


How our data encodes systematic racism

 technologyreview.com/2020/12/10/1013617/racism-data-science-artificial-intelligence-ai-opinion

I've often been told, "The data does not lie." However, that has never been my experience. For me, the data nearly always lies. Google Image search results for "healthy skin" show only light-skinned women, and a query on "Black girls" still returns pornography. The CelebA face data set has labels of "big nose" and "big lips" that are disproportionately assigned to darker-skinned female faces like mine. ImageNet-trained models label me a "bad person," a "drug addict," or a "failure." Data sets for detecting skin cancer are missing samples of darker skin types.

White supremacy often appears violently—in gunshots at a crowded Walmart or church service, in the sharp remark of a hate-fueled accusation or a rough shove on the street—but sometimes it takes a more subtle form, like these lies. When those of us building AI systems continue to allow the blatant lie of white supremacy to be embedded in everything from how we collect data to how we define data sets and how we choose to use them, it signifies a disturbing tolerance.

Non-white people are not outliers. Globally, we are the norm, and this doesn't seem to be changing anytime soon. Data sets so specifically built in and for white spaces represent the constructed reality, not the natural one. To have accuracy calculated in the absence of my lived experience not only offends me, but also puts me in real danger.

Corrupt data

In a research paper titled "Dirty Data, Bad Predictions," lead author Rashida Richardson describes an alarming scenario: police precincts suspected or confirmed to have engaged in "corrupt, racially biased, or otherwise illegal" practices continue to contribute their data to the development of new automated systems meant to help officers make policing decisions.

The goal of predictive policing tools is to send officers to the scene of a crime before it happens. The assumption is that locations where individuals had been previously arrested correlate with a likelihood of future illegal activity. What Richardson points out is that this assumption remains unquestioned even when those initial arrests were racially motivated or illegal, sometimes involving "systemic data manipulation, police corruption, falsifying police reports, and violence, including robbing residents, planting evidence, extortion, unconstitutional searches, and other corrupt practices." Even data from the worst-behaving police departments is still being used to inform predictive policing tools.

As the Tampa Bay Times reports, this approach can provide algorithmic justification for further police harassment of minority and low-income communities. Using such flawed data to train new systems embeds the police department's documented misconduct in the

algorithm and perpetuates practices already known to be terrorizing those most vulnerable to that abuse.

This may appear to describe a handful of tragic situations. However, it is really the norm in machine learning: this is the typical quality of the data we currently accept as our unquestioned “ground truth.”

Related Story

What’s missing from corporate statements on racial injustice? The real cause of racism.

An analysis of 63 recent statements shows that US tech companies repeatedly placed responsibility for racial injustice on Black people.

One day GPT-2, an earlier publicly available version of the automated language generation model developed by the research organization OpenAI, started talking to me openly about “white rights.” Given simple prompts like “a white man is” or “a Black woman is,” the text the model generated would launch into discussions of “white Aryan nations” and “foreign and non-white invaders.”

Not only did these diatribes include horrific slurs like “bitch,” “slut,” “nigger,” “chink,” and “slanteye,” but the generated text embodied a specific American white nationalist rhetoric, describing “demographic threats” and veering into anti-Semitic asides against “Jews” and “Communists.”

GPT-2 doesn’t think for itself—it generates responses by replicating language patterns observed in the data used to develop the model. This data set, named WebText, contains “over 8 million documents for a total of 40 GB of text” sourced from hyperlinks. These links were themselves selected from posts most upvoted on the social media website Reddit, as “a heuristic indicator for whether other users found the link interesting, educational, or just funny.”

However, Reddit users—including those uploading and upvoting—are known to include white supremacists. For years, the platform was rife with racist language and permitted links to content expressing racist ideology. And although there are practical options available to curb this behavior on the platform, the first serious attempts to take action, by then-CEO Ellen Pao in 2015, were poorly received by the community and led to intense harassment and backlash.

Whether dealing with wayward cops or wayward users, technologists choose to allow this particular oppressive worldview to solidify in data sets and define the nature of models that we develop. OpenAI itself acknowledged the limitations of sourcing data from Reddit, noting

that “many malicious groups use those discussion forums to organize.” Yet the organization also continues to make use of the Reddit-derived data set, even in subsequent versions of its language model. The dangerously flawed nature of data sources is effectively dismissed for the sake of convenience, despite the consequences. Malicious intent isn’t necessary for this to happen, though a certain unthinking passivity and neglect is.

