Bias Optimizers

Artificial intelligence readily amplifies human prejudice; making a better AI will require a new approach to both regulation and ethics.

By Damien P. Williams

Recently I learned that men can sometimes be nurses and secretaries, but women can never be doctors or presidents. I learned that Black people are more likely to owe money than to have it owed to them. And I learned that if you need disability assistance, you'll get more of it if you live in a facility than if you receive care at home.

At least, that is what I would think if I took responses from today's new artificial intelligence (AI) systems at face value. It has been less than a year since OpenAI released ChatGPT and mere months since its GPT-4 update and Google's release of a competing AI chatbot, Bard. Creators promise these systems will make our lives easier, removing drudge work like writing emails, filling out forms, and even writing code. But the bias programmed into these systems threatens to spread more prejudice into the world. AI-facilitated biases can affect who gets hired for what jobs, who gets believed as an expert in their field, and who is more likely to be targeted and prosecuted by police.

For some people, the word *bias* is synonymous with prejudice, a bigoted and closed-minded way of thinking that precludes new understanding. But bias also implies a set of fundamental values and expectations. For an AI system, bias may be a set of parameters that allow a system or agent to achieve a goal.

Like all technologies, AI reflects human bias and values, but it also has an unusually great capacity to amplify them as well. This means we must be purposeful about how we build AI systems so that they amplify the values we want them to, rather than the ones accidentally fed into them. We have to ask questions about the source material that trains them, including books, social media posts, news and academic articles, and even police reports and patient information. We must also examine the frameworks into which that data is placed: What is the system doing with that data? Are some patterns or relationships between certain words or phrases given more value than others? Which ones? Why? What are the assumptions and values at play in the design of tools that transform human lived experiences into data, and that data into algorithms that impact human lives?

It is much easier to see through the mystique of ChatGPT and other AI applications once you understand exactly what they are and what they do. The truth about such algorithms is that they're literally just sets of instructions. You have a set of standardized operations within which particular weights and measures can be adjusted. In so doing, you have to adjust every element of the whole to

make sure the final product still turns out the right way. Algorithms are often sold as magical, but they are neither unexplainable, nor even terribly unfamiliar. The recipe for any food — just as for anything you have to make — is an algorithm, too. My favorite algorithm is pumpkin pie. If you go to make a pumpkin pie, you might decide you'd like cinnamon, cardamom, and nutmeg to be a bigger part of the pie. But you can't adjust the proportion of the pie's dry ingredients without considering the rest, or you'll end up with a crumbly, spongy mess; it won't really be a good pie. You must adjust the whole recipe, the whole algorithm.

To the person using it, an algorithm may look like a unitary thing that performs one job: A Google search, for instance, seems like a singular, powerful operation that searches the web. In reality, platforms and search engines work on dozens of algorithms which search, sort, rank, weight, associate, suggest, amplify, and suppress words, concepts, and content. Those algorithms work in concert, but when you take a matrix of algorithms and automate it, it looks as if your computer system is autonomous and self-directed. With the new AI chatbots, it feels a lot like the promise of "true artificial intelligence," a seductive idea that goes back to the dawn of the computer age.

A HISTORY OF BIAS

Since the 1940s, when mathematicians and cryptographers like Joan Clarke, Jane Hughes, Pamela Rose,the other ~8,000 women of Bletchley Park, and Alan Turing broke complex codes to help win World War II, people have wondered about the possibility of intelligence in digital computers. In the 1950s, computer researchers began to ask, "Can Machines Think?" And in the 1960s, a rift formed between two camps of AI researchers at Dartmouth. One group focused on computation and cybernetics, feedback loops that mimic biological processes. The other group worked to replicate human neural networks in electronic form. Neither camp considered machine bodies, emotions, or socialization, however. These researchers firmly believed that the key to AI was to divorce any messy social factors from the purity of rationality and intellect.

As part of this work, scientists developed language models (LMs), a method of determining the probability of words connecting to each other based on context cues such as their starting letter and the preceding word. One of the earliest examples was ELIZA, a program developed by computer scientist Joseph Weizenbaum at MIT in 1964. ELIZA was meant to parody open-ended psychotherapy, so the program would do things like rephrasing the typed inputs from users as parroted questions rather than replying with anything new. Even knowing that they were talking to a computer, humans repeatedly formed emotional bonds with ELIZA, often in as little as one or two short conversations. Weizenbaum was astounded at what he called the "powerful delusional thinking" such a brief engagement could produce.

