When fake data is a good thing – how synthetic data trains AI to solve real problems

Ambuj Tewari, Professor of Statistics, University of Michigan

Published: November 18, 2025 8:17am EDT

These faces are fake – generated by artificial intelligence – but useful for training other AI systems about human faces.

David Beniaguev

You've just finished a strenuous hike to the top of a mountain. You're exhausted but elated. The view of the city below is gorgeous, and you want to capture the moment on camera. But it's already quite dark, and you're not sure you'll get a good shot. Fortunately, your phone has an AI-powered night mode that can take stunning photos even after sunset.

Here's something you might not know: That night mode may have been trained on synthetic nighttime images, computer-generated scenes that were never actually photographed.

As artificial intelligence researchers exhaust the supply of real data on the web and in digitized archives, they are increasingly turning to synthetic data, artificially generated examples that mimic real ones. But that creates a paradox. In science, making up data is a cardinal sin. Fake data and misinformation are already undermining trust in information online. So how can synthetic data possibly be good? Is it just a polite euphemism for deception?

As a machine learning researcher, I think the answer lies in intent and transparency. Synthetic data is generally not created to manipulate results or mislead people. In fact, ethics may require AI companies to use synthetic data: Releasing real human face images, for example, can violate privacy, whereas synthetic faces can offer similar benefit with formal privacy guarantees.

There are other reasons that help explain the growing use of synthetic data in training AI models. Some things are so scarce or rare that they are barely represented in real data. Rather than letting these gaps become an Achilles' heel, researchers can simulate those situations instead.

Another motivation is that collecting real data can be costly or even risky. Imagine collecting data for a self-driving car during storms or on unpaved roads. It is often much more efficient, and far safer, to generate such data virtually.

How synthetic data is made

Training an AI model requires large amounts of data. Like students and athletes, the more an AI is trained, the better its performance tends to be. Researchers have known for a long time that if data is in short supply, they can use a technique known as data augmentation. For example, a given image can be rotated or scaled to yield additional training data. Synthetic data is data augmentation on steroids. Instead of making small alterations to existing images, researchers create entirely new ones.

But how do researchers create synthetic data? There are two main approaches. The first approach relies on rule-based or physics-based models. For example, the laws of optics can be used to simulate how a scene would appear given the positions and orientations of objects within it.

The second approach uses generative AI to produce data. Modern generative models are trained on vast amounts of data and can now create remarkably realistic text, audio, images and videos. Generative AI offers a flexible way to produce large and diverse datasets.

Both approaches share a common principle: If data does not come directly from the real world, it must come from a realistic model of the world.

Downsides and dangers

It is also important to remember that while synthetic data can be useful, it is not a panacea. Synthetic data is only as reliable as the models of reality it comes from, and even the best scientific or generative models have weaknesses.

Researchers have to be careful about potential biases and inaccuracies in the data they produce. For example, researchers may simulate the home-insurance ecosystem to help detect fraud, but those simulations could embed unfair assumptions about neighborhoods or property types. The benefits of such data must be weighed against risks to fairness and equity.

It's also important to maintain a clear distinction between models and simulations on one hand and the real world on the other. Synthetic data is invaluable for training and testing AI systems, but when an AI model is deployed in the real world, its performance and safety should be proved with real, not simulated, data for both technical and ethical reasons.

Future research on synthetic data in AI is likely to face many challenges. Some are ethical, some are scientific, and others are engineering problems. As synthetic data becomes more realistic, it will be more useful for training AI, but it will also be easier to misuse. For example, increasingly realistic synthetic images can be used to create convincing deepfake videos.

I believe that researchers and AI companies should keep clear records to show which data is synthetic and why it was created. Clearly disclosing which parts of the training data are real and which are synthetic is a key aspect of responsibly producing AI models. California's law, "Generative artificial intelligence: training data transparency," set to take effect on Jan. 1, 2026, requires AI developers to disclose if they used synthetic data in training their models.

Researchers should also study how mistakes in simulations or models can lead to bad data. Careful work will help keep synthetic data transparent, trustworthy and reliable.

Keeping it real

Most AI systems learn by finding patterns in data. Researchers can improve their ability to do this by adding synthetic data. But AI has no sense of what is real or true. The desire to stay in touch with reality and to seek truth belongs to people, not machines. Human judgment and oversight in the use of synthetic data will remain essential for the future.

The next time you use a cool AI feature on your smartphone, think about whether synthetic data might have played a role. Our AIs may learn from synthetic data, but reality remains the ultimate source of our knowledge and the final judge of our creations.

Ambuj Tewari receives funding from NSF and NIH.