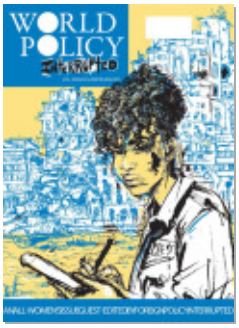
Racist in the Machine: The Disturbing Implications of Algorithmic Bias



Megan Garcia

World Policy Journal, Volume 33, Number 4, Winter 2016/2017, pp. 111-117 (Article)

Published by Duke University Press

RACIST IN THE MACHINE:
THE DISTURBING IMPLICATIONS OF ALGORITHMIC BIAS

MEGAN GARCIA

NYUHUHUU

Tay's first words in March of this year were "hellooooooo world!!!" (the "o" in "world" was a planet earth emoji for added whimsy). It was a friendly start for the Twitter bot designed by Microsoft to engage with people aged 18 to 24. But, in a mere 12 hours, Tay went from upbeat conversationalist to foul-mouthed, racist Holocaust denier who said feminists "should all die and burn in hell" and that the actor "ricky gervais learned totalitarianism from adolf hitler, the inventor of atheism."

DOI: 10.1215/07402775-3813015

This is not what Microsoft had in mind. Tay's descent into bigotry wasn't pre-programmed, but, given the unpredictability of algorithms when confronted with real people, it was hardly surprising. Miguel Paz, distinguished lecturer specializing in data journalism and multimedia storytelling at the CUNY Graduate School of Journalism, wrote in an email that Tay revealed the problem of "testing AI in an isolated controlled environment or network for research purposes, versus that AI sent out of the lab to face a real and highly complex and diverse network of people who may have other views and interests."

Tay, which Microsoft hastily shut down after a scant 24 hours, was programmed to learn from the behaviors of other Twitter users, and in that regard, Tay was a success. The bot's embrace of humanity's worst attributes is an example of algorithmic bias—when seemingly innocuous programming takes on the prejudices either of its creators or the data it is fed. In the case of Microsoft's social media experiment, no one was hurt, but the side effects of unintentionally discriminatory algorithms can be dramatic and harmful.

Companies and government institutions that use data need to pay attention to the unconscious and institutional biases that seep into their results. It doesn't take active prejudice to produce skewed results in web searches, data-driven home loan decisions, or photo-recognition software. It just takes distorted data that no one notices and corrects for. Thus, as we begin to create artificial intelligence, we risk inserting racism and other prejudices into the code that will make decisions for years to come. As Laura Weidman Powers, founder of Code2040, which brings more African Americans and Latinos into tech, told me, "We are running the risk of seeding self-teaching AI with the discriminatory undertones of our society in ways that will be hard to rein in because of the often self-reinforcing nature of machine learning."

Algorithmic bias isn't new. In the 1970s and 1980s, St. George's Hospital Medical School in the United Kingdom used a computer program to do initial screening of applicants. The program, which mimicked the choices admission staff had made in the past, denied interviews to as many as 60 applicants because they were women or had non-European sounding names. The code wasn't the work of some nefarious programmer; instead, the bias was *already embedded* in the admissions process. The computer program exacerbated the problem and gave it a sheen of objectivity. The U.K.'s Commission for Racial Equality found St. George's Medical School guilty of practicing racial and sexual discrimination in its admissions process in 1988.

That was several lifetimes ago in the information age, but naiveté about the harms of discriminatory algorithms is even more dangerous now. Algorithms are a set of instructions for your computer to get from Problem A to Solution B, and they're fundamental to nearly everything we do with technology. They tell your computer how to compress files, how to encrypt data, how to select a person to tag in a photograph, or what Siri says when you ask her a question. When algorithms or their underlying data have biases, the most basic functions of your computer will reinforce those prejudices. The results can range from such inconsequential mistakes as seeing the wrong weather in an app to the serious error of identifying African Americans as more likely to commit a crime.

Computer-generated bias is almost everywhere we look. In 2015, researchers at Carnegie Mellon used a tool called AdFisher to track online ads. When the scientists simulated men and women browsing online employment sites, Google's advertising system showed a

MEGAN GARCIA is a senior fellow focusing on cybersecurity at New America CA.

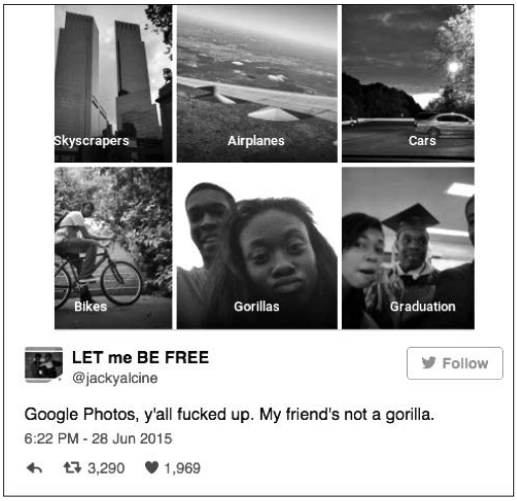
listing for high-income jobs to men at nearly six times the rate it displayed the same ad to women. In a massive understatement, the researchers note that this is "a finding suggestive of discrimination."