Little white lies

White supremacy is the false belief that white individuals are superior to those of other races. It is not a simple misconception but an ideology rooted in deception. Race is the first myth, superiority the next. Proponents of this ideology stubbornly cling to an invention that privileges them.

I hear how this lie softens language from a “war on drugs” to an “opioid epidemic,” and blames “mental health” or “video games” for the actions of white assailants even as it attributes “laziness” and “criminality” to non-white victims. I notice how it erases those who look like me, and I watch it play out in an endless parade of pale faces that I can’t seem to escape—in film, on magazine covers, and at awards shows.

| Data sets so specifically built in and for white spaces represent the constructed reality, not the natural one.

This shadow follows my every move, an uncomfortable chill on the nape of my neck. When I hear “murder,” I don’t just see the police officer with his knee on a throat or the misguided vigilante with a gun by his side—it’s the economy that strangles us, the disease that weakens us, and the government that silences us.

Tell me—what is the difference between overpolicing in minority neighborhoods and the bias of the algorithm that sent officers there? What is the difference between a segregated school system and a discriminatory grading algorithm? Between a doctor who doesn’t listen and an algorithm that denies you a hospital bed? There is no systematic racism separate from our algorithmic contributions, from the hidden network of algorithmic deployments that regularly collapse on those who are already most vulnerable.

Resisting technological determinism

Technology is not independent of us; it’s created by us, and we have complete control over it. Data is not just arbitrarily “political”—there are specific toxic and misinformed politics that data scientists carelessly allow to infiltrate our data sets. White supremacy is one of them.

We’ve already inserted ourselves and our decisions into the outcome—there is no neutral approach. There is no future version of data that is magically unbiased. Data will always be a subjective interpretation of someone’s reality, a specific presentation of the goals and

perspectives we choose to prioritize in this moment. That's a power held by those of us responsible for sourcing, selecting, and designing this data and developing the models that interpret the information. Essentially, there is no exchange of "fairness" for "accuracy"—that's a mythical sacrifice, an excuse not to own up to our role in defining performance at the exclusion of others in the first place.

Stay updated on MIT Technology Review initiatives and events?

Those of us building these systems will choose which subreddits and online sources to crawl, which languages to use or ignore, which data sets to remove or accept. Most important, we choose who we apply these algorithms to, and which objectives we optimize for. We choose the labels we create, the data we take in, the methods we use. We choose who we welcome as data scientists and engineers and researchers—and who we do not. There were many possibilities for the design of the technology we built, and we chose this one. We are responsible.

So why can't we be more careful? When will we finally get into the habit of disclosing data provenance, deleting problematic data sets, and explicitly defining the limitations of every model's scope? At what point can we condemn those operating with an explicit white supremacist agenda, and take serious actions for inclusion?

An uncertain path forward

Distracted by corporate condolences, abstract technical solutions, and articulate social theories, I've watched peers congratulate themselves on invisible progress. Ultimately, I envy them, because they have a choice in the same world where I, like every other Black person, cannot opt out of caring about this.

As Black people now die in a cacophony of natural and unnatural disasters, many of my colleagues are still more galvanized by the latest product or space launch than the jarring horror of a reality that chokes the breath out of me.

| The fact is that AI doesn't work until it works for all of us.

For years, I've watched this issue extolled as important, but it's clear that dealing with it is still seen as a non-

priority, "nice to have" supplementary action—secondary always to some definition of model functionality that doesn't include me.

Models clearly still struggling to address these bias challenges get celebrated as breakthroughs, while people brave enough to speak up about the risk get silenced, or worse. There's a clear cultural complacency with things as usual, and although disappointing, that's not particularly surprising in a field where the vast majority just don't understand the stakes.

The fact is that AI doesn't work until it works for all of us. If we hope to ever address racial injustice, then we need to stop presenting our distorted data as "ground truth." There's no rational and just world in which hiring tools systematically exclude women from technical roles, or where self-driving cars are more likely to hit pedestrians with darker skin. The truth of any reality I recognize is not in these models, or in the data sets that inform them.

The machine-learning community continues to accept a certain level of dysfunction as long as only certain groups are affected. This needs conscious change, and that will take as much effort as any other fight against systematic oppression. After all, the lies embedded in our data are not much different from any other lie white supremacy has told. They will thus require just as much energy and investment to counteract.

Deborah Raji is a Mozilla fellow interested in algorithmic auditing and evaluation. She has worked on several award-winning projects to highlight cases of bias in computer vision and improve documentation practices in machine learning.