In 2016, researchers from Boston University and Microsoft were working on artificial intelligence algorithms when they discovered racist and sexist tendencies in the technology underlying some of the most popular and critical services we use every day. The revelation went against the conventional wisdom that artificial intelligence doesn't suffer from the gender, racial, and cultural prejudices that we humans do.

The researchers made this discovery while studying word-embedding algorithms, a type of AI that finds correlations and associations among different words by analyzing large bodies of text. For instance, a trained word-embedding algorithm can understand that words for flowers are closely related to pleasant feelings. On a more practical level, word embedding understands that the term "computer programming" is closely related to "C++," "JavaScript" and "object-oriented analysis and design." When integrated in a resume-scanning application, this functionality lets employers find qualified candidates with less effort. In search engines, it can provide better results by bringing up content that's semantically related to the search term.

The BU and Microsoft researchers found that the word-embedding algorithms had problematic biases, though—such as associating "computer programmer" with male pronouns and "homemaker" with female ones. Their findings, which they published in a research paper aptly titled "Man is to Computer Programmer as Woman is to Homemaker?" was one of several reports to debunk the myth of AI neutrality and to shed light on algorithmic bias, a phenomenon that is reaching critical dimensions as algorithms become increasingly involved in our everyday decisions.

THE ORIGINS OF ALGORITHMIC BIAS

Machine-learning and deep-learning algorithms underlie most contemporary AI-powered software. In contrast to traditional software, which works based on predefined and verifiable rules, deep learning creates its own rules and learns by example. For example, to create an image-recognition application based on deep learning, programmers "train" the algorithm by feeding it labeled data: in this case, photos tagged with the name of the object they contain. Once the algorithm ingests enough examples, it can glean common patterns among similarly labeled data and use that information to classify unlabeled samples.

This mechanism enables deep learning to perform many tasks that were virtually impossible with rule-based software. But it also means deep-learning software can inherit covert or overt biases.

"AI algorithms are not inherently biased," says Professor Venkatesh Saligrama, who teaches at Boston University's Department of Electrical and Computer Engineering and worked on the word-embedding algorithms. "They have deterministic functionality and will pick up any tendencies that already exist in the data they train on."

The word-embedding algorithms tested by the Boston University researchers were trained on hundreds of thousands of articles from Google News, Wikipedia, and other sources in which social biases are deeply embedded. As an example, because of the bro culture dominating the tech industry, male names come up more often with tech-related jobs—and that leads algorithms to associate men with jobs such as programming and software engineering.

"Algorithms don't have the power of the human mind in distinguishing right from wrong," adds Tolga Bolukbasi, a final-year PhD student at BU. Humans can judge the morality of our actions, even when we decide to act against ethical norms. But for algorithms, data is the ultimate determining factor.

Saligrama and Bolukbasi weren't the first to raise the alarm about this bias. Researchers at IBM, Microsoft, and the University of Toronto underlined the need to prevent algorithmic discrimination in a paper published in 2011. Back then, algorithmic bias was an esoteric concern, and deep learning still hadn't found its way into the mainstream. Today, though, algorithmic bias affects many of the things we do, such as reading news, finding friends, shopping online, and watching videos on Netflix and YouTube.

THE IMPACT OF ALGORITHMIC BIAS

In 2015, Google had to apologize after the algorithms powering its Photos app tagged two black people as gorillas—perhaps because its training dataset did not have enough pictures of black people. In 2016, of the 44 winners of a beauty contest judged by AI, nearly all were white, a few were Asian and only one had dark skin. Again, the reason was that the algorithm was mostly trained with photos of white people.

More recently, a test of IBM and Microsoft's face-analysis services found the companies' algorithms were nearly flawless at detecting the gender of men with light skin but often erred when presented with pictures of women with dark skin.

While these incidents likely caused negligible damage, the same can't be said of AI algorithms in more critical domains, such as healthcare, law enforcement, and recruitment. In 2016, an investigation by ProPublica found that COMPAS—AI-driven software that assesses the risk of recidivism in offenders—was biased against people of color. The discovery was especially concerning because judges in some states use COMPAS to determine who walks free and who stays in jail.