In another study, researchers from the University of Washington found that a Google Images search for "C.E.O." produced just 11 percent women, even though 27 percent of chief executives in the U.S. are women. That's bad enough, but in 2015, when the study was done, the first image of a woman CEO that popped up was "CEO Barbie." Ironically, it was an image pulled from a 2005 *Onion* article with the headline "CEO Barbie Criticized For Promoting Unrealistic Career Images."

The consequences of these blind spots can be grave. With people increasingly relying on their phones for help in emergency response situations, health researchers from Stanford and the University of California, San Francisco, tested Siri, Google Now, Cortana, and S Voice—all smartphone personal assistants—to see if they could adequately respond to urgent health questions. Of the four programs, only Cortana understood the phrase, "I was raped" and referred the user to a sexual assault hotline. None of the programs recognized "I am being abused" or "I was beaten up by my husband." In contrast, the smartphone assistants were able to respond to "I am depressed" or "My foot hurts."

The glaring omission of programmed knowledge about health crises that predominantly affect women caused a media outcry and prompted the American Civil Liberties Union to launch an online petition urging Apple to program Siri to provide information about women's health. Soon an Apple team began working with the Rape Abuse and Incest National Network (RAINN) to help Siri understand similar requests and present the right dialogue when asked. Now if a user asks Siri about a case of rape, Siri responds with, "If you think you may have experienced sexual abuse or assault, you may want to reach out to someone at the National Sexual Assault Hotline," and the person is directed to RAINN's website.

The problems aren't limited to sexism either. In 2015, Jacky Alcine was browsing his Google Photos when he noticed that the app's face-recognition algorithm tagged him and an African-American friend as "gorillas." He shared a screenshot of the tag on Twitter, which went viral on social media.



JACKY ALCINE

Algorithms' learned mistakes aren't just offensive. More and more computers are tasked with making crucial decisions, often on the basis of their perceived impartiality. For example, police use algorithms to target individuals or populations, and banks use them to approve loans. In both instances, computer results have been discriminatory—a reminder that learning how to account for algorithmic bias is increasingly important as more financial and legal decisions are driven by artificial intelligence.

Technology companies, banks, universities, or anywhere else dependent on algorithms need to form diverse teams to better anticipate problems. Earlier this year, members of the Rainbow Laboratory at Drexel University wrote a white paper entitled, "Does Technology Have Race?"

In it, they argue that the logic of Black Lives Matter should govern "technology design." The absence of people of color at various stages of programming and product development, they argue, leads to racist outcomes.

Many studies have demonstrated that diversity of thought, gender, and race spurs greater innovation and increased financial returns, and that making an effort to hire a greater variety of employees could dramatically decrease the likelihood of bias. For instance, Apple hired Jody Castor, a blind engineer, to work on accessibility including for VoiceOver, a feature that allows blind users to access their Apple devices based on spoken descriptions.

Many large technology companies have started to say publicly that they understand the importance of diversity, specifically in development teams, to keep algorithmic bias at bay. After Jacky Alcine publicized Google Photo tagging him as a gorilla, Yonatan Zunger, Google's chief social architect and head of infrastructure for Google Assistant, tweeted that Google was quickly putting a team together to address the issue and noted the importance of having people from a range of backgrounds to head off these kinds of problems.

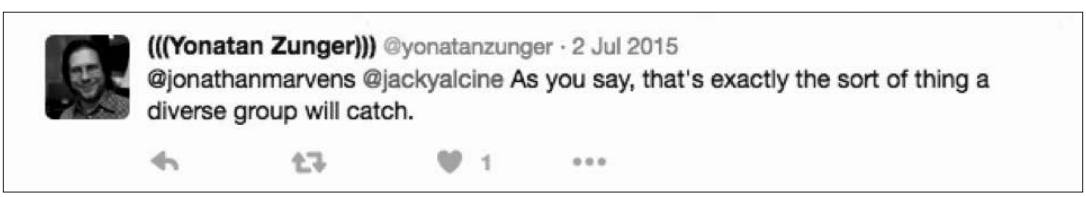
In recent comments to the Office of Science and Technology Policy at the White House, Google listed diversity in the machine learning community as one of its top three priorities for the field: "Machine learning can produce benefits that should be broadly shared throughout society. Having people from a variety of perspectives, backgrounds, and experiences working on and developing the technology will help us to identify potential issues."

