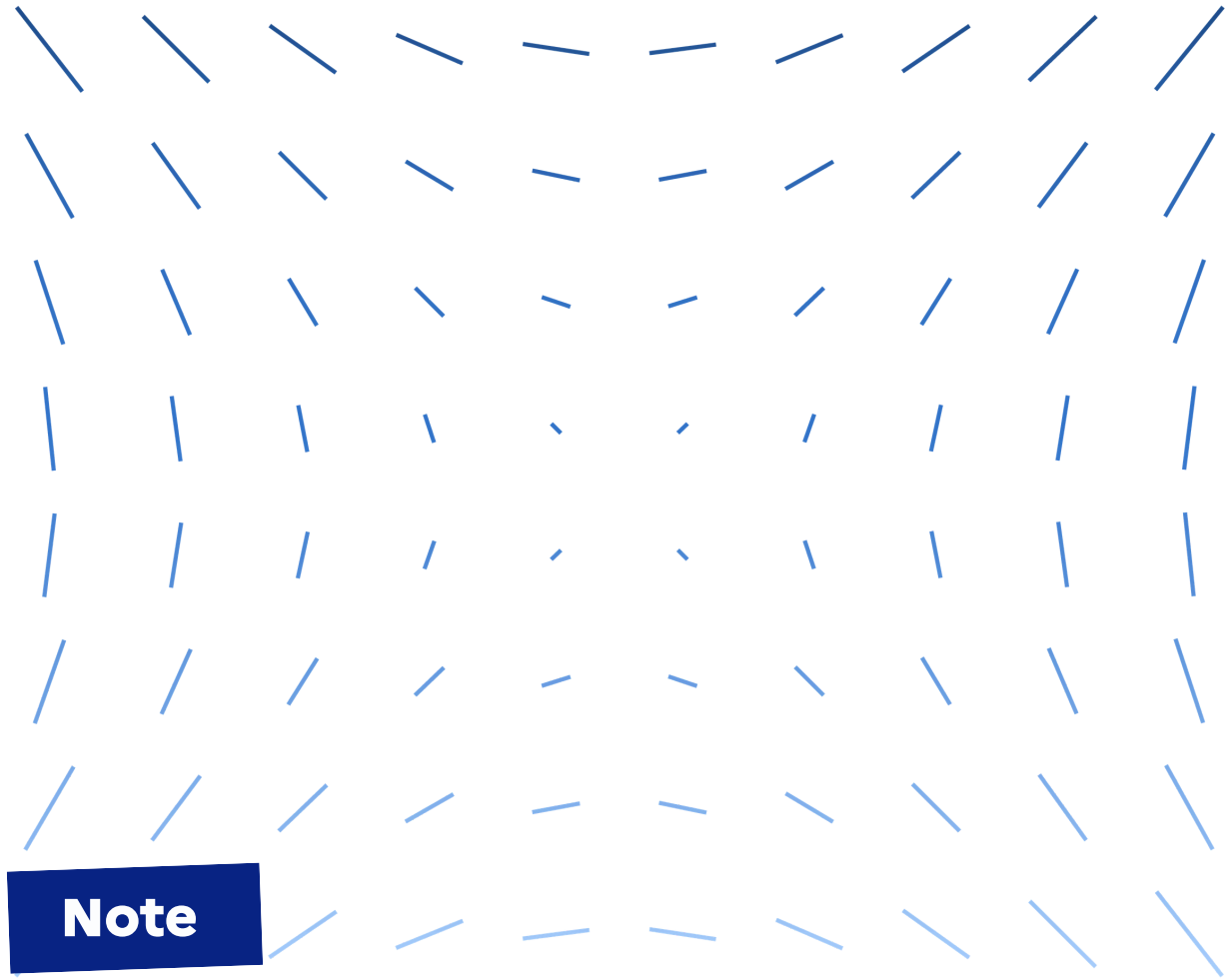
RÉPUBLIQUE FRANÇAISE

Liberté
Égalité
Fraternité

Conseil de l'IA et du Numérique



Note

AI & Cybersecurity

Anticipating Today to Master Tomorrow

June 2026

Observing the varied reactions – ranging from panic to denial – following the announcements surrounding Mythos, an artificial intelligence model developed by Anthropic reported to stand out for its strong capacity to detect and exploit security flaws in software, browsers, and computer operating systems, the French AI and Digital Council sets out in this short note its initial reflections on the links between artificial intelligence (AI) and cybersecurity.

