All CCS and Cross-sectoral Media & Press Greece Artificial Intelligence

An arms race over disinformation: using AI to detect AI

In a world full of deepfakes and fake news, EU-funded researchers are building tools to help journalists tell what's real and what's not.

"We are in a continuous loop of trying to be able to understand and catch up with the latest technology. Riccardo Gallotti, AI4Trust

By Hannah Docter-Loeb

Last winter, as Christmas markets opened across Europe, social media was flooded with alarming videos. Posts claimed that radical Islamists were "invading" Christmas markets.

One clip appeared to show people "disrupting" the opening of the Brussels Christmas market, while a separate photo showed a market surrounded by heavy security. The message was clear: Christian traditions were supposedly under threat.

But the reality was different. The videos came from peaceful demonstrations, and the photo had been generated using AI. What looked convincing at first glance was misleading – or entirely fake.

This is the new information landscape.

According to a recent European Commission survey, nearly two-thirds of respondents said they had encountered disinformation or fake news within the previous week. With AI tools now able to generate highly realistic images, videos and text, it has become harder than ever to differentiate between what's real and what's not.

In response, a multinational team of researchers and media specialists supported by EU funding decided to fight fire with fire.

A first line of defence

In 2020, experts from universities, media houses and technology companies teamed up in a four-year EU-funded initiative called AI4Media. The aim was to create AI tools to help journalists and fact-checkers verify digital content quickly and reliably.

“There is an urgent need to develop AI techniques for the media sector,” said Yiannis Kompatsiaris, research director at the Centre for Research & Technology Hellas (CERTH), who coordinated the initiative.

AI has dramatically lowered the barrier to producing convincing fake content. Anyone with access to generative AI can now create fabricated images, cloned voices or realistic-looking news articles. Social media platforms amplify that content at speed.

“When a fake story is supported by realistic images, it becomes much easier to believe – and more tempting to share because the content generates higher views,” Kompatsiaris added.

The AI4Media team built verification tools designed to fit directly into newsroom workflows. Media organisations such as Deutsche Welle in Germany and VRT in Belgium tested them in real-world settings.

“Fact-checkers and journalists face suspicious images every day,” said Akis Papadopoulos, a researcher at CERTH who worked on the project. He described the technology as a “first line of defence”, not a replacement for human judgement but a way to flag potentially manipulated content quickly.

“It’s important to equip journalists across Europe – and globally – with tools that help them identify suspicious material fast,” he said.

According to the European Digital Media Observatory, an independent, EU-funded hub that monitors disinformation campaigns across all EU countries, AI-generated disinformation has increased steadily in recent months.

And this goes well beyond isolated hoaxes. Coordinated campaigns can influence elections, distort public debate and undermine trust in institutions.

Spotting disinformation patterns

Identifying manipulated content is only part of the challenge. Understanding how disinformation spreads – who amplifies it, how narratives evolve and whether campaigns are coordinated – is just as important.

“We are in a continuous loop of trying to be able to understand and catch up with the latest technology,” said Riccardo Gallotti, head of the Complex Behavior Unit at Fondazione Bruno Kessler (FBK).

Based in Trento, Italy, the FBK is a research centre known for its work in digital innovation, AI and the study of complex social systems. In a parallel EU-funded project to AI4Media called AI4Trust, FBK partnered with universities and media organisations across Europe to analyse the wider dynamics of online disinformation.

Partners included Euractiv in Belgium, Sky Italia, and the fact-checking services Maldita.es in Spain, Ellenika Hoaxes in Greece and Demagog in Poland.

While AI4Media focused on detecting manipulated media and integrating verification tools into newsrooms, AI4Trust built a hybrid human-machine system to monitor and analyse disinformation at scale.

Its platform tracks multiple social media and news sites in near real time, using advanced AI algorithms to process multilingual and multimodal content – text, audio and images.

Because the volume of online material far exceeds human capacity, the system filters and flags posts that carry a high risk of being fake. Professional fact-checkers then review this material, and their verified assessments feed back into the system to improve performance.

The two projects are complementary. One focuses on detecting manipulated content, while the other examines how it spreads. Together, they offer both the microscope and the wide-angle lens needed to understand and counter AI-powered disinformation.

An arms race

Using AI to detect AI might sound ironic, but it is serious business. “It is indeed funny, but it’s like an arms race,” Kompatsiaris said.

Generative AI models are evolving at extraordinary speed. When AI4Media began, tools like ChatGPT were still in their infancy. Since then, the quality and realism of AI-generated content have advanced dramatically.

“We have entered a new era where the acceleration is hard for the human mind to keep up with,” Papadopoulos said. “To keep up with AI, you need to be using AI.”

As generative models grow more powerful, detection systems must constantly adapt. That was one of the biggest challenges the researchers faced.

“The technology has progressed so fast that it’s difficult even for us as researchers to keep up,” Papadopoulos explained. “We had to continuously update our models to detect newly generated images.”

The team automated parts of the verification process and regularly retrained their systems. But staying ahead demands continued investment – in both research and the media sector that depends on these technologies.

The future of AI

Yet technology alone is not enough. “We need tools, but we also need policies and rules,” Kompatsiaris said.

Under the EU’s Digital Services Act, very large online platforms must assess and mitigate systemic risks, including the spread of disinformation, and increase transparency about how their systems operate. The Artificial Intelligence Act introduces transparency obligations for certain generative AI systems, including requirements to label AI-generated content.

At the same time, a draft Code of Practice on transparency for AI-generated content aims to encourage clearer disclosure and watermarking standards.

Protecting independent journalism is another priority. The European Media Freedom Act sets out safeguards to ensure that professional media content is recognised and protected on major online platforms.

Large platforms must notify recognised media outlets before removing journalistic content and explain their reasoning, giving organisations time to respond. The goal is to prevent legitimate

reporting from being taken down without justification.

Together, these measures and systems form a wider shield: technology to detect manipulation, regulation to improve transparency and accountability, and safeguards to protect responsible journalism.

Public awareness remains just as vital.

“There is no single solution,” Kompatsiaris said. “We need a combination of AI tools, transparency, regulation and awareness if we want to be more effective against disinformation.”

Research in this article was funded by the EU’s Horizon Programme. The views of the interviewees don’t necessarily reflect those of the European Commission. If you liked this article, please consider sharing it on social media.