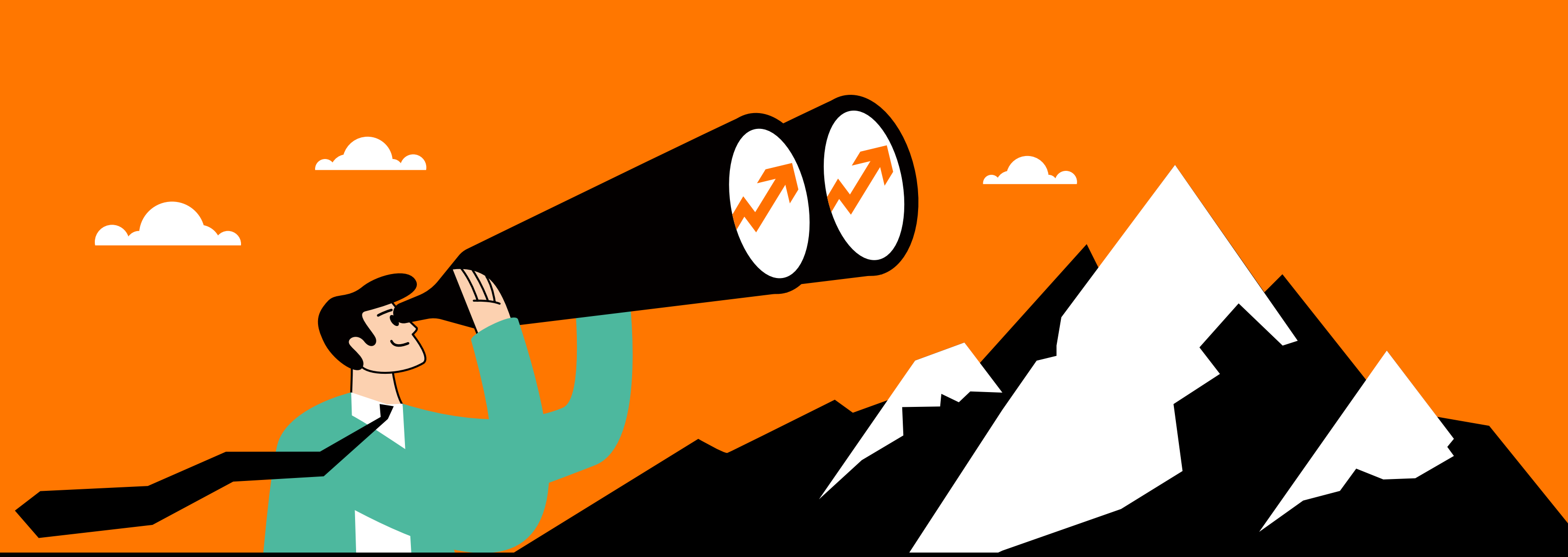
Business

Cyberdefense

Security in an age of deepfakes—and what you can do about it

Ivan's CISO is anxious about the quality of deepfakes. But Ivan knows how to identify AI-generated content.

Artificial Intelligence.
Real Wisdom.



The fast evolution of GenAI disrupts forecasting, monitoring, and the implementation of cybersecurity countermeasures. Vivien Mura, CTO at Orange Cyberdefense, shares insights on the imminent and future threats related to the illicit exploitation of artificial intelligence technology and applications.

Why does GenAI, at the heart of deepfakes and deepaudios, represent a cybersecurity threat?

The overwhelming majority of cyberattacks begins with social engineering, which primarily relies on fake email content or fraudulent websites. Regarding phishing attempts, the best defense so far has been to train individuals and raise awareness.

So far, teaching them how to pinpoint certain anomalies in content—suspicious URLs, spelling mistakes, downgraded and pixelated logos and visuals, requests for credit card or bank account details—was the first line of defense. The second technical barrier came afterward.

However, the rise of GenAI allows the faking of seamless content in a way that even the most vigilant individuals can be fooled. What is coming in the months and years ahead goes way beyond the manipulations we can currently easily identify. Fake content is no longer just textual. It is “multimodal,” meaning it can also process voice, images, and video feeds. GenAI facilitates the production of advanced multimedia content and will likely allow for increased performance at lower costs in the coming months.

Currently, there is still a limitation: producing convincing fake audio/video content without latency or glitches still requires premium models with significant computing power behind them. We have seen tech demos that are still showing irregularities, latency and glitches, which can raise suspicions. But it is only a matter of months before it becomes harder to separate authentic from fake content, produced with more accessible and affordable tools.

The business models associated with AI are constantly changing, not to mention the appearance of a new GenAI tool every two to three months, touted as superior to its predecessors. We are thus observing a cycle of innovation that has become very, very short. It is very likely that highly efficient AI technologies will be used by malicious individuals to conduct sophisticated attacks at lower costs and with minimal skillset levels. As this situation will become the norm, awareness alone will not suffice. It will become complicated, even for an expert, to make the distinction.