1. The balance between attack and defence has always driven the cybersecurity sector. Advances in one of the two domains systematically prompt a reaction from the other side, and this has been the case for more than 30 years. Each advance in attack capabilities (such as the explosion of computer viruses in the early 2000s) pushes defenders to innovate in order to protect themselves. Conversely, each improvement in defence systems (such as the mass adoption of communications encryption from the 1990s onwards) prompts the offensive side to find new ways of circumventing these protections. **This balance is dynamic and never stable**: it evolves according to the technologies, skills, and motivations of the forces at play.

2. In cybersecurity as in other fields, AI is generating rapid transformations whose consequences will almost certainly be major and lasting. As specified by the French Cybersecurity Agency (ANSSI), the impact of AI on the security of information systems is threefold:

    a. The cybersecurity of AI: AI systems are first and foremost information systems and, consequently, the target of attacks. These systems may be attacked in a 'traditional' manner but also have their own vulnerabilities, which may give rise to attacks of a novel nature (prompt injection, poisoning through training data, etc.).

    b. Cybersecurity through AI: AI offers new tools as well as the automation of certain tasks for the benefit of defenders. For example, the detection of attacks in the cyber field now relies on phenomenal quantities of technical data in seeking to identify anomalies, a task far better suited to the machine than to the human.

    c. Cybersecurity in the face of AI refers to the symmetrical use of AI by attackers, who can automate their attacks and gain in both effectiveness and speed.

3. The progress of AI in searching for vulnerabilities in software components, proprietary and open source alike, is unprecedented. This is the very subject of the debates surrounding Mythos, whose performance in terms of offensive capabilities impresses and provokes

lively debate, particularly in the banking sector[1][2], even as the model is not currently available, Anthropic deeming it too effective to risk seeing it fall into the wrong hands. The American company has confined itself to sharing the model with 40 actors, almost all of them American, such as J.P. Morgan, Nvidia, Amazon, Apple, or Chase Bank. The riposte was immediate on OpenAI's side, which launched GPT 5.4 Cyber on 14 April, ostensibly less capable than Mythos. Other models will follow in the coming months, but the lesson lies elsewhere: the interdependence between AI and cybersecurity continues to strengthen. Whilst it is still too early to say whether this technology will benefit attackers or defenders more, certain common-sense observations are worth making:

a. **Those actors, public and private alike, who fail to make the ongoing effort to understand and integrate these new uses will quickly be left behind.** To guard against this, the challenge is not to wait for the release of the most advanced AI models; rather, it is urgent to bring oneself up to standard in terms of compliance and to apply the fundamental principles of cybersecurity, such as regular security updates or the implementation of rigorous access controls (referring, for example, to ANSSI's recommendations here, in French).

b. Whilst the human is today less and less in the loop, and whilst AI constitutes a valuable aid for developing, correcting, and testing software, **dispensing entirely and immediately with human expertise would appear to be a guarantee of failure**, as AI also generates vulnerabilities that other AI systems are not certain to detect in time. In the same way, agentic AI applied to software development, security, and operations ('DevSecOps') has already demonstrated its value but also significant risks if it is not sufficiently supervised.

c. One certainty nonetheless: the pace at which software vulnerabilities are discovered and corrected will keep on growing, at the risk of becoming entirely unmanageable with current methods. At first glance, this is rather good news for the security of information systems, since they will be increasingly free of flaws. This optimistic view could, however, prove naive and misleading if the services responsible for information systems fail to keep up with the frenetic pace of patching or, worse, if such mechanisms are not yet systematised, as remains all too often the case in the field of industrial computing. Many actors are already stretched to capacity and unable to keep their information systems up to date. Yet a published patch effectively reveals the vulnerability of which it is the consequence, which will push attackers, boosted by AI, to exploit these patches and target all those who have not yet had the time to bring themselves up to standard (and they will probably be numerous).

