ARTIFICIAL INTELLIGENCE

The AI lab waging a guerrilla war over exploitative AI

The tools Glaze and Nightshade are giving artists hope that they can fight back against AI that hoovers internet data to train. Are they enough?

By Melissa Heikkilä

November 13, 2024

DIMA KASHTALYAN

Ben Zhao remembers well the moment he officially jumped into the fight between artists and generative AI: when one

artist asked for AI bananas.

A computer security researcher at the University of Chicago, Zhao had made a name for himself by building tools to protect images from facial recognition technology. It was this work that caught the attention of Kim Van Deun, a fantasy illustrator who invited him to a Zoom call in November 2022 hosted by the Concept Art Association, an advocacy organization for artists working in commercial media.

On the call, artists shared details of how they had been hurt by the generative AI boom, which was then brand new. At that moment, AI was suddenly everywhere. The tech community was buzzing over image-generating AI models, such as Midjourney, Stable Diffusion, and OpenAI's DALL-E 2, which could follow simple word prompts to depict fantasylands or whimsical chairs made of avocados.

But these artists saw this technological wonder as a new kind of theft. They felt the models were effectively stealing and replacing their work. Some had found that their art had been scraped off the internet and used to train the models, while others had discovered that their own names had become prompts, causing their work to be drowned out online by AI knockoffs.

Zhao remembers being shocked by what he heard. "People are literally telling you they're losing their livelihoods," he told me one afternoon this spring, sitting in his Chicago living room. "That's something that you just can't ignore."

So on the Zoom, he made a proposal: What if, hypothetically, it was possible to build a mechanism that would help mask their art to interfere with AI scraping?

"I would love a tool that if someone wrote my name and made a prompt, like, garbage came out," responded Karla Ortiz, a prominent digital artist. "Just, like, bananas or some weird stuff."

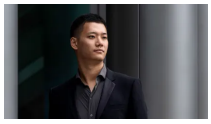
That was all the convincing Zhao needed—the moment he joined the cause.

Fast-forward to today, and millions of artists have deployed

two tools born from that Zoom: Glaze and Nightshade, which were developed by Zhao and the University of Chicago's SAND Lab (an acronym for "security, algorithms, networking, and data").

Arguably the most prominent weapons in an artist's arsenal against nonconsensual AI scraping, Glaze and Nightshade work in similar ways: by adding what the researchers call "barely perceptible" perturbations to an image's pixels so that machine-learning models cannot read them properly. Glaze, which has been downloaded more than 6 million times since it launched in March 2023, adds what's effectively a secret cloak to images that prevents AI algorithms from picking up on and copying an artist's style. Nightshade, which I wrote about when it was released almost exactly a year ago this fall, cranks up the offensive against AI companies by adding an invisible layer of poison to images, which can break AI models; it has been downloaded more than 1.6 million times.

RELATED STORY



2024 Innovator of the Year: Shawn Shan builds tools to help artists fight back against exploitative AI

Thanks to the tools, "I'm able to post my work online," Ortiz says, "and that's pretty huge." For artists like her, being seen online is crucial to getting more work. If they are uncomfortable about ending up in a massive for-profit AI model without compensation, the only option is to delete their work from the internet. That would mean career suicide. "It's really dire for us," adds Ortiz, who has become one of the most vocal advocates for fellow artists and is part of a class action lawsuit against AI companies, including Stability AI, over copyright infringement.

But Zhao hopes that the tools will do more than empower individual artists. Glaze and Nightshade are part of what he sees as a battle to slowly tilt the balance of power from large corporations back to individual creators.

"It is just incredibly frustrating to see human life be valued so little," he says with a disdain that I've come to see as pretty typical for him, particularly when he's talking about Big Tech. "And to see that repeated over and over, this prioritization of profit over humanity … it is just incredibly frustrating and maddening."

As the tools are adopted more widely, his lofty goal is being

put to the test. Can Glaze and Nightshade make genuine security accessible for creators—or will they inadvertently lull artists into believing their work is safe, even as the tools themselves become targets for haters and hackers? While experts largely agree that the approach is effective and Nightshade could prove to be powerful poison, other researchers claim they've already poked holes in the protections offered by Glaze and that trusting these tools is risky.

