

ASU researcher pushes standards for AI-generated media

'YZ' Yang is helping unite researchers, tech companies and policy groups to create reliable ways to detect and even forget AI-generated images

By Kelly deVos, ASU News
April 7, 2026

Flip a coin. Heads: The photo you're looking at is real. Tails: It's been generated by artificial intelligence, or AI. Either way, you would probably be guessing, and that's the problem.

Last year, a [research study](#) published in [Communications of the Association for Computing Machinery](#), a magazine by one of the largest and most influential professional organizations for computer scientists, found that people could distinguish AI-generated media from authentic content only about 51% of the time, roughly the same accuracy as random chance. As generative tools improve, the human eye is quickly losing the ability to tell what's real online and what's fabricated.

The consequences are already visible. Online retailers are experiencing a [surge of fraudulent](#) returns using AI-generated images. [Deepfake-related financial losses](#) exceeded \$200 million in just three months of 2025.

While the tools that create synthetic media are advancing quickly, the systems to verify them are still fragmented or missing entirely.

At Arizona State University, [Yezhou "YZ" Yang](#) is working to change that.

Setting standards for AI-generated content

In the [School of Computing and Augmented Intelligence](#), part of the [Ira A. Fulton Schools of Engineering](#) at ASU, Yang is helping lead efforts to develop technical standards that make AI-generated media identifiable.

The idea is simple in concept: require generative AI systems to embed detectable signals, similar to digital fingerprints, directly into the content they produce.

“It’s like a wireless protocol,” Yang says. “If everyone agrees to the protocol, then every model generating images would embed something like a watermark that detectors can read later.”

Yang’s team began studying this problem as early as 2020, focusing on subtle statistical patterns left behind by generative models, patterns invisible to humans but detectable by machines.

“There are digital traces,” he says. “Humans can’t see them, but computers can.”

But those traces are becoming harder to find. As models improve, the detection problem risks becoming an ongoing technological arms race. That realization pushed Yang to look beyond detection alone.

From detection to correction

If detection is about identifying AI-generated content after it appears, a newer line of research asks a deeper question: What if AI systems could correct themselves?

That’s where Yang’s work on machine unlearning comes in.

Machine unlearning focuses on teaching AI systems to selectively forget specific data, concepts or behaviors — whether that’s copyrighted material, sensitive personal data or harmful content. Instead of retraining massive models from scratch, which can take months and cost millions, unlearning methods target and remove unwanted information directly.

“Whatever data is learned — the good and the bad — it sticks,” Yang says. “Unlearning gives us a way to go back and fix that.”

In the Fulton Schools, Yang’s group has been among the early contributors applying unlearning techniques to text-to-image models, an area that has received far less attention than large language models.

In one recent project, his team developed a method called Robust Adversarial Concept Erasure, or RACE, designed to remove sensitive concepts, such as explicit imagery, from generative models while resisting attempts to bring them back through adversarial prompts.

The work addresses a key weakness in earlier approaches. Even when models are trained to eliminate a concept, users can often recover it with cleverly crafted prompts. Yang’s method strengthens that erasure by anticipating and blocking those attempts.

A second project, EraseFlow, takes the idea further by treating unlearning as a process of reshaping how an AI model generates images over time. Instead of simply blocking outputs, the system redirects the model away from unwanted concepts while preserving overall image quality.

Together, these approaches point toward a future where AI systems are not only transparent, but also editable after deployment — a capability with major implications for privacy, safety and regulation.

Unlearning could help companies comply with laws like the “right to be forgotten,” remove copyrighted material when licenses expire, or eliminate harmful biases discovered after a model is

released.

Building global consensus

At the same time, Yang is working to ensure these technical advances don't remain isolated in research labs.

His group collaborates with initiatives like the [Coalition for Content Provenance and Authenticity](#) and organizations such as the [World Privacy Forum](#), helping shape international conversations around AI transparency, governance and data rights.

The goal is to create shared standards, not just for detecting AI-generated media, but for how systems should behave across their entire lifecycle.

"The technology starts with computer scientists," Yang says. "But the impact on society requires a much bigger conversation."

Why it matters

As AI-generated media becomes more realistic and more widespread, the challenge is no longer just identifying what's fake. It's maintaining trust in an environment where anything can be fabricated or altered after the fact.