ELIZA was one of the first mainstream language models, but the work didn't end there. The dream of artificial intelligence grew up alongside the dream of natural language processing. Researchers

working on natural language processing sought to combine linguistics, computer science, artificial neural networking, and AI to find ways for computers to interpret, process, and communicate in human-like, conversational language. In the 2010s, the programs Global Vectors for Word Representation (GloVe) and Word2Vec were two of the foremost examples of natural language processing programs. They work by statistically mapping the relationships between words, embedding layers of associative meaning between them.

LMs could represent the semantic connections between words like "dog" and "dig" or "plane" and "flight." These early programs used so-called machine learning, a process of encoding various elements of English language as data, and then training the system to hit particular predictive targets and reinforce associations between the datapoints. Those associations are then mapped as mathematical representations of how strongly they're associated. In a sense they are complex autocomplete programs: They predict the ways words are likely to be strung together based on the ways language is typically organized in books, stories, articles, and so on.

But Word2Vec and GloVe had two major problems. First, their outputs often contained prejudicial bias. This occurred because the most readily available language sets on which they could be trained included things like the over 600,000 emails generated by 158 employees of the Enron Corporation in the years before the company collapsed. This particular data set was full to the brim with human beings speaking in bigoted, immoral, or even just unconsciously biased ways about certain groups of other humans. Within what researchers now call the 'Enron Corpus,' you will find people trading and rating pictures of women; slurs against anyone of perceived Muslim background; and stereotypical "jokes" about the sexual proclivities of Black and Asian peoples. Tools using this material replicated and iterated the same prejudices, resulting in outcomes such as automated résumé sorters rejecting the applications of women and certain minorities at higher rates than white men.

The second problem was that Word2Vec and GloVe could not map associations across larger reams of text. The number of associations they could make actually decreased the larger the quantity of text got. These models group related words into compact, easily embedded representations; repeated word clusters translate into more strongly related associations. Thus, the larger the corpus, the more difficulty these older programs have mapping connections across the whole text, rather than the small, repeated clusters. Using more text as input requires different solutions— and thus the transformer framework was born.

BIRTH OF THE TRANSFORMER

The "GPT" in ChatGPT stands for "Generative Pretrained Transformer." Its name describes a system of interoperable algorithms that weigh, arrange, and create associative distributions of text. They're built on large language models (LLMs), a subtype of LMs developed over the past five years or so, with datasets millions, billions, and now even trillions of words in size. LLMs are

trained through deep learning — multiple layers of machine learning operating on and refining each other.

LLMs and the applications that use them, much like the forerunner language-model systems, are a form of automated word association, in which words and phrases known as "language corpora" are turned into mathematical representations known as "tokens." The system is then trained on the tokens to predict the association between them. Well-trained natural language processing systems can interact with and guide a human through any number of tasks, from navigating a website to completing a complicated administrative application — or so the theory goes.

This approach often appears to work. You can use GPTs to generate a story, summarize a book, or even just have a conversation. When someone types in a collection of words, the transformer measures those words against the tokens, and then generates a collection of words and phrases in a particular form, all with a high likelihood of fidelity to what the user sought. But the new systems retain the same prejudicial problems as Word2Vec, only now those problems multiply faster and more extensively than ever before.

Prejudicial bias not only informs the input and output of these systems, but the very structures on which they are built. If Google image recognition is trained on more examples of cats than Black people; or if the testing group for a digital camera's blink detection includes no people of Asian descent; or if the very basis of photographic technology doesn't see dark skin very well, how can you possibly be surprised at the biased results?

Because of those embedded biases, predictive policing systems tied to algorithmic facial recognition regularly misidentify Black subjects and recommend over-policing in Black communities. Algorithmic benefits distribution systems meant to serve disabled populations are dependent on outdated notions about standards of care for disability, both in the training data and in the weights and operations within the models themselves. AI applications in healthcare and health insurance routinely recommend lower standards of care to already vulnerable and marginalized individuals and groups. Rua Williams at Purdue University and independent AI researcher Janelle C. Shane have shown that GPT checkers have problems with original text written by neurodivergent individuals. Entering such text into automated plagiarism-checking software, which already endangers disabled and otherwise marginalized students, has a high likelihood of producing harmful false positives—something admitted to by automated plagiarism company Turnitin in late May of 2023.