But Neil Turkewitz, a copyright lawyer who used to work at the Recording Industry Association of America, offers a more sweeping view of the fight the SAND Lab has joined. It's not about a single AI company or a single individual, he says: "It's about defining the rules of the world we want to inhabit."

Poking the bear

The SAND Lab is tight knit, encompassing a dozen or so researchers crammed into a corner of the University of Chicago's computer science building. That space has accumulated somewhat typical workplace detritus—a Meta

Quest headset here, silly photos of dress-up from Halloween parties there. But the walls are also covered in original art pieces, including a framed painting by Ortiz.

Years before fighting alongside artists like Ortiz against "AI bros" (to use Zhao's words), Zhao and the lab's co-leader, Heather Zheng, who is also his wife, had built a record of combating harms posed by new tech.

When I visited the SAND Lab in Chicago, I saw how tight knit the group was. Alongside the typical workplace stuff were funny Halloween photos like this one. (Front row: Ronik Bhaskar, Josephine Passananti, Anna YJ Ha, Zhuolin Yang, Ben Zhao, Heather Zheng. Back row: Cathy Yuanchen Li, Wenxin Ding, Stanley Wu, and Shawn Shan.)

COURTESY OF SAND LAB

Though both earned spots on *MIT Technology Review*'s 35 Innovators Under 35 list for other work nearly two decades ago, when they were at the University of California, Santa Barbara (Zheng in 2005 for "cognitive radios" and Zhao a year later for peer-to-peer networks), their primary research focus has become security and privacy.

The pair left Santa Barbara in 2017, after they were poached by the new co-director of the University of Chicago's Data Science Institute, Michael Franklin. All eight PhD students from their UC Santa Barbara lab decided to follow them to Chicago too. Since then, the group has developed a "bracelet of silence" that jams the microphones in AI voice assistants like the Amazon Echo. It has also created a tool called Fawkes—"privacy armor," as Zhao put it in a 2020 interview with the *New York Times*—that people can apply to their photos to protect them from facial recognition software. They've also studied how hackers might steal sensitive information through stealth attacks on virtual-reality headsets, and how to distinguish human art from AI-generated images.

“Ben and Heather and their group are kind of unique because they’re actually trying to build technology that hits right at some key questions about AI and how it is used,” Franklin tells me. “They’re doing it not just by asking those questions, but by actually building technology that forces those questions to the forefront.”

It was Fawkes that intrigued Van Deun, the fantasy illustrator, two years ago; she hoped something similar might work as protection against generative AI, which is why she extended that fateful invite to the Concept Art Association’s Zoom call.

That call started something of a mad rush in the weeks that followed. Though Zhao and Zheng collaborate on all the lab’s projects, they each lead individual initiatives; Zhao took on what would become Glaze, with PhD student Shawn Shan (who was on this year’s Innovators Under 35 list) spearheading the development of the program’s algorithm.

