: when foreign languages are naively processed by machine learning programs, leading to gender-stereotyped sentences. The Turkish language uses a gender-neutral, third person pronoun, "o." Plugged into the well-known, online translation service Google Translate, however, the Turkish sentences "o bir doktor" and "o bir hem?ire" with this gender-neutral pronoun are translated into English as "he is a doctor" and "she is a nurse."

"This paper reiterates the important point that machine learning methods are not 'objective' or 'unbiased' just because they rely on mathematics and algorithms," said Hanna Wallach, a senior researcher at Microsoft Research New York City, who was not involved in the study. "Rather, as long as they are trained using data from society and as long as society exhibits biases, these methods will likely reproduce these biases."



Predicted percentage of women with a certain occupation. Credit: Aylin Caliskan

Another objectionable example harkens back to a well-known 2004 paper by Marianne Bertrand of the University of Chicago Booth School of Business and Sendhil Mullainathan of Harvard University. The economists sent out close to 5,000 identical resumes to 1,300 job advertisements, changing only the applicants' names to be either traditionally European American or African American. The former group was 50 percent more likely to be offered an interview than the latter. In an apparent corroboration of this bias, the new Princeton study demonstrated that a set of African American names had more unpleasantness associations than a European American set.

Computer programmers might hope to prevent cultural stereotype perpetuation through the development of explicit, mathematics-based instructions for the machine learning programs underlying AI systems. Not unlike how parents and mentors try to instill concepts of fairness and equality in children and students, coders could endeavor to make machines reflect the better angels of human nature.

"The biases that we studied in the paper are easy to overlook when designers are creating systems," said Narayanan. "The biases and stereotypes in our society reflected in our language are complex and longstanding. Rather than trying to sanitize or eliminate them, we should treat biases as part of the language and establish an explicit way in machine learning of determining what we consider acceptable and unacceptable."

More information: "Semantics derived automatically from language corpora contain human-like biases," *Science* (2017). science.sciencemag.org/cgi/doi … 1126/science.aal4230

Provided by Princeton University