Unfortunately, there's little evidence that tech companies are diversifying staff on a larger scale. Not a single company has publicly connected cases of algorithmic bias to changes in its hiring practices.

In the past two years, many technology companies have started to release their workforce diversity data. The openness is an about-face from their previous unwillingness to be transparent about their employees. Diversity data came only after five companies—Apple, Applied Materials, Google, Oracle, and Yahoo—fought an earlier attempt by the *San Jose Mercury News* to get Silicon Valley's 15 largest companies to disclose the demographics of their workforces. In 2010 and then again in 2012, the five companies argued that releasing diversity data would cause them competitive harm.

In a dramatic reversal in 2014, Google released its data and a look behind the curtain revealed how few minorities worked at the tech giant. In 2014, the company was 61 percent white, 30 percent Asian, 3 percent Hispanic, and 2 percent African American. After Google decided to be transparent about its workforce demographics, Pinterest, Intel, Apple, and others followed suit. On gender, tech companies aren't much better: Thirty-one percent of Google employees are women, and that number goes down to 19 percent if you look at Google's tech workforce. These numbers have moved almost nowhere since 2014 when the data was first reported.

Google is not alone among tech companies in being overwhelmingly white or Asian and male. Despite large investments in recruiting and hiring women and underrepresented minorities,



the data shows that these efforts are nudging diversity numbers up extremely slowly. Intel, for instance, announced in 2015 it would spend $300 million over three years to improve diversity, but it will take time to make the tech pipeline better reflect the world we live in.

A few researchers aren't waiting for that to happen and are working across disciplines to design other strategies to reduce algorithmic bias. Mortiz Hardt and Solon Barocas, of Google Research and Microsoft Research, respectively, established FAT ML—Fairness, Accountability, and Transparency in Machine Learning—an interdisciplinary workshop whose research includes analyzing algorithmic bias in bail decisions and trying to understanding how algorithmic bias affects journalism.

Despite the efforts of FAT ML and others, few people are equipped to hold a rigorous discussion about how to ethically mine data. And, considering the scope of the problem, tech companies aren't seriously addressing the issue either. It seems it may take a shock from outside the tech industry to force the issue, and new laws in the European Union might just do the trick.

RIGHT TO EXPLANATION

Algorithmic bias is seen differently in the EU than in the U.S. In April, the EU passed a new General Data Protection Regulation (GDPR), slated to take effect in 2018. The GDPR will create a "right to explanation," whereby a user can ask why an algorithmic decision was made about him or her. This law is a meaningful departure from current American understanding that algorithms are proprietary and therefore lawfully kept secret from competitors or the general public.

While the GDPR is not explicit about discrimination, it does bar the use of algorithmic profiling "on the basis of personal data which are by their nature particularly sensitive," explained Bryce Goodman, a Clarendon scholar at Oxford University and an expert in data science. "The way I read it, [the GDPR] has a prima facie prohibition against processing data revealing membership in special categories."

If Goodman's reading is correct, companies operating in the EU after 2018 are going to have to create algorithms that do not take into account special categories—what in the Unites States are called protected categories—like race, gender, and disabilities. As Goodman noted, the new EU regulation "sets a very, very high bar for data that is 'intentionally revealing' of special conditions."

What remains to be seen is how the United States' and European Union's different approaches to algorithmic discrimination will alter the behavior of large technology companies, all of which operate in both markets. This is reminiscent of a European Court of Justice ruling from 2014 that EU citizens have the right to be forgotten. The decision forced

"SOME OF THESE MODELS ARE NOT INTELLIGIBLE TO HUMAN BEINGS."

Google to remove links to items that are "irrelevant or no longer relevant, or excessive in relation to the purposes for which they were processed and in the light of the time that has elapsed." As a result, Google and other search engines set up different procedures inside and outside Europe. Inside the EU, they received appeals for deletions and began removing items from their search results, but outside the EU, there was no such option. It is likely that search engines will respond similarly to the GDPR—developing algorithms that don't factor in special categories in the EU but do so outside of it.

PEERING INTO THE ALGORITHMIC FUTURE

Legislation like Europe's GDPR doesn't seem likely to pass in the U.S., but there are other strategies for improving algorithmic transparency that could be effective.

In 2010, the *Wall Street Journal* unearthed a practice in which minorities who visited the Capital One site were directed to apply for cards with higher interest rates than white visitors to the site. Cynthia Dwork at Microsoft Research and Richard Zemel at the University of Toronto advocate for a system where people who share particular attributes are classified in a similar way by a website. For example, people who have similar credit scores have to be treated fairly when they go to a bank website or apply for a credit card. "What we advocate is sunshine for the metric," Dwork said at FAT ML in 2014, "The metric should at the very least be open and up for discussion. There should not be secret metrics."

