

The rise of AI suspicion

What Reddit reveals about AI detection and the growing suspicion facing human creators

By Kelly deVos, ASU News
October 1, 2026

As images, videos and text generated by artificial intelligence become more common online, people are increasingly being asked to make difficult judgments: [Is this real?](#) Was it made by a person? Can it be trusted?

Those questions are no longer limited to computer scientists or technology companies. They now shape how people encounter news, social media, art and books.

The stakes are especially visible in publishing, where writers and readers are debating the use of AI-detection tools. A [recent analysis claimed](#) that one in five randomly selected Kindle eBooks in a large dataset contained more than 25% AI-generated text, and that the viral dark dystopian romance “Daggermouth” by H.M. Wolfe [was 60% AI-written](#).

But AI detection is itself controversial, raising questions about evidence, false positives and what it means to call a work “AI-like” in a culture already anxious about authenticity.

[Jaron Mink](#) is an assistant 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. Their research focuses on security, privacy, machine learning and trustworthy human interaction with AI systems.

In a [recent paper](#), Mink and collaborators analyzed more than 13,000 posts and 222,000 comments from two Reddit communities devoted to determining whether online media is AI-generated.

In this Q&A, Mink explains what those communities reveal about AI detection, how public strategies for identifying AI-generated content are changing and what the rise of AI suspicion means for writers, artists and anyone trying to decide what is real online.

Note: Answers have been edited for length and/or clarity.

Question: What first made you want to study Reddit communities like [r/isthisAI](#) and [r/RealOrAI](#)? What did you see happening there that felt important enough to examine closely?

Answer: A lot of my work focuses on understanding how people encounter emerging technologies in the real world and what the implications are. During my doctoral research, I began studying user perceptions of deepfake media, and I became fascinated by the way communities were starting to form around detection. People were gathering in online spaces to figure out whether they were

interacting with bots, or whether the images, videos and text they were seeing online were real or AI-generated.

Most of my earlier research on this topic took place in controlled studies, where we asked participants to evaluate specific examples of media. The Reddit communities offered something different: a much more natural look at how people were already making these judgments on their own. Instead of responses gathered in a research setting, we were looking at people's actual thoughts, posted publicly over a long period of time. It seemed like an obvious place to understand, in a broader sense, how people are trying to make sense of what is real and what is fake online, especially as AI-generated images, videos and audio have improved.

Q: In the paper, you and your co-authors identify a whole taxonomy of strategies people use to decide whether something was made with AI. What are the main ways people are trying to make those judgments?

A: One set of strategies is based on people's beliefs about how the real world works. They look for physical details that do not make sense, such as lighting that seems wrong or objects that morph in ways that feel unnatural.

Another strategy is content plausibility. Something may not be physically impossible, but it may seem very unlikely, such as an image of someone making 10,000 cupcakes in a room. People also compare media to what they think AI looks like. That might mean a particular visual style, a cadence in writing or a way that photos seem to be composed.

We also saw people reasoning about motive. Who would use AI, and why? For example, if something looks like an unrealistic product and someone appears to be trying to sell it, people may infer that it is probably fake.

Finally, people look outside the media itself. They ask whether it came from a source they trust, whether other sources confirm it or whether it was posted before AI systems were capable of producing that kind of content.

Q: Were there any strategies that surprised you, either because they were clever, strange or more common than you expected?

A: One thing that surprised me was how much people used social media trends as evidence. That was not something I had really seen before. People would point to trends like doorbell prank videos or unrealistic Airbnb-style videos and say, in effect, "I have seen this pattern on TikTok, and people are copying it."

In those cases, it is not necessarily that the media violates physics or looks technically wrong. Instead, people are drawing on their knowledge of the current social media environment. AI detection has entered the social media echo chamber in a way it had not before, and it was interesting to see people use that cultural context as part of their reasoning.

Q: One thing I found fascinating is that people's detection strategies changed over time. As generative AI tools got better, how did the crowd's methods for spotting AI change?

A: Early on, people might focus on image artifacts or on the fact that videos and audio were not as convincing. As those systems improved, the strategies had to shift.

People started relying less on a single obvious clue and more on broader context, such as whether the content fits a trend, whether the source is trustworthy or whether the timing makes sense.

That is part of what makes this such an interesting moment. People are trying to update their mental models in real time, but the technology is changing quickly, and the signs people learn to look for may stop being useful.

Q: A lot of people seem to look for telltale signs, including weird hands and odd textures in images, unnatural wording in text or just a feeling that something is “off.” How reliable are those instincts, and where do they start to break down?