This brings a challenge in the field of cybersecurity, as we will need to find other ways to detect such illicit activities. We are on the brink of a bona fide paradigm shift. In this context, finding countermeasure solutions to address the misuses of deepfake and deepvoice will represent a significant challenge. We will need to innovate quickly to detect and respond to these new categories of threats.

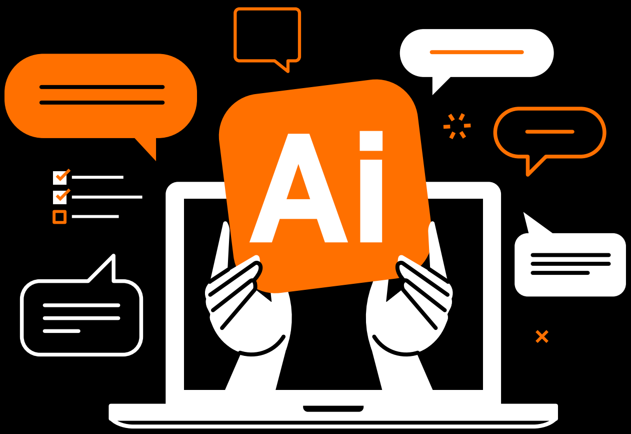
What steps are Orange Cyberdefense currently undertaking to identify suitable countermeasures?

This type of cyberattack won't be countered by a single technological solution, but rather through a series of countermeasures. That is why we actively monitor, identify and benchmark new solutions entering the market, to see which one would make sense and yield efficient results. Orange Cyberdefense's goal is to combine, integrate, and offer these solutions to our clients, to provide a tangible and appropriate response in helping to detect sophisticated fraudulent content.

Beyond the fraudulent uses of GenAI technologies, what are the other risks associated with the rise of these platforms?

The democratization of LLM platforms indeed increases the attack surface and includes vulnerabilities. We believe that cyber attackers will become very interested in hacking AI solutions. Why? Because these AI solutions are interacting with information systems, which represents a potential threat issue in accessing big chunks of data.

Let's put ourselves in the shoes of a cyber attacker who has previously exploited usual vulnerabilities to penetrate an IT or network infrastructure. It is essential to understand that an AI service is interconnected with a whole set of systems. When this population realizes that the shortest path to access a significant volume of data is no longer to exploit a firewall vulnerability but rather that of an AI platform, their approach will likely change direction. In cybersecurity, we will need to find solutions to counter the exploitation of such vulnerabilities.



Do you foresee other threats related to the use of these tools in the future?

In the next five years, we will need to remain very vigilant regarding these models' ability to automate, plan, and execute offensive scenarios.

The level of automation will likely increase in our daily digital lives. This technological evolution could also be used for illicit purposes. The current level of AI has not yet reached this degree of automation, but we must be wary of what will soon be possible:

- Conduct a recon phase
- Identify critical assets
- Identify exploitable vulnerabilities
- Establish a compromise chain
- Even build a whole attack infrastructure
- Ultimately conduct an end-to-end cyberattack

I am indeed referring to the setup of a genuine battle plan that the tool will fully execute or assist with. We are not there yet. When we look at the planning capabilities of AI at the moment, they remain very limited.

What can you tell us regarding disinformation tactics?

We are observing a trend in cyberattacks hybridization, blending traditional cyberattacks with informational attacks. It is obvious that those engaging in such activities are aligned with states to destabilize the digital space of organizations, other states, companies, etc.

The goal is to destabilize as much as possible, manipulate opinions, make some noise, and amplify the crisis.

This trend has several roots:

1. The current geopolitical context of a global "multipolar" crisis: the war in Ukraine, the Israeli-Palestinian conflict, the industrial war between the United States and China… These tensions encourage certain powers or criminal organizations to venture into this territory.
2. The democratization of GenAI, which facilitates the creation of fake content with tools that are almost free.

Disinformation, manipulation of information, fake news… These tactics are used to amplify the crisis and weaken the target. Regarding cyber extortion, or Cy-X, simply making someone believe they have been the victim of a cyberattack, such as a data breach or ransomware, can be as critical as if the attack was real. A business can suddenly find itself needing to respond on social media, to the press, to its partners, and to its clients, all while conducting an internal investigation to ensure there has been no alert, leak, or actual attack.

In this case, the first thing the company must do is:

1. Verify if the threat is genuine. This requires time for investigation and qualification, which can take a week or more.
2. Mobilize its resources on two battlefields: the technical field of cybersecurity and the informational field.

From a doctrinal standpoint, the United Kingdom and English-speaking countries have long considered that the security of the digital space integrates both information systems and information itself. In France, we have separated the two: cybersecurity on one side, and crisis information management on the other. One might wonder if this doctrine will need to evolve, with all necessary precautions to limit abuses.