For Yang, solving that problem will require both sides of the equation: systems that can identify synthetic content and systems that can adapt, correct and improve themselves over time.

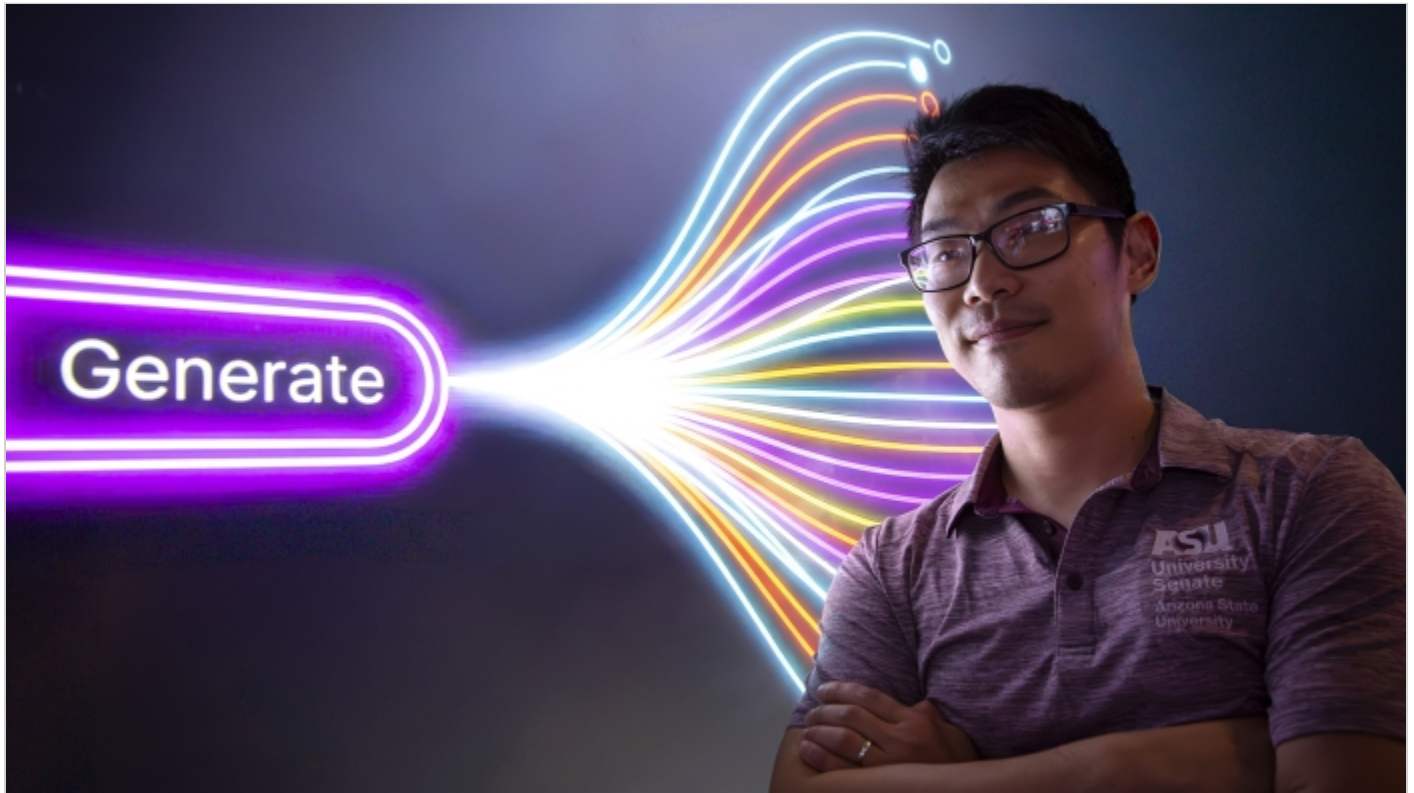
"At some point, society will have to solve this," he says. "We can't have a world where anyone can generate convincing fake evidence."

[Ross Maciejewski](#), director of the School of Computing and Augmented Intelligence, says that's exactly why work like Yang's is critical.

"Addressing the risks of AI isn't just a technical problem. It's a societal one," Maciejewski says. "Our school is uniquely positioned to bring together the research, policy and real-world partnerships needed to tackle these issues. YZ's work exemplifies how we're helping lead important conversations while developing solutions that can scale."

