Futurism

THINGS FALL APART / ARTIFICIAL INTELLIGENCE

# When AI Is Trained on AI-Generated Data, Strange Things Start to Happen

"Loops upon loops."

UPDATED AUG 2 *by* MAGGIE HARRISON

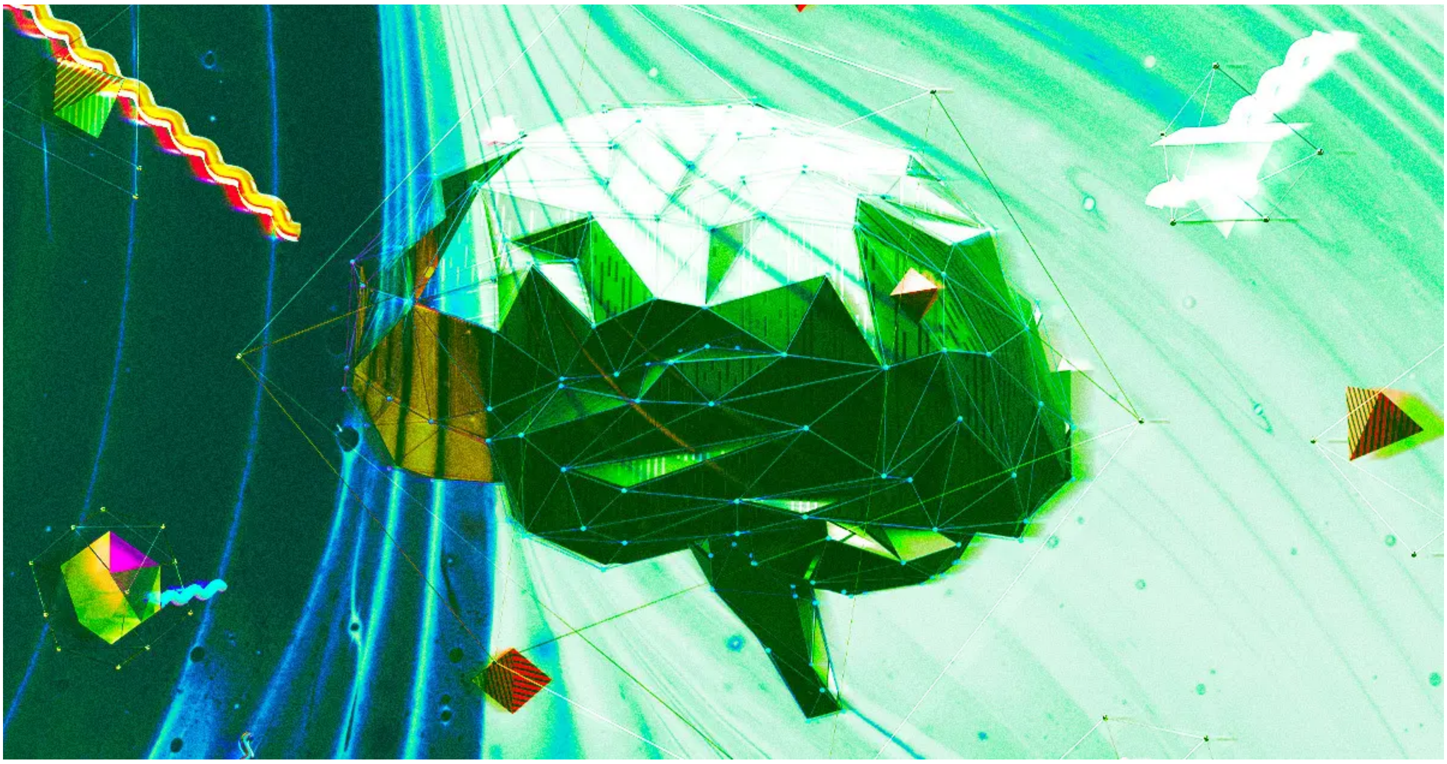

Image by Getty / Futurism

It hasn't even been a year since OpenAI released ChatGPT, and already generative AI is everywhere. It's in classrooms; it's in political advertisements; it's in entertainment and journalism and a growing number of AI-powered content farms. Hell, generative AI has even been integrated into search engines, the great mediators and organizers of the open web. People have already lost work to the tech, while new and often confounding AI-related careers seem to be on the rise.

Though whether it sticks in the long term remains to be seen, at least for the time being generative AI seems to be cementing its place in our digital *and* real lives. And as it becomes increasingly ubiquitous, so does the synthetic content it produces. But in an ironic twist, those same synthetic outputs might also stand to be generative AI's biggest threat.