In general, systems trained on the "natural language" people use on the internet when they talk about marginalized groups is likely to cast those groups as lesser. Expressions of prejudicial values and biases are not restricted to explicit slurs and threats of physical violence; they can also emerge more subtly as webs of ideas and beliefs that may show up in all kinds of speech, actions, and systems.

Such prejudices are inherent in the data used to train AI systems. The factual and structural wrongness is then reinforced as the AI tools then issue outputs which are labeled "objective" or "just math." These systems behave the way that they do because they encode prejudicial and even outright bigoted beliefs about other humans during training and use. When it comes to systems like ChatGPT, these problems will only increase as they get more powerful and seem more "natural." Their ability to associate, exacerbate, and iterate on perceived patterns — the foundation of how LLMs work —will continue to increase the bias within them.

Because machine learning reinforces these processes, the technology becomes a confirmation bias optimizer. The systems generate responses that seem like factual answers in fluid language, but that output is just matching what it's been trained to associate as the most correct-seeming collection of tokens. GPTs do not care when they get something wrong or perpetuate a harmful prejudice. They are designed only to give you an answer you're statistically more likely to accept.

That innocent-sounding goal contains immense potential for harm. Imagine integrating an AI that discerns the ethnicity of a patient from a set of x-rays, and then combining it with another AI that consistently misdiagnoses signs of renal illness in Black patients — or with one that recommends lower standards of care. Now add a chat integration intended to help patients understand their diagnoses and treatment options. Then feed all of that back to human doctors as suggestions and recommendations for how they ought to interact with the human patient in front of them.

AI models have been called as revolutionary as the internet itself. They've also been compared to precocious children. But at present, these children are the spawn of hegemonic corporations fundamentally motivated by maximizing profit. Should we really give them the authority to control what we consider real knowledge in the world?

RETHINKING THE SYSTEM

If generative AI systems like ChatGPT and Bard are meant to merely reflect the world as it has been, then they are extremely well-suited to that task. But if we want them to help us make decisions toward a better future, one in which we're clear about which values we want in our technologies and our cultures, then we need to rethink everything about them.

We know that we can mitigate AI's replication and iteration of prejudicial bias by intentionally altering the weights and associative tokens. In colloquial terms, doing so would tell the system to model the world in a different way. To do that — to engage in a process known as "bias bracketing" — these systems would have to be built on a framework that constantly checks, rechecks, and reevaluates the associations it has, and actively seeks out other, alternative associations.

Self-evaluating for bias, including implicit bias, is something that even most humans do not do well. Learning how to design, build, and train an algorithmic system to do it automatically is by no means a small task. Before that work could begin, the builders would also have to confront the fact that even after mitigation, some form of bias will always be present.

We also need to take a step back and reconsider the question, "What are these 'AI' tools meant to do?" and understand that human values, beliefs, and assumptions will always influence our answers. Used differently, GPTs could help us recognize and interrogate the biases in our language and our social structures, then generate new ideas, riffing and remixing from what already exists.

Imagine how much fairer and more constructive these tools might be if the data used to train them were sourced ethically from public domain works, or from people who volunteer their data, with a record of provenance, so we could be sure AI is not generating text or art that is essentially stolen from human creators. Imagine if GPTs had to obtain permission to use someone's data, and if data collection were opt-in rather than opt-out. Imagine how much more we could trust these systems if regulations required them to clearly state that they aren't truth-telling machines but are instead bullshit engines — systems designed only to spit out collections of words that are statistically likely to jibe with our inputs. Imagine if the architectures of these GPT tools were shaped not primarily by corporate interests but by those most marginalized and most likely to be subject to and negatively impacted by them.

To build these systems differently will require more than a "pause" on development, as some AI researchers have repeatedly suggested. It will require AI systems creators to be fully honest about what these systems are and what they do. It will require a reformulation of values, real oversight and regulation, and an ethic that sees marginalized people not as test subjects but as design leaders. Above all, it will require all of us to push hard against the prejudices that inform our creations and that flow out to us.