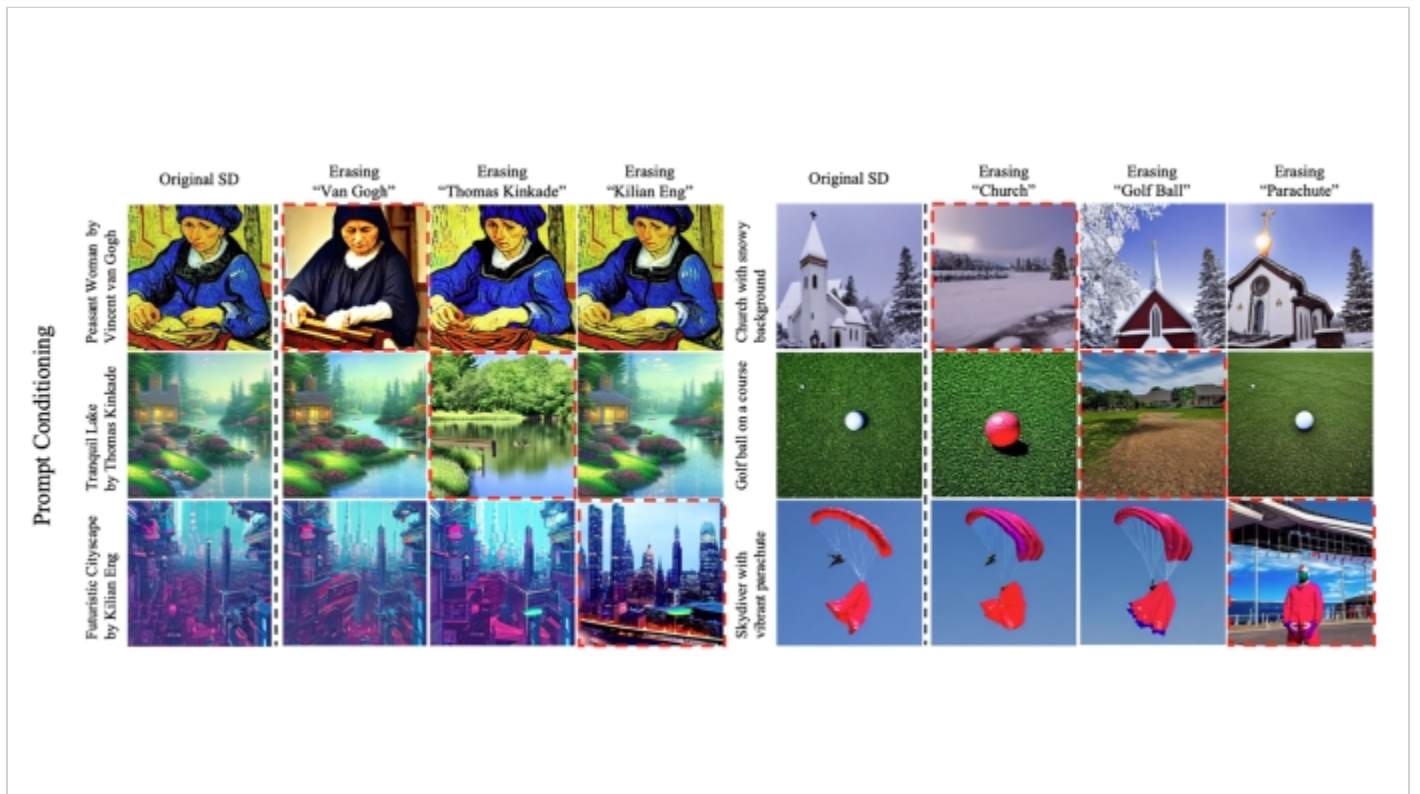
This story originally appeared on [ASU News](#).

Main image



Yezhou “YZ” Yang poses in front of a screen displaying an image generation prompt. Yang is an associate professor of computer science and engineering in the School of Computing and Augmented Intelligence, part of the Ira A. Fulton Schools of Engineering at Arizona State University, who is working to build a global consensus on artificial intelligence, or AI, privacy and ethical standards. Photographer: Erika Gronek/ASU

Text image(s)



A figure demonstrating how Yang and his team can teach AI models to forget specific content. The highlighted images show targeted elements being removed, while the remaining images show that everything else stays the same. The original model's outputs are included for comparison. Figure courtesy of the Active Perception Group