That's because underpinning the growing generative AI economy is human-made data. Generative AI models don't just cough up human-like content out of thin air; they've been trained to do so using troves of material that actually *was* made by humans, usually scraped from the web. But as it turns out, when you feed synthetic content back to a generative AI model, strange things start to happen. Think of it like data inbreeding, leading to increasingly mangled, bland, and all-around bad outputs. (Back in February, Monash University data researcher Jathan Sadowski described it as "Habsburg AI," or "a system that is so heavily trained on the outputs of other generative AI's that it becomes an inbred mutant, likely with exaggerated, grotesque features.")

It's a problem that looms large. AI builders are continuously hungry to feed their models more data, which is generally being scraped from an internet that's increasingly laden with synthetic content. If there's too much destructive inbreeding, could everything just... fall apart?

To understand this phenomenon better, we spoke to machine learning researchers Sina Alemohammad and Josue Casco-Rodriguez, both PhD students in Rice University's Electrical and Computer Engineering department, and their supervising professor, Richard G. Baraniuk. In collaboration with researchers at Stanford, they recently published a fascinating — though yet to be peer-reviewed — paper on the subject, titled "Self-Consuming Generative Models Go MAD."

MAD, which stands for Model Autophagy Disorder, is the term that they've coined for AI's apparent self-allergy. In their research, it took only five cycles of training on synthetic data for an AI model's outputs to, in the words of Baraniuk, "blow up."

It's a fascinating glimpse at what just might end up being generative AI's Achilles heel. If so, what does it all mean for regular people, the burgeoning AI industry, and the internet itself?

*This interview has been edited for length and clarity.*

***Futurism: So you coined a term for the phenomenon of AI self-consumption: MAD. Can you tell us what that acronym stands for, and what that phenomenon entails?***

**Richard G. Baraniuk:** So, AI is a big field. Generative AI is one important part, and one that the public has become really aware of lately. Generative models generate or synthesize data. So in ChatGPT a person types in a prompt, then the GPT model synthesizes text to write a response. Or if you're using image generators like DALL-E or Stable Diffusion, you put in a text prompt, and the system generates a digital image.

So engineers develop these generative models. They're basically a computer program, and the program that needs to be trained. Right? And it needs to be trained with huge amounts of data from the internet. ChatGPT was trained on as much of the world wide web as OpenAI could find — basically everything. DALL-E was similarly trained on as many digital images out on the web that could be found.

And the crux is that increasingly, AI models are being trained on not just natural data, or real data sourced from the real world. Now, they're also being trained with data that's been synthesized by other generative models. So consequently, we've entered an age where either wittingly or unwittingly, generative models are increasingly consuming the outputs from other generative models. Some companies are willingly training generative models on synthetic data, particularly when they're in an area where there just isn't enough real data. Some people are unwittingly training on generative models — for example, it turns out that a lot of the big training datasets that are used today for learning actually contain synthetic images generated by other generative models.

We summarize this in what we call an autophageous loop. That's a technical term that basically just means self-consuming. Maybe think of an animal, not just chasing his tail, but eating his tail. We also like the analogy to Mad Cow Disease — feeding cows to other young cows in an ever-repeating cycle that leads to brain-destroying pathogens. So basically, our work has been studying these self-consuming loops, and understanding when models go MAD, if you will. When bad things happen, and what to do if you don't want bad things to happen.

***Futurism: Is there a certain threshold where synthetic content starts to cause problems? As you've said, synthetic content is already making its way into AI training sets. But how much synthetic content does it take for a model to go MAD and break down?***

***Josue Casco-Rodriguez:*** It definitely varies from model to model situation to situation.

***Sina Alemohammad:*** Let's look at this intuitively. Say you have 1 billion pieces of natural data, and you have one piece of synthetic data. MAD won't happen. But one year later, if you have 1 billion pieces of synthetic data, then definitely it'll go MAD in five iterations. We have found this ratio in Gaussian modeling.

***Baraniuk:*** Right. And just to be clear, there absolutely is a threshold for every model. But figuring out for DALL-E versus Midjourney, what the exact balance of real and synthetic data needs to be to keep everything safe and not going MAD, that's still a subject of research. But now the question for these big industrial models is, well, what is it?

***Futurism: So what are some of the implications for AI companies, then? We could say, ideally, that none of this ever gets into datasets. But that's obviously happened — after all, people are already using AI to do Mechanical Turk work.***

***Casco-Rodriguez:*** So when you're unwittingly using synthetic data — that also applies to practitioners, people who are generating images and putting them on the internet — you're probably not going to be conscious of the fact that what you produce is going to be in the future training of generative models. We see this with the dataset Laion-5B, for example, which was used to train Stable Diffusion: generative images that people made in the past are being used to train new generative models. So if people are producing synthetic data, they need to be conscious of this fact. On the company side, your best shot is using something like watermarking to be able to detect synthetic data and maybe remove it.

