Learning From History: AI Gender Bias

AJ Starita • March 27, 2024 • 5 min read

Application Security

Share article

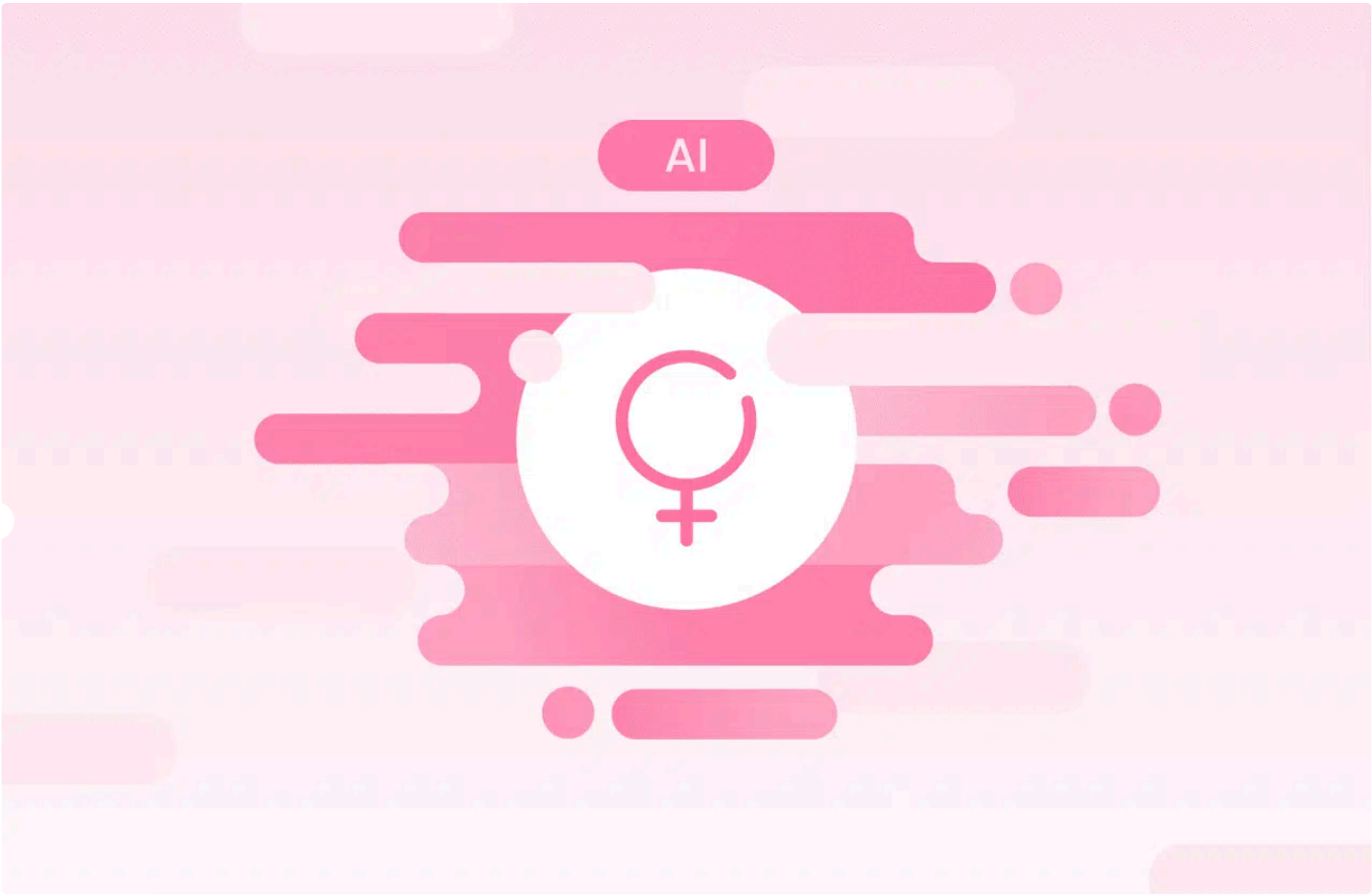

Table of Contents

From Data to Discrimination

Looking to an unbiased future

March 8th was International Women's Day; here in the U.S., the entire month of March is Women's History Month. It's a good time to talk about AI gender bias, specifically in large language models (LLMs). Why? Because one of the biggest hurdles to achieving unbiased artificial intelligence (AI) is, well, history--specifically, historical data about the world and how women have been featured in it.