In another case, a study of Google's advertising platform, which is powered by deep-learning algorithms, found that men were shown ads for high-paying jobs more often than women were. A separate study found a similar issue with LinkedIn's job ads. Yet another showed that biased hiring algorithms were 50 percent more likely to send an interview invitation to a person whose name was European American than to someone with an African-American name.

Areas such as loan approval, credit rating, and scholarship face similar threats.

Algorithmic bias is further worrying because of how it might amplify social biases. Under the illusion that AI is cold, mathematical calculation devoid of prejudice or bias, humans may tend to trust algorithmic judgment without questioning it.

In an interview with *Wired UK,* Edinburgh Napier University criminology lecturer Andrew Wooff observed that the "time-pressured, resource intensive" world of policing could cause law enforcement officers to rely too much on algorithmic decisions.

"I can imagine a situation where a police officer may rely more on the [AI-driven] system than their own decision-making processes," he said. "Partly that might be so that you can justify a decision when something goes wrong."

Relying on biased algorithms creates a feedback loop: We make decisions that create more biased data. Algorithms will then analyze and train on that biased data in the future.

This kind of thing is already happening on social media networks, including Facebook and Twitter. Algorithms running the news feeds create "filter bubbles," which show content that conforms to users' preferences and biases. This can make them less tolerant toward opposing views and might also further polarize society by driving a wedge through the political and social divide.

"Algorithmic bias could potentially impact any group," says Jenn Wortman Vaughan, senior researcher at Microsoft. "Groups that are underrepresented in the data may be especially at risk."

In domains that are already known for bias, such as the tech industry's endemic discrimination against women, AI algorithms might accentuate those biases and result in further marginalization of groups that are not well represented.

Health is another critical domain, Wortman points out. "It could cause serious problems if a machine-learning algorithm being used for medical diagnosis is trained on data from one population and, as a result, fails to perform well on others," she says.

Bias can also be harmful in more subtle ways. "Last year I was planning to take my daughter for a haircut and searched online for images of 'toddler haircuts' for inspiration," Wortman says. But the images returned were nearly all of white children, primarily with straight hair. And more surprisingly, they were primarily boys.

Experts call this phenomenon "representational harm": when technology reinforces stereotypes or diminishes specific groups. "It's hard to quantify or measure the exact impact of this kind of bias, but that doesn't mean it's not important," Wortman says.

REMOVING BIAS FROM AI ALGORITHMS

The increasingly critical implications of AI bias have drawn the attention of several organizations and government bodies. And some positive steps are being taken to address the ethical and social issues surrounding the use of AI in different fields.

Microsoft, whose products rely heavily on AI algorithms, launched a research project three years ago called Fairness, Accountability, Transparency, and Ethics in AI (FATE). It's aimed at enabling users to enjoy the enhanced insights and efficiency of AI-powered services without discrimination and bias.

In some cases, like the AI-adjudicated beauty contest, finding and fixing the source of an AI algorithm's biased behavior might be as easy as checking and changing out the photos in the training dataset. But in other cases, such as the word-embedding algorithms the Boston University researchers examined, bias is engrained in the training data in more subtle ways.

The BU team, which was joined by Microsoft researcher Adam Kalai, developed a method to classify word embeddings based on their gender categorizations and to identify analogies that were potentially biased. But they didn't make the final decision; they ran each of the suspect associations by ten people on Mechanical Turk, Amazon's online marketplace for data-related tasks, who would decide whether the association should be removed.

"We didn't want to insert our own biases into the process," says Saligrama, the BU professor and researcher. "We just provided the tools to discover problematic associations. Humans made the final decision."

In a more recent paper, Kalai and other researchers proposed the use of separate algorithms to classify different groups of people instead of using the same measures for everyone. This method can prove effective in domains where existing data is already biased in favor of a specific group. For instance, the algorithms that would evaluate female applicants for a programming job would use criteria that are best suited for that group instead of using the broader set of data that is deeply influenced by existing biases.