[1] Financial Times, Latest AI models could threaten world banking system, financial officials warn, April 17, 2026.
[2] Les Échos, IA : l'Europe s'inquiète à son tour de la menace que Mythos fait peser sur les banques, April 16, 2026.

*

4. AI, like any rapidly developing technological field, is the object of a great deal of hype and announcement-driven marketing, with each actor seeking to convince others that it is ideally placed to win the frantic race to AI and to raise colossal sums (for the record, OpenAI's latest fundraising round in March 2026 amounted to 122 billion dollars). **Touting the 'dangerousness' of one's models should they fall into the wrong hands is a clever way of showcasing their performance and generating keen interest, including among investors.** The method is not new: in February 2019[3], OpenAI claimed that its GPT 2.0 model was too dangerous to be released, before finally making it public six months later. In the same way, calling upon religious dignitaries[4] to 'regulate' AI is an equally clever way of suggesting that, in addition to being 'intelligent', these models are on the verge of attaining a form of 'spirituality'.

5. Whilst the use of AI is becoming indispensable, it risks creating new and particularly significant technological dependencies, even a 'sovereignty dilemma' for those actors who, anxious to remain at the cutting edge of uses, would undermine their mastery of risks and the immunity of the architecture of their information systems by considerably extending the cyber risk surface. The consequences may be financial for enterprises, whether directly on account of data breaches, industrial sabotage, high remediation costs to restore compromised systems, or through the cost of technological dependence (lock-in by proprietary solutions). The risks of massive leaks of data and intellectual property may grow rapidly, either directly or as a result of 'shadow AI', which remains highly widespread. However, **the question arises in even more serious terms for those actors whose activities relate to national sovereignty and who risk being confronted with delicate trade-offs between performance and autonomy.**

6. Furthermore, these developments highlight the growing and imperative need to strengthen the capabilities for evaluating AI models in France and in Europe. In Silicon Valley, evaluation is now more than a subject of interest; certain major research laboratories, such as the Stanford Data Science Institute, headed by the world-leading statistician Emmanuel Candès, have pivoted towards the evaluation of AI models. In London, the British AI Security Institute, created in 2023, conducted an independent audit[5] of Mythos's capabilities, tempering on that occasion the collective excitement: "our testing shows that Mythos can exploit systems with weak security posture […] This highlights the importance of cybersecurity basics". **This nonetheless raises acutely the question of French and European capabilities in the evaluation of AI models.** France thus suffers from a strong information asymmetry compared to the British, American, and Chinese ecosystems. In the medium term, the consequences will go beyond the issues of model safety and security;

[3] "Due to our concerns about malicious applications of the technology, we are not releasing the trained model", February, 2019.
[4] Clubic, Anthropic invite 15 chrétiens pour un sommet sur la moralité de l'intelligence artificielle, April 13, 2026.
[5] "Our evaluation of Claude Mythos Preview's cyber capabilities", April 13, 2026.

they will bear on the quality of the ecosystem and of the talent, which will necessarily gravitate around laboratories and enterprises 'at the frontier'.

7. In conclusion, AI is agnostic with regard to cybersecurity, in the sense that it assists both attackers and defenders. In the current context, **the French AI and Digital Council forcefully reiterates that cybersecurity is an eminently important subject, vital for public and private entities alike, and that a genuine seriousness in this field proves more indispensable than ever.** In this regard, with the aim of encouraging reflection and action from now on in a rapidly expanding field, we can only call for:

    - actors not to give in to the surrounding panic, but rather to consider AI governance and security frameworks from the moment solutions are adopted. In the short term, enterprises could envisage the creation of a permanent function dedicated to the autonomous discovery, qualification, and remediation of vulnerabilities, as an extension of the 'DevOps' method;
    - the implementation of general frameworks which, although not specifically designed for AI, remain entirely apposite, such as the one associated with the European NIS2 Directive;
    - the accelerated structuring of a European public-private ecosystem for the evaluation of AI models around the French National Institute for the Evaluation and Security of AI (INESIA) and leading-edge AI laboratories in Europe.