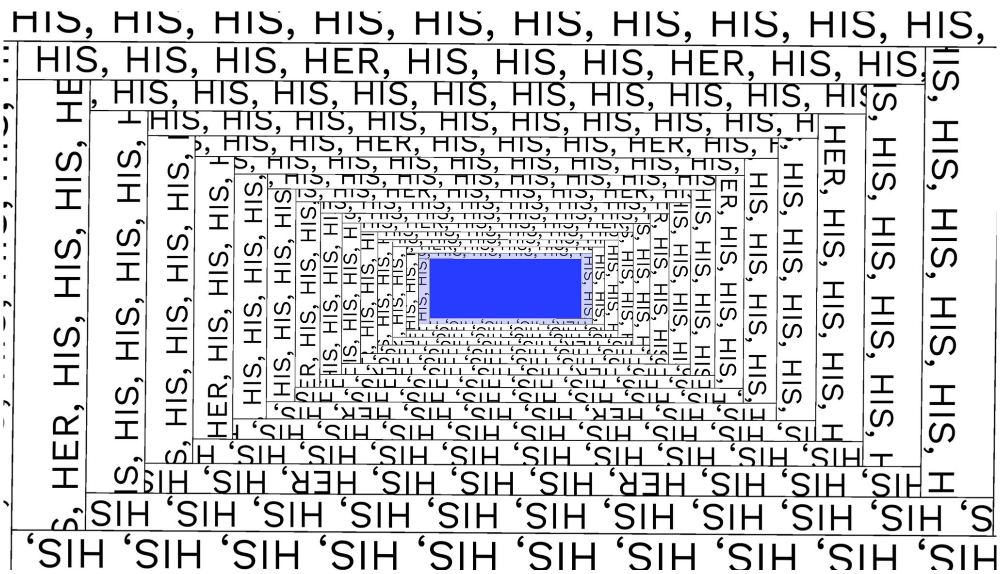
We Teach A.I. Systems Everything, Including Our Biases (Published 2019)

Researchers say computer systems are learning from lots and lots of digitized books and news articles that could bake old attitudes into new technology.

- Give this article



Credit Credit... By Kiel Mutschelknaus

Nov. 11, 2019

SAN FRANCISCO — Last fall, Google unveiled a breakthrough artificial intelligence technology called BERT that changed the way scientists build [systems that learn how people write and talk.]

But BERT, which is now being deployed in services like Google's internet search engine, has a problem: It could be picking up on biases in the way a child mimics the bad behavior of his parents.

BERT is one of a number of A.I. systems that learn from lots and lots of digitized information, as varied as old books, Wikipedia entries and

news articles. Decades and even centuries of biases — along with a few new ones — are probably baked into all that material.

BERT and its peers are more likely to associate men with computer programming, for example, and generally don't give women enough credit. One program decided almost everything written about President Trump was negative, even if the actual content was flattering.

As new, more complex A.I. moves into an increasingly wide array of products, like online ad services and business software or talking digital assistants like Apple's Siri and Amazon's Alexa, tech companies will be pressured to guard against the unexpected biases that are being discovered.

But scientists are still learning how technology like BERT, called "universal language models," works. And they are often surprised by the mistakes their new A.I. is making.

On a recent afternoon in San Francisco, while researching a book on artificial intelligence, the computer scientist Robert Munro fed 100 English words into BERT: "jewelry," "baby," "horses," "house," "money," "action." In 99 cases out of 100, BERT was more likely to associate the words with men rather than women. The word "mom" was the outlier.

"This is the same historical inequity we have always seen," said Dr. Munro, who has a Ph.D. in computational linguistics and previously oversaw natural language and translation technology at Amazon Web Services. "Now, with something like BERT, this bias can continue to perpetuate."

“This is the same historical inequity we have always seen,” said the computer scientist Robert Munro. Credit... Cayce Clifford for The New York Times

In a blog post this week, Dr. Munro also describes how he examined cloud-computing services from Google and Amazon Web Services that help other businesses add language skills into new applications. Both services failed to recognize the word “hers” as a pronoun, though they correctly identified “his.”