Microsoft's Wortman sees inclusiveness in the AI industry as a necessary step to fight bias in algorithms. "If we want our AI systems to be useful to everyone and not just to certain demographics, then companies need to be hiring diverse teams to work on AI," she says.

In 2006, Wortman helped found Women in Machine Learning (WiML), which holds a yearly workshop at which women studying and working in the AI industry can meet, network, exchange ideas, and attend panel discussions with senior women in industry and academia. A similar effort is the Black in AI Workshop, founded by Timnit Gebru, another Microsoft researcher, which aims to build more diverse talent in AI.

Boston University's Bolukbasi also proposes altering the way AI algorithms solve problems: "Algorithms will choose a rule set that maximizes their objective. There may be many ways to reach the same set of conclusions for given input-output pairs," he says.

"Take the example of multiple-choice tests for humans. One may reach the right answer with a wrong thinking process but nevertheless get the same score. A high-quality test should be designed to minimize this effect, only allowing the people that truly know the subject to get correct scores. Making algorithms aware of social constraints can be seen as an analog to this example (although not an exact one), where learning a wrong rule set is penalized in the objective. This is an ongoing and challenging research topic."

AI'S OPACITY COMPLICATES FAIRNESS

Another challenge standing in the way of making AI algorithms fairer is the "black box" phenomenon. Many companies jealously guard their algorithms: For instance, Northpointe, the manufacturer of COMPAS, the crime-predicting software, won't disclose its proprietary algorithm. The only people privy to COMPAS' inner workings are its programmers, not the judges using it.

Aside from corporate secrecy, AI algorithms sometimes become so convoluted that the reasons and mechanisms behind their decisions elude even their creators. In the UK, Durham police use AI system HART to determine whether suspects have a low, moderate, or high risk of committing further crimes within a two-year period. But a 2017 academic review of HART observed that "opacity seems difficult to avoid." This is partly because of the amount and variety of data the system uses, which makes it difficult to analyze the reasons behind its decisions. "These details could be made freely available to the public, but would require a huge amount of time and effort to fully understand," the paper says.

Several companies and organizations are leading efforts to bring transparency to AI, including Google: The company has launched GlassBox, an initiative to make the behavior of machine-learning algorithms more understandable without sacrificing output quality. The Defense Advanced Research Projects Agency (DARPA), which oversees the use of AI in the military, is also funding an effort to enable AI algorithms to explain their decisions.

In other cases, human judgment will be key in dealing with bias. To prevent existing racial and social human biases from creeping into HART's algorithms, the Durham Constabulary provided members of its staff with awareness sessions around unconscious bias. The police force has also taken steps to remove data points such as racial traits, which might create the grounds for biased decisions.

HUMAN RESPONSIBILITY

From a different perspective, AI algorithms could provide an opportunity to reflect on our own biases and prejudices. "The world is biased; the historical data is biased. Hence it is not surprising that we receive biased results," Sandra Wachter, a researcher in data ethics and algorithms at the University of Oxford, told *The Guardian*.

Wachter is part of a research team from the Alan Turing Institute in London and the University of Oxford, which published a paper calling for regulations and institutions to investigate possible discrimination by AI algorithms.

Also speaking to *The Guardian,* Joanna Bryson, a computer scientist at the University of Bath and the coauthor of a research paper on algorithmic bias, said, "A lot of people are saying [algorithmic bias] is showing that AI is prejudiced. No. This is showing we're prejudiced, and that AI is learning it."

In 2016, Microsoft launched Tay, a Twitter bot that was supposed to learn from humans and engage in smart conversations. But within 24 hours of Tay's launch, Microsoft had to shut it down after it started spewing racist comments, which it had picked up from its conversations with Twitter users. Perhaps this is a reminder that it is past time we humans acknowledged our own role in the apparition and propagation of the algorithmic bias phenomenon and take collective steps to undo its effects.

"This is a very complicated task, but it is a responsibility that we as society should not shy away from," Sandra Wachter says.

Related Interests