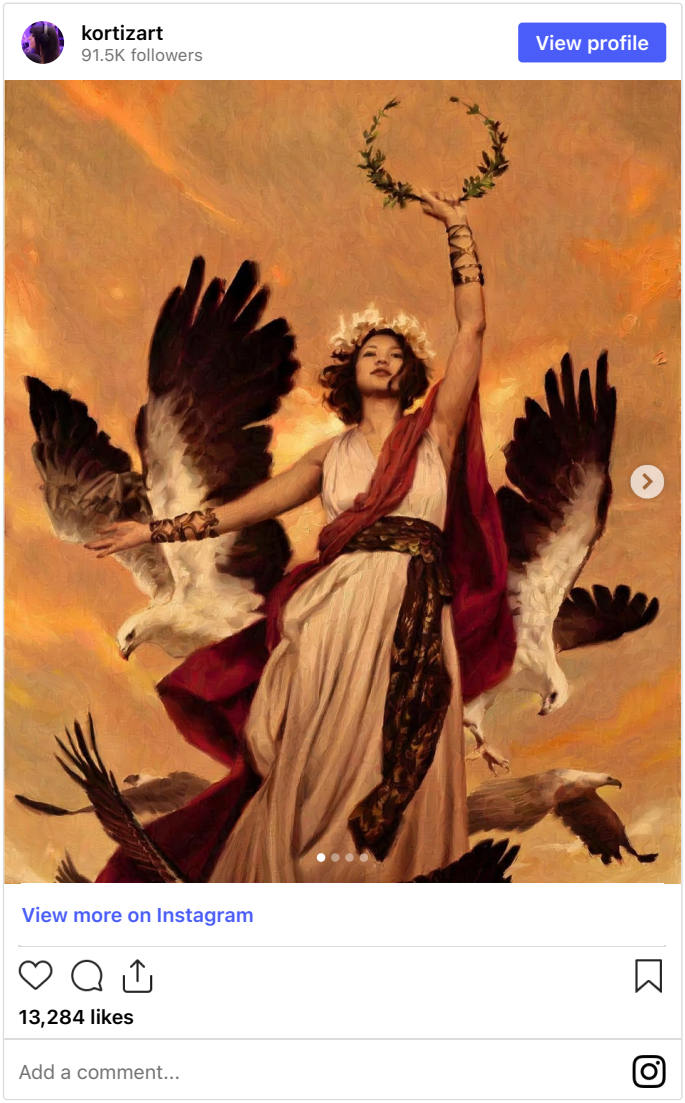
In parallel to Shan’s coding, PhD students Jenna Cryan and Emily Wenger sought to learn more about the views and needs of the artists themselves. They created a user survey that the team distributed to artists with the help of Ortiz. In replies from more than 1,200 artists—far more than the average number of responses to user studies in computer science—the team found that the vast majority of creators had read about art being used to train models, and 97% expected AI to decrease some artists’ job security. A quarter said AI art had already affected their jobs.

Almost all artists also said they posted their work online, and more than half said they anticipated reducing or removing that online work, if they hadn’t already—no matter the professional and financial consequences.

The first scrappy version of Glaze was developed in just a month, at which point Ortiz gave the team her entire catalogue of work to test the model on. At the most basic level, Glaze acts as a defensive shield. Its algorithm identifies features from the image that make up an artist's individual style and adds subtle changes to them. When an AI model is trained on images protected with Glaze, the model will not be able to reproduce styles similar to the original image.

A painting from Ortiz later became the first image publicly released with Glaze on it: a young woman, surrounded by flying eagles, holding up a wreath. Its title is *Musa Victoriosa*, "victorious muse."

It's the one currently hanging on the SAND Lab's walls.



Despite many artists' initial enthusiasm, Zhao says, Glaze's launch caused significant backlash. Some artists were skeptical because they were worried this was a scam or yet another data-harvesting campaign.

The lab had to take several steps to build trust, such as offering the option to download the Glaze app so that it adds the protective layer offline, which meant no data was being transferred anywhere. (The images are then shielded when artists upload them.)

Soon after Glaze's launch, Shan also led the development of the second tool, Nightshade. Where Glaze is a defensive mechanism, Nightshade was designed to act as an *offensive* deterrent to nonconsensual training. It works by changing the pixels of images in ways that are not noticeable to the human eye but manipulate machine-learning models so they interpret the image as something different from what it actually shows. If poisoned samples are scraped into AI training sets, these samples trick the AI models: Dogs become cats, handbags become toasters. The researchers

say only a relatively few examples are enough to permanently damage the way a generative AI model produces images.

Currently, both tools are available as free apps or can be applied through the project's website. The lab has also recently expanded its reach by offering integration with the new artist-supported social network Cara, which was born out of a backlash to exploitative AI training and forbids AI-produced content.

In dozens of conversations with Zhao and the lab's researchers, as well as a handful of their artist-collaborators, it's become clear that both groups now feel they are aligned in one mission. "I never expected to become friends with scientists in Chicago," says Eva Toorenent, a Dutch artist who worked closely with the team on Nightshade. "I'm just so happy to have met these people during this collective battle."

Images online of Toorenent's *Belladonna* have been treated with the SAND Lab's Nightshade tool.

EVA TOORENENT

Her painting *Belladonna*, which is also another name for the nightshade plant, was the first image with Nightshade's poison on it.

