'Intersectional hallucinations': Why AI struggles to understand that a six-year-old can't be a doctor or claim a pension

July 31 2024, by Ericka Johnson

Credit: Pixabay/CC0 Public Domain

When you go to the hospital and get a blood test, the results are put in a dataset and compared with other patients' results and population data.

This lets doctors compare you (your blood, age, sex, health history, scans, etc) to other patients' results and histories, allowing them to predict, manage and develop new treatments.

For centuries, this has been the bedrock of scientific research: identify a problem, gather data, look for patterns, and build a model to solve it. The hope is that Artificial Intelligence (AI) —the kind called Machine Learning that makes models from data—will be able to do this far more quickly, effectively and accurately than humans.

However, training these AI models needs a LOT of data, so much that some of it has to be synthetic—not real data from real people, but data that reproduces existing patterns. Most synthetic datasets are themselves generated by Machine Learning AI.

Wild inaccuracies from image generators and chatbots are easy to spot, but synthetic data also produces hallucinations—results that are unlikely, biased, or plain impossible. As with images and text, they can be amusing, but the widespread use of these systems in all areas of public life means that the potential for harm is massive.

What is synthetic data?

AI models need much more data than the real world can offer. Synthetic data provides a solution—generative AI that examines the statistical distributions in a real dataset and creates a new, synthetic one to train other AI models.

This synthetic "pseudo" data is similar but not identical to the original, meaning it can also ensure privacy, skirt data regulations, and be freely shared or distributed.

Synthetic data can also supplement real datasets, making them big

enough to train an AI system. Or, if a real dataset is biased (has too few women, for example, or over-represents cardigans instead of pullovers), synthetic data can balance it out. There is ongoing debate around how far synthetic data can stray from the original.

Glaring omissions

Without proper curation, the tools that make synthetic data will always over-represent things that are already dominant in a dataset and under-represent (or even omit) less common "edge-cases".

This was what initially sparked my interest in synthetic data. Medical research already under-represents women and other minorities, and I was concerned that synthetic data would exacerbate this problem. So, I teamed up with a machine learning scientist, Dr. Saghi Hajisharif, to explore the phenomenon of disappearing edge-cases.

In our research, we used a type of AI called a GAN to create synthetic versions of 1990 US adult census data. As expected, edge-cases were missing in the synthetic datasets. In the original data we had 40 countries of origin, but in a synthetic version, there were only 31—the synthetic data left out immigrants from 9 countries.

Once we knew about this error, we were able to tweak our methods and include them in a new synthetic dataset. It was possible, but only with careful curation.

'Intersectional hallucinations'—AI creates impossible data

We then started noticing something else in the data—intersectional hallucinations.

Intersectionality is a concept in gender studies. It describes power dynamics that produce discrimination and privilege for different people in different ways. It looks not just at gender, but also at age, race, class, disability, and so on, and how these elements 'intersect' in any situation.

This can inform how we analyze synthetic data—all data, not just population data—as the intersecting aspects of a dataset produce complex combinations of whatever that data is describing.

In our synthetic dataset, the statistical representation of separate categories was quite good. Age distribution, for example, was similar in the synthetic data to the original. Not identical, but close. This is good because synthetic data should be similar to the original, not reproduce it exactly.

Then we analyzed our synthetic data for intersections. Some of the more complex intersections were being reproduced, too. For example, in our synthetic dataset, the intersection of *age-income-gender* was reproduced quite accurately. We called this accuracy "intersectional fidelity."

But we also noticed the synthetic data had 333 datapoints labeled "husband/wife and single"—an intersectional hallucination. The AI had not learned (or been told) that this is impossible. Of these, more than 100 datapoints were "never-married-husbands earning under 50,000 USD a year", an intersectional hallucination that did not exist in the original data.

On the other hand, the original data included multiple "widowed females working in tech support", but they were completely absent from the synthetic version.

This means that our synthetic dataset could be used for research on age-income-gender questions (where there was intersectional fidelity) but not

if one were interested in "widowed females working in tech support". And one should watch out for "never-married-husbands" in the results.

The big question is: where does this stop? These hallucinations are 2-part and 3-part intersections, but what about 4-part intersections? Or 5-part? At what point (and for what purposes) would the synthetic data become irrelevant, misleading, useless or dangerous?

Embracing intersectional hallucinations

Structured datasets exist because the relationships between the columns on a spreadsheet tell us something useful. Remember the blood test. Doctors want to know how your blood compares to normal blood, and to other diseases and treatment outcomes. That is why we organize data in the first place, and have done for centuries.

However, when we use synthetic data, intersectional hallucinations are always going to happen because the synthetic data must be slightly different to the original, otherwise it would simply be a copy of the original data. Synthetic data therefore *requires* hallucinations, but only the right kind—ones that amplify or expand the dataset, but do not create something impossible, misleading or biased.

The existence of intersectional hallucinations means that one synthetic dataset cannot work for lots of different uses. Each use-case will need bespoke synthetic datasets with labeled hallucinations, and this needs a recognized system.

Building reliable AI systems

For AI to be trustworthy, we have to know which intersectional hallucinations exist in its training data, especially when it is used to

predict how people will act, or to regulate, govern, treat or police us. We need to ensure they are not trained on dangerous or misleading intersectional hallucinations—like a six-year-old medical doctor receiving pension payments.

But what happens when synthetic datasets are used carelessly? Right now there is no standard way to mark them, and they are often mixed up with real data. When a dataset is shared for others to use, it is impossible to know if it can be trusted, and to know what is a hallucination and what is not. We need clear, universally recognizable ways to identify synthetic data.

Intersectional hallucinations may not be as amusing as a hand with 15 fingers, or recommendations to put glue on a pizza. They are boring, unsexy numbers and statistics, but they will affect us all—sooner or later, synthetic data is going to spread everywhere, and it will always, by its very nature, contain intersectional hallucinations. Some we want, some we don't, but the problem is telling them apart. We need to make this possible before it is too late.

Provided by The Conversation

Citation: 'Intersectional hallucinations': Why AI struggles to understand that a six-year-old can't be a doctor or claim a pension (2024, July 31) retrieved 7 August 2024 from https://techxplore.com/news/2024-07-intersectional-hallucinations-ai-struggles-year.html