Viewpoint: I unintentionally created a biased AI algorithm 25 years ago—tech companies are still making the same mistake

May 9 2023, by John MacCormick

The author's 1998 head-tracking algorithm used skin color to distinguish a face from the background of an image. Credit: John MacCormick, CC BY-ND

In 1998, I unintentionally created a racially biased artificial intelligence

algorithm. There are lessons in that story that resonate even more strongly today.

The dangers of bias and errors in AI algorithms are now well known. Why, then, has there been a flurry of blunders by tech companies in recent months, especially in the world of AI chatbots and image generators? Initial versions of ChatGPT produced racist output. The DALL-E 2 and Stable Diffusion image generators both showed racial bias in the pictures they created.

My own epiphany as a white male computer scientist occurred while teaching a computer science class in 2021. The class had just viewed a video poem by Joy Buolamwini, AI researcher and artist and the self-described poet of code. Her 2019 video poem "AI, Ain't I a Woman?" is a devastating three-minute exposé of racial and gender biases in automatic face recognition systems—systems developed by tech companies like Google and Microsoft.

The systems often fail on women of color, incorrectly labeling them as male. Some of the failures are particularly egregious: The hair of Black civil rights leader Ida B. Wells is labeled as a "coonskin cap"; another Black woman is labeled as possessing a "walrus mustache."

Echoing through the years

I had a horrible déjà vu moment in that computer science class: I suddenly remembered that I, too, had once created a racially biased algorithm. In 1998, I was a doctoral student. My project involved tracking the movements of a person's head based on input from a video camera. My doctoral adviser had already developed mathematical techniques for accurately following the head in certain situations, but the system needed to be much faster and more robust. Earlier in the 1990s, researchers in other labs had shown that skin-colored areas of an image

could be extracted in real time. So we decided to focus on skin color as an additional cue for the tracker.

I used a digital camera—still a rarity at that time—to take a few shots of my own hand and face, and I also snapped the hands and faces of two or three other people who happened to be in the building. It was easy to manually extract some of the skin-colored pixels from these images and construct a statistical model for the skin colors. After some tweaking and debugging, we had a surprisingly robust real-time head-tracking system.

Not long afterward, my adviser asked me to demonstrate the system to some visiting company executives. When they walked into the room, I was instantly flooded with anxiety: the executives were Japanese. In my casual experiment to see if a simple statistical model would work with our prototype, I had collected data from myself and a handful of others who happened to be in the building. But 100% of these subjects had "white" skin; the Japanese executives did not.

Miraculously, the system worked reasonably well on the executives anyway. But I was shocked by the realization that I had created a racially biased system that could have easily failed for other nonwhite people.

Privilege and priorities

How and why do well-educated, well-intentioned scientists produce biased AI systems? Sociological theories of privilege provide one useful lens.

Ten years before I created the head-tracking system, the scholar Peggy McIntosh proposed the idea of an "invisible knapsack" carried around by white people. Inside the knapsack is a treasure trove of privileges such as "I can do well in a challenging situation without being called a credit to my race," and "I can criticize our government and talk about how much I

fear its policies and behavior without being seen as a cultural outsider."

In the age of AI, that knapsack needs some new items, such as "AI systems won't give poor results because of my race." The invisible knapsack of a white scientist would also need: "I can develop an AI system based on my own appearance, and know it will work well for most of my users."

One suggested remedy for white privilege is to be actively anti-racist. For the 1998 head-tracking system, it might seem obvious that the anti-racist remedy is to treat all skin colors equally. Certainly, we can and should ensure that the system's training data represents the range of all skin colors as equally as possible.

Unfortunately, this does not guarantee that all skin colors observed by the system will be treated equally. The system must classify every possible color as skin or nonskin. Therefore, there exist colors right on the boundary between skin and nonskin—a region computer scientists call the decision boundary. A person whose skin color crosses over this decision boundary will be classified incorrectly.

Scientists also face a nasty subconscious dilemma when incorporating diversity into machine learning models: Diverse, inclusive models perform worse than narrow models.

A simple analogy can explain this. Imagine you are given a choice between two tasks. Task A is to identify one particular type of tree—say, elm trees. Task B is to identify five types of trees: elm, ash, locust, beech and walnut. It's obvious that if you are given a fixed amount of time to practice, you will perform better on Task A than Task B.

In the same way, an algorithm that tracks only white skin will be more accurate than an algorithm that tracks the full range of human skin

colors. Even if they are aware of the need for diversity and fairness, scientists can be subconsciously affected by this competing need for accuracy.

$$\begin{pmatrix} 610.3 & 710.6 & 744.3 & 210.3 \\ 710.6 & 1038.2 & 1140.5 & 167.0 \\ 744.3 & 1140.5 & 1313.7 & 152.7 \end{pmatrix}$$

This matrix is at the heart of the author's 1998 skin color model. Can you spot the racism? Credit: John MacCormick, CC BY-ND

Hidden in the numbers

My creation of a biased algorithm was thoughtless and potentially offensive. Even more concerning, this incident demonstrates how bias can remain concealed deep within an AI system. To see why, consider a particular set of 12 numbers in a matrix of three rows and four columns. Do they seem racist? The head-tracking algorithm I developed in 1998 is controlled by a matrix like this, which describes the skin color model. But it's impossible to tell from these numbers alone that this is in fact a racist matrix. They are just numbers, determined automatically by a computer program.

The problem of bias hiding in plain sight is much more severe in modern machine-learning systems. Deep neural networks—currently the most

popular and powerful type of AI model—often have millions of numbers in which bias could be encoded. The biased face recognition systems critiqued in "AI, Ain't I a Woman?" are all deep neural networks.

The good news is that a great deal of progress on AI fairness has already been made, both in academia and in industry. Microsoft, for example, has a research group known as FATE, devoted to Fairness, Accountability, Transparency and Ethics in AI. A leading machine-learning conference, NeurIPS, has detailed ethics guidelines, including an eight-point list of negative social impacts that must be considered by researchers who submit papers.

Who's in the room is who's at the table

On the other hand, even in 2023, fairness can still be the victim of competitive pressures in academia and industry. The flawed Bard and Bing chatbots from Google and Microsoft are recent evidence of this grim reality. The commercial necessity of building market share led to the premature release of these systems.

The systems suffer from exactly the same problems as my 1998 head tracker. Their training data is biased. They are designed by an unrepresentative group. They face the mathematical impossibility of treating all categories equally. They must somehow trade accuracy for fairness. And their biases are hiding behind millions of inscrutable numerical parameters.

So, how far has the AI field really come since it was possible, over 25 years ago, for a doctoral student to design and publish the results of a racially biased algorithm with no apparent oversight or consequences? It's clear that biased AI systems can still be created unintentionally and easily. It's also clear that the bias in these systems can be harmful, hard to detect and even harder to eliminate.

These days it's a cliché to say industry and academia need diverse groups of people "in the room" designing these algorithms. It would be helpful if the field could reach that point. But in reality, with North American computer science doctoral programs graduating only about 23% female, and 3% Black and Latino students, there will continue to be many rooms and many algorithms in which underrepresented groups are not represented at all.

That's why the fundamental lessons of my 1998 head tracker are even more important today: It's easy to make a mistake, it's easy for bias to enter undetected, and everyone in the room is responsible for preventing it.

Provided by The Conversation

Citation: Viewpoint: I unintentionally created a biased AI algorithm 25 years ago—tech companies are still making the same mistake (2023, May 9) retrieved 8 May 2024 from https://techxplore.com/news/2023-05-viewpoint-unintentionally-biased-ai-algorithm.html