"It's so symbolic," she says. "People taking our work without our consent, and then taking our work without consent can ruin their models. It's just poetic justice."

No perfect solution

The reception of the SAND Lab's work has been less harmonious across the AI community.

After Glaze was made available to the public, Zhao tells me, someone reported it to sites like VirusTotal, which tracks malware, so that it was flagged by antivirus programs. Several people also started claiming on social media that the tool had quickly been broken. Nightshade similarly got a fair share of criticism when it launched; as *TechCrunch* reported in January, some called it a "virus" and, as the story explains, "another Reddit user who inadvertently went viral on X questioned Nightshade's legality, comparing it to 'hacking a vulnerable computer system to disrupt its operation.'"

"We had no idea what we were up against," Zhao tells me. "Not knowing who or what the other side could be meant that every single new buzzing of the phone meant that maybe someone did break Glaze."

Both tools, though, have gone through rigorous academic peer review and have won recognition from the computer security community. Nightshade was accepted at the IEEE Symposium on Security and Privacy, and Glaze received a distinguished paper award and the 2023 Internet Defense Prize at the Usenix Security Symposium, a top conference in the field.

"In my experience working with poison, I think [Nightshade is] pretty effective," says Nathalie Baracaldo, who leads the AI security and privacy solutions team at IBM and has studied data poisoning. "I have not seen anything yet—and the word *yet* is important here—that breaks that type of defense that Ben is proposing." And the fact that the team has released the source code for Nightshade for others to probe, and it hasn't been broken, also suggests it's quite secure, she adds.

At the same time, at least one team of researchers does claim to have penetrated the protections of Glaze, or at least an old version of it.

As researchers from Google DeepMind and ETH Zurich detailed in a paper published in June, they found various ways Glaze (as well as similar but less popular protection tools, such as Mist and Anti-DreamBooth) could be circumvented using off-the-shelf techniques that anyone could access—such as image upscaling, meaning filling in pixels to increase the resolution of an image as it's enlarged. The researchers write that their work shows the "brittleness of existing protections" and warn that "artists may believe they are effective. But our experiments show *they are not*."

Florian Tramèr, an associate professor at ETH Zurich who was part of the study, acknowledges that it is "very hard to come up with a strong technical solution that ends up really making a difference here." Rather than any individual tool, he ultimately advocates for an almost certainly unrealistic ideal: stronger policies and laws to help create an environment in which people commit to buying only human-created art.

What happened here is common in security research, notes Baracaldo: A defense is proposed, an adversary breaks it, and—ideally—the defender learns from the adversary and makes the defense better. "It's important to have both ethical attackers and defenders working together to make our AI systems safer," she says, adding that "ideally, all defenses should be publicly available for scrutiny," which would both "allow for transparency" and help avoid creating a false sense of security. (Zhao, though, tells me the researchers have no intention to release Glaze's source code.)

Still, even as all these researchers claim to support artists and their art, such tests hit a nerve for Zhao. In Discord chats that were later leaked, he claimed that one of the researchers from the ETH Zurich–Google DeepMind team "doesn't give a shit" about people. (That researcher did not respond to a request for comment, but in a blog post he said it was important to break defenses in order to know how to fix them. Zhao says his words were taken out of context.)

RELATED STORY

This new data poisoning tool lets artists fight back against generative AI

Zhao also emphasizes to me that the paper's authors mainly evaluated an earlier version of Glaze; he says its new update is more resistant to tampering. Messing with images that have current Glaze protections would harm the very style that is being copied, he says, making such an attack useless.

This back-and-forth reflects a significant tension in the computer security community and, more broadly, the often adversarial relationship between different groups in AI. Is it wrong to give people the feeling of security when the protections you've offered might break? Or is it better to have *some* level of protection—one that raises the threshold for an attacker to inflict harm—than nothing at all?

Yves-Alexandre de Montjoye, an associate professor of applied mathematics and computer science at Imperial College London, says there are plenty of examples where similar technical protections have failed to be bulletproof. For example, in 2023, de Montjoye and his team probed a digital mask for facial recognition algorithms, which was meant to protect the privacy of medical patients' facial images; they were able to break the protections by tweaking just one thing in the program's algorithm (which was open source).