A: Those instincts can be useful, but they are not reliable enough to treat as proof. People often form ideas about what AI looks or sounds like, but those ideas can be inconsistent or wrong. In writing, for example, people sometimes point to punctuation or a polished style as evidence of AI, even though those are not telltale signs.

Q: Your study also seems to be about more than detection. It’s about how people are debating authenticity together and building shared beliefs in public. How can that kind of collective sensemaking help people, and how can it go wrong?

A: I think the existence of these communities shows that this is a real concern for people. Some people may enjoy the collective deduction aspect, but in other work we are doing, we see that the context can be more serious. People are worried about scams, about a family member falling for something, or about whether a piece of media is being used to deceive them.

Collective sensemaking is a normal human response to uncertainty. People come together to understand a situation, even if they cannot fully solve it. That can be helpful because people bring different knowledge and perspectives. But it can also go wrong when socially formed beliefs become too strong, or when people reinforce one another’s assumptions. If everyone is looking for signs of AI, they may find them even where they do not exist.

Q: Some AI companies are introducing watermarking or provenance systems to help identify AI-generated content. For readers who are not computer scientists, how do those systems work, and what can they prove?

A: There are a few different systems people talk about. One is an invisible watermark. In that case, when an AI system generates content, it does so in a very specific pattern that appears normal to people but can later be detected by the company that made the system. If the content is unaltered, the company may be able to say with a good degree of confidence that it was produced by its system.

But those systems are probabilistic. They are not a secure guarantee. Real content could theoretically be misclassified, and AI-generated content can often be modified so the watermark is harder or impossible to detect.

A different approach is media provenance. That uses cryptographic methods to create a record of where a piece of media came from and how it was modified. In theory, you could see that an image was captured by a certain camera and then edited by particular people or organizations. That can provide much stronger evidence about the history of a file. But if something does not have provenance, that does not prove it is fake. It only means you do not know where it came

from.

It is also important to separate whether something was made with AI from whether it is true. Misinformation existed before AI, and it will exist after AI. AI-generated, fake and misleading forms of media are related issues, but they are not the same thing.

Q: Many writers are now afraid of being accused of using AI. If AI systems are trained on human writing and can imitate an author's style, what does it truly mean when a detector says something sounds "AI-like"? Could original human work be treated as suspicious because AI has learned to mimic human style?

A: I think that fear is very understandable. Human work can absolutely be treated as suspicious because AI has learned patterns from human style. We have seen people incorrectly believe that human-created content was AI-generated, and those beliefs are often socially formed. They can also change depending on what the public thinks AI looks like at a particular moment.

That concerns me because people can use inconsistent evidence to support the same accusation. If a piece of writing is polished, someone may say that is a sign of AI. If it has mistakes, someone else may say that is a sign of AI. Those patterns do not concretely exist the way people sometimes imagine.

I say this as an academic who uses em dashes a lot and now sees them being taken out of papers because of this association. There is a real cost to being seen as nonconforming or suspicious. I think we need a more nuanced conversation. The important question should not only be whether AI was used, but whether it was used in a trustworthy and appropriate way.

Q: In publishing and creative communities, people are starting to talk about AI detectors almost like forensic tools. Books have been questioned, authors have been accused and people are using detection to try to make judgments. What should we understand about the limits of these tools before reputations or careers are put on the line?

A: We should be very cautious about treating AI detectors as forensic tools. In security, even a system that works 99% of the time can still be a problem if someone who understands the system can consistently fool it.

Many AI detectors rely on machine learning to classify patterns that appear more common in human writing or more common in AI-generated writing. If you generate a piece of text and put it directly into a detector, it may classify it accurately. But if someone edits that text, or uses another tool designed to evade detection, it can become much easier to fool the system.

That is why I do not think detection systems are robust enough to serve as conclusive proof, especially when reputations or careers are involved. Provenance-based systems are more promising for long-term trust because they can provide stronger evidence about where content came from. But even then, we are usually talking about degrees of trust, not absolute certainty.