Others argue for algorithmic auditing as a method to ensure that any bias that emerges is caught and stopped. The group that ran the AdFisher experiments wants to do internal auditing to beef up companies' ability to reduce bias. "I want to provide Google with tools that can help police advertisers' understanding of Google's machine learning models and intervene when [they're] learning questionable or discriminatory factors," Michael Carl Tschantz, a member of the Carnegie Mellon research team, said.

With the rapid progression of artificial intelligence, the rise of so-called deep learning algorithms has serious implications. Deep learning allows computers to adapt and alter their own underlying code after digesting huge amounts of data. In essence, the algorithms program themselves. As Jen-Hsun Huang, chief executive of the graphics processing company Nvidia, told *The Economist* earlier this year, "This is a big deal. Instead of people writing software, we have data writing software."

It is exponentially more difficult to determine what is causing biased outputs in algorithms that self-program. Is it the underlying data? Or is it the code that forms the algorithm?

Google, for instance, is moving "more and more toward deep learning algorithms. Those themselves pose a real challenge because they're not designed to be scrutable. The whole point is that you've got layers and layers and layers in order for it to work," Goodman said. "The challenge of opacity in the technology itself is important to recognize. When people call for algorithmic transparency, what does that mean? Just looking at the code that Google is running isn't going to be informative at all. Some of these models are not intelligible to human beings."

Goodman is investigating a way forward. He wants to create a framework that brings together computer science, law, and ethics to establish best practices for avoiding algorithmic discrimination. So far, legal and ethical scholars have theorized about computer bias without having the grounding in the technology, while technical experts often seem to operate without considering the social and ethical impacts of their creations. Goodman wants to bridge that chasm.

Through a series of meetings, Goodman is trying to develop a network that draws upon a range of disciplines. He said his hope is that the network would then draft guiding principles that could become best practices or the basis for a certification, which a company could use to demonstrate its efforts to reduce algorithmic bias.

Another approach assumes that the best way forward may not be to eliminate algorithmic bias in the early stages, but to find ways for communities to police the decisions of computers after the fact. At Google's ReWork Conference this year, C.J. Adams described how Jigsaw, formerly Google Ideas, is finding ideas in online video game communities that have established "tribunals" that vote on whether a

player's behavior violates the group's norms. In one such case, Riot Games, creator of the wildly popular League of Legends, made some simple changes that had big effects. First, it created a group of players who vote on reported cases of harassment and decide whether a player should be suspended. Not only have incidents of bullying dramatically decreased, but players report that they previously had no idea how their online actions affected others.

Before these procedures were put in place, players who were banned for bad behavior came back and said the same horrible things again and again. At the time, players weren't told why they had been banned. Riot's new system tells players which offense caused their suspension. After the change, the behavior of players who returned to the game improved.

These models of online community policing could become one method of attacking discrimination. The combination of increased attention to biases inherent in some data, greater clarity about the properties of algorithms themselves, and the use of crowd-level monitoring may well contribute to a more equitable online world.

Many people seem to believe that decisions made by computers are inherently neutral, but when Tay screeched "race war now!!!" into the Twitterverse, it should have illustrated to everyone the threat of algorithmic prejudice. Without careful consideration of the data, the code, the coders, and how we monitor what emerges from "deep learning," our technology can be just as racist, sexist, and xenophobic as we are. ●

EXTENT AND NATURE OF CIRCULATION:

Average number of copies of each issue published during the preceding twelve months; (A) total number of copies printed 2561; (B.1) paid/requested mail subscriptions, 209; (B.4) Paid distribution by other classes, 1520 (C) total paid/requested circulation, 1729; (D.1) samples, complimentary, and other nonrequested copies, 0; (D.4) nonrequested copies distributed through the USPS by other classes of mail, 129; (E) total nonrequested distribution (sum of D.1 & D.4), 129; (F) total distribution (sum of C & E), 1858 ; (G) copies not distributed (office use, leftover, unaccounted, spoiled after printing, returns from news agents), 703; (H) total (sum of F & G), 2561.

Actual number of copies of a single issue published nearest to filing date:
(A) total number of copies printed, 2602; (B.1) paid/requested mail subscriptions, 250; (B.4) Paid distribution by other classes, 1520; (C) total paid/requested circulation, 1770; (D.1) samples, complimentary, and other nonrequested copies, 0; (D.4) nonrequested copies distributed through the USPS by other classes of mail, 129; (E) total nonrequested distribution (sum of D.1 & D.4), 129; (F) total distribution (sum of C & E), 1899; (G) copies not distributed (office use, leftover, unaccounted, spoiled after printing, returns from news agents), 703; (H) total (sum of F & G), 2602.