As far as when you're knowingly using synthetic data, you need to be conscious that these generative models aren't perfect, and that they oftentimes do things like sacrifice synthetic diversity for synthetic quality. If that's happening in your model, and you're training on it, then you need to be cognizant that that's happening.

***Baraniuk:*** Say there are companies that, for whatever reason — maybe it's cheaper to use synthetic data, or they just don't have enough real data — and they just throw caution to the wind. They say, "we're going to use synthetic data." What they don't realize is that if they do this generation after generation, one thing that's going to happen is the artifacts are going to be amplified. Your synthetic data is going to start to drift away from reality.

That's the thing that's really the most dangerous, and you might not even realize it's happening. And by drift away from reality, I mean you're generating images that are going to become increasingly, like, monotonous and dull. The same thing will happen for text as well if you do this — the diversity of the generated images is going to steadily go down. In one experiment that we ran, instead of artifacts getting amplified, the pictures all converge into basically the same person. It's totally freaky.

***Futurism: What are the implications for users of these systems?***

***Baraniuk:*** It's difficult for the users to protect themselves. If these models are being used in a loop like this, unfortunately, from a user perspective, the content they're generating is just going to become increasingly dull. And that's going to be disappointing, right? That's just fact.

So what can users really do to help the situation? One thing they can do is not turn off watermarking where it exists. There are some downsides to watermarking, but if someone's training a new model, they could seek synthetic images with watermarks and throw them out. That would really help with this threshold effect that we talked about. The second thing that users need to know is that their outputs, if they put them on the web, are going to invariably leak into training datasets for future systems. Some things are just inevitable.

***Futurism: What are the downsides of watermarking?***

***Baraniuk:*** It intentionally introduces an artifact. And compounded over generations, those can blow up like the AI-generated images in our paper.

***Alemohammad:*** Yeah, the problem is unknown. We don't know how the watermark can or will be amplified. But definitely, the benefits outweigh the downside. Right now, watermarking is the solution that we need to find synthetic data.

***Futurism: AI is already being integrated into services throughout the internet, most notably search engines. And search engines, to some capacity, are how we do almost everything online — they're central to the mediation and navigation of the***

***web. Are you at all concerned about the future of the web's usability, if generated AI models are integrated into the web and into our daily lives, and then start to degrade because they keep swallowing synthetic material?***

***Baraniuk:*** This a really important long-term question. There's no question that MADness has the potential to significantly reduce the quality of the data on the internet. Just the quality of the data. And our work in this particular paper hasn't really dealt with the kind of AI systems used, let's say, in search engines. So it's a bit too early to tell. But there is some other work out there that actually shows that if you train a different, non-generative AI system — some other kind of AI system like the kind used in search engines — if you train that AI system using synthetic data in addition to real data, performance actually goes down.

So this supports the hypothesis that the more synthetic data that's out there, it could actually lower the performance of a whole host of tools, search engines included, that are trained on all of this data out there on the internet — some of which is real, but some of which is synthetic. Folks are starting to connect those dots.

***Casco-Rodriguez:*** One thought I've had is that the ping-ponging back and forth between models can be really freaky. And since generative AI is already being used to do things like generate websites entirely, you could wind up having generative models leading you to results that are also synthetic, that have hyperlinks to other synthetic websites. There could be a whole synthetic ecosystem that you find through search engines, which is kind of crazy.

***Baraniuk:*** Yeah, you'd get trapped in that world. It connects back to how people are using ChatGPT to do Mechanical Turk work. In order to do supervised learning — which is one of the big ways people learn from data to build these kinds of models — you need to label data. This kind of data annotation has been the gold standard, but now they're finding that when you put these labeling tasks out on a service like, say, Mechanical Turk, people aren't doing it anymore. They're just asking AI systems to do the labeling for them. It's more efficient, but it puts us exactly in another one of these loops we've been talking about. Loops upon loops.

***Futurism:*** Snake eating its tail.

***Baraniuk:*** No question. Again, it's this idea of loops on top of loops, and that can make it extremely hard to ultimately track down the source of where any problems in AI models are coming from.

A quick story there: the whole jumping off point for our research happened when one of our group members was at a conference presenting a poster — not on this stuff, but on related work — and there was a researcher from industry who walked by who, just offhandedly, remarked that pretty soon, there's gonna be more synthetic images on the internet than real images. This was about a year and a half ago, and he said that there are going to be more synthetic websites than real websites. There's going to be more fake text than real text.

**More on AI:** *AI Developers Are Already Quietly Training AI Models Using AI-Generated Data*

SHARE THIS ARTICLE

READ THIS NEXT

KICKIN' ASS AND MELTING EGGS

Google's Search AI Says You Can Melt Eggs. Its Source Will Make You Facepalm

INSULT TO INJURY

Gizmodo Is Stripping Journalists' Bylines Off AI-Translated Articles

TRUST ISSUES