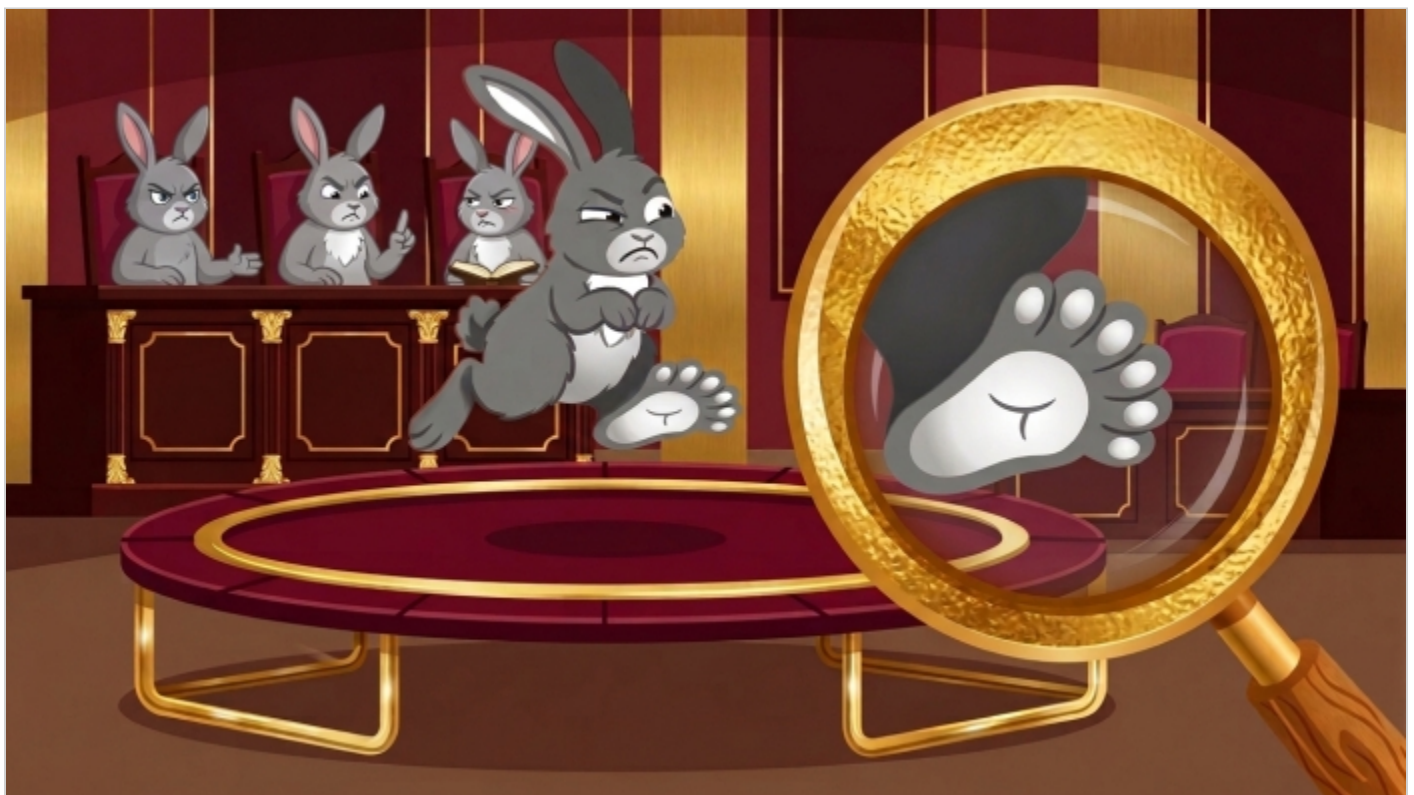
Q: As AI-generated content becomes more common, how do we protect people from deception without creating a culture where everyone is constantly suspicious of everyone else's work?

A: I think people may increasingly care less about whether AI was used at all and more about whether a person, organization or system is trustworthy. Was the work produced responsibly? Is there a verification process? Does the source have a history of producing reliable things? Those questions may matter more than a simple AI-versus-human label.

We are in a strange cultural moment. AI poses real risks, and it is not surprising that people are concerned. But we must be careful in ensuring our responses themselves do not cause unintentional or undue harm, particularly to those already made vulnerable by AI. People already know that not everything online is trustworthy. We may end up treating images, videos and text online with similar caution: not assuming everything is fake, but understanding that context, source and verification matter.

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

Main image



Inspired by the viral “AI Bunnies Jumping on a Trampoline” meme, a drawing shows a disgruntled rabbit with too many toes hopping as a disapproving panel of judges looks on. The meme was one of several examined in new research by Jaron Mink, an assistant professor of computer science and engineering in ASU's School of Computing and Augmented Intelligence, which analyzed online communities dedicated to detecting artificial intelligence use. Graphic generated by Erika Gronek/ASU using Adobe Firefly and Gemini