“We are aware of the issue and are taking the necessary steps to address and resolve it,” a Google spokesman said. “Mitigating bias from our systems is one of our A.I. principles, and is a top priority.” Amazon, in a statement, said it “dedicates significant resources to ensuring our technology is highly accurate and reduces bias, including rigorous benchmarking, testing and investing in diverse training data.”

Researchers have long warned of bias in A.I. that learns from large amounts data, including the facial recognition systems that are used by police departments and other government agencies as well as popular internet services from tech giants like Google and Facebook.

In 2015, for example, the Google Photos app was caught labeling African-Americans as “gorillas.” The services Dr. Munro scrutinized also showed bias against women and people of color.

BERT and similar systems are far more complex — too complex for anyone to predict what they will ultimately do.

“Even the people building these systems don’t understand how they are behaving,” said Emily Bender, a professor at the University of Washington who specializes in computational linguistics.

BERT is one of many universal language models used in industry and academia. Others are called ELMO, ERNIE and GPT-2. As a kind of inside joke among A.I. researchers, they are often named for Sesame Street characters. (BERT is short for Bidirectional Encoder Representations from Transformers.)

They learn the nuances of language by analyzing enormous amounts of text. A system built by OpenAI, an artificial intelligence lab in San Francisco, analyzed thousands of self-published books, including romance novels, mysteries and science fiction. BERT analyzed the same library of books along with thousands of Wikipedia articles.

In analyzing all this text, each system learned a specific task. OpenAI’s system learned to predict the next word in a sentence. BERT learned to identify the missing word in a sentence (such as “I want to ____ that car because it is cheap”).

Through learning these tasks, BERT comes to understand in a general way how people put words together. Then it can learn other tasks by analyzing more data. As a result, it allows A.I. applications to improve at a rate not previously possible.

“BERT completely changed everything,” said John Bohannon, director of science at Primer, a start-up in San Francisco that specializes in natural language technologies. “You can teach one pony all the tricks.”

Google itself has used BERT to improve its search engine. Before, if you typed “Do estheticians stand a lot at work?” into the Google search engine, it did not quite understand what you were asking. Words like “stand” and “work” can have multiple meanings, serving either as nouns or verbs. But now, thanks to BERT, Google correctly responds to the same question with a link describing the physical demands of life in the skin care industry.

But tools like BERT pick up bias, according to a recent research paper from a team of computer scientists at Carnegie Mellon University. The paper showed, for instance, that BERT is more likely to associate the word “programmer” with men than with women. Language bias can be a particularly difficult problem in conversational systems.

As these new technologies proliferate, biases can appear almost anywhere. At Primer, Dr. Bohannon and his engineers recently used BERT to build a system that lets businesses automatically judge the sentiment of headlines, tweets and other streams of online media. Businesses use such tools to inform stock trades and other pointed decisions.

But after training his tool, Dr. Bohannon noticed a consistent bias. If a tweet or headline contained the word “Trump,” the tool almost always judged it to be negative, no matter how positive the sentiment.

“This is hard. You need a lot of time and care,” he said. “We found an obvious bias. But how many others are in there?”

Dr. Bohannon said computer scientists must develop the skills of a biologist. Much as a biologist strives to understand how a cell works, software engineers must find ways of understanding systems like BERT.

In unveiling the new version of its search engine last month, Google executives acknowledged this phenomenon. And they said they

tested their systems extensively with an eye toward removing any bias.

Researchers are only beginning to understand the effects of bias in systems like BERT. But as Dr. Munro showed, companies are already slow to notice even obvious bias in their systems. After Dr. Munro pointed out the problem, Amazon corrected it. Google said it was working to fix the issue.

Primer's chief executive, Sean Gourley, said vetting the behavior of this new technology would become so important, it will spawn a whole new industry, where companies pay specialists to audit their algorithms for all kinds of bias and other unexpected behavior.

"This is probably a billion-dollar industry," he said.

More on A.I. and bias: