

Why AI Watermarks and Detectors Could Backfire

By Nadav Ziv | August 25, 2026

Nadav Ziv is a researcher at the Digital Inquiry Group.

Originally published in TIME. [Read the original article.](#)

Claude now [watermarks](#) AI-generated text to comply with European Union transparency rules. [OpenAI](#) and [Google](#) add invisible fingerprints to AI-generated images. And Substack is [touting](#) a feature that scans pieces for signs of AI. Will we finally be able to tell what's real on the Internet? My take: not even close.

In fact, AI watermarks and detectors may leave us worse off by creating a false sense of confidence in content marked as genuine.

Watermarks and detectors are gaining traction as we lose our ability to trust our senses online. Look up the [Will Smith eating spaghetti test](#), and you'll see just how far AI has come. A 2023 AI-generated video shows the actor slurping spaghetti, face distorted, in a way that breaks physics. By 2025, AI was producing lifelike renditions. Deepfakes are so good that experts recommend families [develop](#) secret codewords to identify one another.

"But I know a fake when I see it," someone might say.

Unfortunately, [research](#) consistently shows that you do not. This can feel especially hard to accept given the abundance of AI slop rocketing around the Internet. You may even start to think you can sniff out offending content. It might work, for a little bit. It almost [never lasts](#). Any signal that becomes discernible is one a sophisticated actor will find ways to avoid.

We've seen this story before. During the earliest days of the Internet, visual polish at least told you something. Major institutions had the [resources](#) needed to produce well-designed websites. Janky-looking sites, on the other hand, screamed "scam!" Information experts [directed](#) Internet users to dwell on features such as design, broken links, and typos. But when the Internet changed, the advice didn't.

A [study](#) I led, published in 2022, found that 96% of America's leading colleges and universities offered outdated advice on how to evaluate online information—long after platforms like Wix, Squarespace, and Photoshop made it easier for bad actors to create fake but convincing-looking websites. Inexpensive software made slick graphics ubiquitous. Educators, however, continued to instruct Internet users to search for visual clues like a game of Where's Waldo?

The most dangerous legacy of this aesthetic fixation is the inverse illusion: the cognitive tendency to believe that if the presence of a signal proves one thing, its absence proves the opposite. Yes, a site with misspellings that claims to show aliens still isn't legit. But a beautiful site with a [dot-org domain](#) can also be harmful. In 2019, our research group [found](#) that nearly half of hate groups had dot-org domains. Bad actors know how to adopt the trappings of credibility.

The same is true with AI. Even if visible flaws sometimes linger, their absence doesn't mean content is genuine. Yet, too often, experts offer surface-level clues to identifying AI-generated content. This is why in the lead-up to the 2024 elections, Stanford Professor Sam Wineburg and I [warned](#) about public officials who advised citizens to pay attention to lighting, strange shadows, or other visual cues to identify deepfakes, even after AI content stopped making these errors. Many [2026 guides](#) to spotting AI content mislead readers with the same poor advice.

Which brings us to AI watermarks and detectors. These approaches, based on hidden signals in content, promise that while we can't always spot the signs, their algorithms can.

I'm not a software engineer. Yet I was able to easily strip metadata from some AI-generated images just by screenshotting them. Anthropic [confirms](#) that file metadata can be "stripped through format conversion, re-saving, screenshots, or other means." Watermarks like SynthID are stronger and can persist after screenshots. But I was able to use a free online tool to remove a SynthID watermark.

Google [admits](#) that the accuracy of detecting watermarked AI text is "greatly reduced" when users thoroughly rewrite what they generate, and that it "is not designed to directly stop motivated adversaries from causing harm." More broadly, [open-weight AI models](#) that can run locally, outside platform terms and conditions, guarantee the spread of unmarked content.

Third-party detectors, too, have a spotty track record. I've regularly run AI-generated text through detectors that said it was human and vice versa. [Many studies](#) of text, image, and audio [detectors](#) find that they don't work very consistently, and yet, their [findings](#) are used as the basis for public accusations. Every detector must confront an arms race with humanizer tools and other workarounds motivated actors find.

I would argue that the biggest problem for detectors and watermarks remains the inverse illusion. Just because content lacks a watermark doesn't mean it wasn't produced or edited with AI. As Anthropic [notes](#): "lack of a detected mark doesn't mean the content wasn't AI-generated or processed." Deferring judgment to AI detectors leaves us vulnerable to bad actors who know how to launder content and make it pass muster.

This is a confusing time. Many of us are, understandably, uncertain. In one recent [pilot](#), our research group showed 117 students a confident chatbot answer about local history with hallucinated facts. Half said they weren't sure if it was true. One student said AI is sometimes right and sometimes wrong and "you never know which is which."

But just because we can't trust our eyes or place full faith in detectors doesn't mean we can't trust anything. Rather than hunt for visual clues or outsource judgment to detectors and watermarks, we can turn to reputation and context. It's easy to fake content. It's much harder to fake a good reputation that's validated by credible sources.

The next time you see unfamiliar content online, resist the urge to ask, "Does this look like AI?" or run the content through a detector. Instead, ask yourself, "Do I trust where this information is coming from?" Open a new tab and check if reputable people and organizations confirm what you're seeing.

In an era of dwindling trust, we should not fork over ours to cheap signals or cheap software.