Using such defenses is still sending a message, he says, and adding some friction to data profiling. "Tools such as TrackMeNot"—which protects users from data profiling—"have been presented as a way to protest; as a way to say *I do not consent*."

"But at the same time," he argues, "we need to be very clear with artists that it is removable and might not protect against future algorithms."

While Zhao will admit that the researchers pointed out some of Glaze's weak spots, he unsurprisingly remains confident that Glaze and Nightshade are worth deploying, given that "security tools are never perfect." Indeed, as Baracaldo points out, the Google DeepMind and ETH Zurich researchers showed how a highly motivated and sophisticated adversary will almost certainly always find a way in.

Yet it is "simplistic to think that if you have a real security problem in the wild and you're trying to design a protection tool, the answer should be it either works perfectly or don't deploy it," Zhao says, citing spam filters and firewalls as examples. Defense is a constant cat-and-mouse game. And

he believes most artists are savvy enough to understand the risk.

Offering hope

The fight between creators and AI companies is fierce. The current paradigm in AI is to build bigger and bigger models, and there is, at least currently, no getting around the fact that they require vast data sets hoovered from the internet to train on. Tech companies argue that anything on the public internet is fair game, and that it is "impossible" to build advanced AI tools without copyrighted material; many artists argue that tech companies have stolen their intellectual property and violated copyright law, and that they need ways to keep their individual works out of the models—or at least receive proper credit and compensation for their use.

So far, the creatives aren't exactly winning. A number of companies have already replaced designers, copywriters, and illustrators with AI systems. In one high-profile case, Marvel Studios used AI-generated imagery instead of human-created art in the title sequence of its 2023 TV series *Secret Invasion*. In another, a radio station fired its human presenters and replaced them with AI. The technology has become a major bone of contention between unions and film, TV, and creative studios, most recently leading to a strike by video-game performers. There are numerous ongoing lawsuits by artists, writers, publishers, and record labels against AI companies. It will likely take

That's why Zhao and Zheng see Glaze and Nightshade as necessary interventions—tools to defend original work, attack those who would help themselves to it, and, at the very least, buy artists some time. Having a perfect solution is not really the point. The researchers need to offer *something* now because the AI sector moves at breakneck speed, Zheng says, means that companies are ignoring very real harms to humans. "This is probably the first time in our entire technology careers that we actually see this much conflict," she adds.

On a much grander scale, she and Zhao tell me they hope that Glaze and Nightshade will eventually have the power to overhaul how AI companies use art and how their products produce it. It is eye-wateringly expensive to train

POPULAR

A fundamental flaw leaves LLMs strikingly vulnerable to attack

Will Douglas Heaven

AI models, and it's extremely laborious for engineers to find and purge poisoned samples in a data set of billions of images. Theoretically, if there are enough Nightshaded images on the internet and tech companies see their models breaking as a result, it could push developers to the negotiating table to bargain over licensing and fair compensation.

That's, of course, still a big "if." *MIT Technology Review* reached out to several AI companies, such as Midjourney and Stability AI, which did not reply to requests for comment. A spokesperson for OpenAI, meanwhile, did not confirm any details about encountering data poison but said the company takes the safety of its products seriously and is continually improving its safety measures: "We are always working on how we can make our systems more robust against this type of abuse."

In the meantime, the SAND Lab is moving ahead and looking into funding from foundations and nonprofits to keep the project going. They also say there has also been interest from major companies looking to protect their intellectual property (though they decline to say which), and Zhao and Zheng are exploring how the tools could be applied in other industries, such as gaming, videos, or music. In the meantime, they plan to keep updating Glaze and Nightshade to be as robust as possible, working closely with the students in the Chicago lab—where, on another wall, hangs Toorenent's *Belladonna*. The painting has a heart-shaped note stuck to the bottom right corner: "Thank you! You have given hope to us artists."

This story has been updated with the latest download figures for Glaze and Nightshade.

by Melissa Heikkilä

AI is more likely than humans to form biases when hiring

Michelle Kim

Here's why AI agents lie and cheat to reach their goals

Grace Huckins

Bill Gates says we've passed AI's danger thresholds. Now what?

Mat Honan