From Data to Discrimination

When we train AI models to answer questions like "who would be a good candidate for this job?", we have to feed them some existing data to learn from, such as the resumes of people who were good hires in the past and how long they stayed with the company. Sounds straightforward, right? If you want to filter thousands of resumes to get the most likely great candidates, you should look for the qualities present in the resumes of great employees of the past. But if the job happens to be in a historically male-dominated field (hello, tech sector!), the data may teach the AI model that being male is a positive indicator for a good candidate, and being female is a negative one. That's exactly what happened at Amazon before they scrapped their project.

Resumes for people in technical roles isn't the only data set that's lacking in female representation. Much of the data we have about health, including medicine interactions and symptom expression, comes from studies and records that often exclude women. In 1993, the United States Congress mandated women and minority participation in National Institute of Health studies and the U.S. Food and Drug Administration reversed guidelines that excluded women from the early stages of drug trials. Prior to that, almost all health data gathered in America came from men. But even since then, health data parity has not been—and will likely never be—reached. One large source of medical data in the U.S. comes from the U.S. military, and only 17% of members of the armed forces are female.

Not getting a job due to gender discrimination is a terrible thing, but AI biases in health may end up putting women's lives in danger.

The main reason women were excluded from health studies for so long was that researchers commonly felt that women's hormone levels were a variable that would be too cumbersome to account for. They figured it wasn't worth the effort when they could simply extrapolate data they got from men to apply it to women. But women aren't just smaller men. Differences between the sexes are more than differences in anatomy, and women's symptoms don't always manifest the same way as they do in men. Examples of this include cardiovascular disease (particularly signs of heart attacks) and sleep apnea, where women are often mis- or underdiagnosed because their symptoms don't show up in a "textbook" way. Those looking to build AI models to assist in diagnosing patients will need to account for this.

Data collected from crash test dummies also leaves women out. In a car crash, women are 17% more likely to die and 78% more likely to sustain serious injury. Some of this disparity is inevitable due to women's lower bone density and muscle mass, but how much is unknown. Car manufacturers are only required to use the "average male" dummy for testing, which means that safety features like seatbelts and airbags are thus tuned to perform well for men, not women. Although one female dummy exists, it is simply a scaled-down version of the male dummy. It does not account for the differences between men and women in bone structure or bone mass, and represents only the bottom 5th percentile of women by height and weight. Testing with the small female dummy is entirely optional and currently only done by a small number of manufacturers. That means that when we focus AI models on the task of designing safe cars for all, they're going to need more information than we can give them at this time, or they may just design cars that are safer for men with little to no regard to safety performance for women.

Looking to an unbiased future

There is sometimes a hope that AI can take humans "out of the equation" for some tasks and therefore remove bias, but the unfortunate truth is that because AI is human-built using human-collected data, we can't truly take humans out of the equation. As we strive for equality, it's important to remember that bias can be not only simply present in data and repeated, but actually ***amplified*** by AI models when deep learning learns outdated notions about men and women too deeply.

The teams building AI models today almost assuredly do not intend to proliferate bias against women. With proper controls and thoughtful training, gender bias in AI models can be alleviated, if not outright eliminated. Governments, research institutions, non-profits, and corporations are all beginning to see a need to improve their processes for

collecting meaningful data pertaining to women. With more and better data, improvement is on the horizon for both AI models and their impact on the lives of women.

About the author

AJ Starita

AJ Starita is fascinated by the challenges and triumphs of cybersecurity and open source software. When not writing about technology, AJ can usually be found exploring nature or reading detective novels.

Recent resources

What is the KEV Catalog?

AJ Starita • 19 Sep 2024

Application Security

Application Security — The

AJ Starita • 12 Sep 2024

Application Security