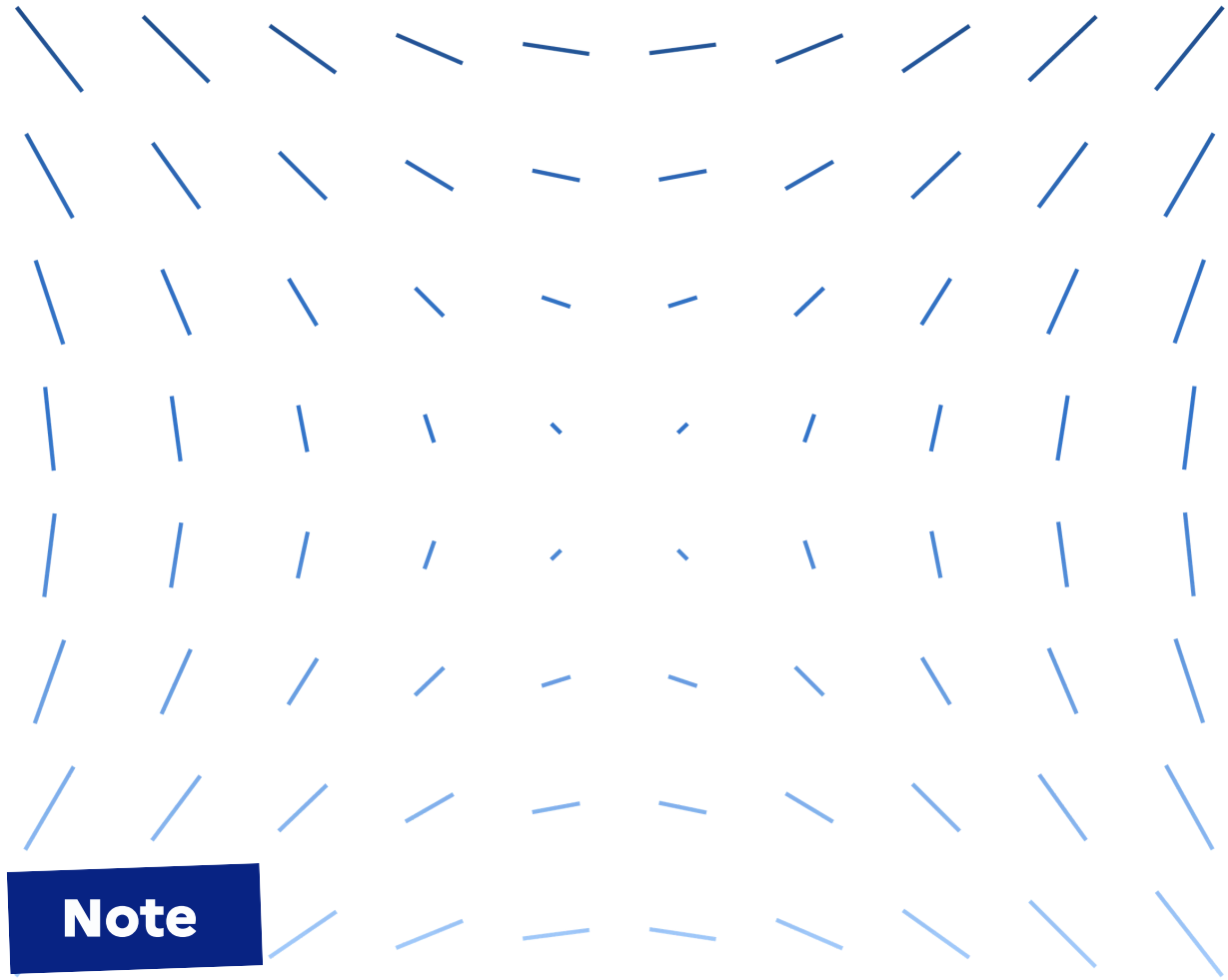
RÉPUBLIQUE FRANÇAISE

Liberté
Égalité
Fraternité

Conseil de l'IA et du Numérique



Note

IA & cybersécurité anticiper aujourd'hui pour maîtriser demain

avril 2026

Observant les réactions variées – allant de la panique au déni – à la suite des annonces autour de Mythos, un modèle d'intelligence artificielle développé par Anthropic qui se distinguerait par sa forte capacité à détecter et exploiter des failles de sécurité dans les logiciels, navigateurs et systèmes d'exploitation informatiques, le Conseil de l'IA et du numérique livre au travers de cette courte note ses premières réflexions sur les liens entre intelligence artificielle (IA) et cybersécurité.

1. L'équilibre entre attaque et défense anime depuis toujours le secteur de la cybersécurité. Les progrès dans l'un des deux domaines entraînent systématiquement la réaction de l'autre camp et ce depuis plus de 30 ans. Chaque avancée en matière d'attaque (comme l'explosion des virus informatiques au début des années 2000) pousse les défenseurs à innover pour se protéger. À l'inverse, chaque amélioration des systèmes de défense (à l'image de l'adoption massive du chiffrement des communications dès les années 1990) incite la partie offensive à trouver de nouvelles façons de contourner ces protections. **Cet équilibre est dynamique et n'est jamais stable** : il évolue en fonction des technologies, des compétences et des motivations des forces en présence.

2. En matière de cybersécurité comme dans d'autres domaines, l'IA génère des transformations rapides dont les conséquences seront très certainement majeures et durables. Comme précisé par l'Agence nationale de la sécurité des systèmes d'information (ANSSI), l'impact de l'IA sur la sécurité des systèmes d'information est triple :

    a. La cybersécurité de l'IA : les systèmes d'IA sont avant tout des systèmes d'information et, par conséquent, la cible d'attaques. Ces systèmes peuvent être attaqués de manière « classique » mais disposent également de leurs propres vulnérabilités, qui peuvent engendrer des attaques de nature originale (*prompt injection*, empoisonnement par les données d'apprentissage, *etc.*).

    b. La cybersécurité par l'IA : l'IA propose de nouveaux outils ainsi qu'une automatisation de certaines tâches au profit des défenseurs. Par exemple, la détection d'attaques dans le champ cyber s'appuie désormais sur des quantités phénoménales de données techniques pour chercher à identifier des anomalies, tâche bien plus adaptée à la machine qu'à l'humain.

    c. La cybersécurité face à l'IA fait référence à l'usage symétrique de l'IA par les attaquants, qui peuvent automatiser leurs attaques et gagner à la fois en efficacité et en rapidité.

3. Les progrès de l'IA pour rechercher des vulnérabilités dans les briques logicielles, propriétaires comme *open source*, sont sans précédent. C'est tout l'objet des débats autour de Mythos, dont les performances en matière de capacités offensives

impressionnent et suscitent de vifs débats, en particulier dans le secteur bancaire[1] [2], alors que le modèle n'est aujourd'hui pas disponible, Anthropic le jugeant trop efficace pour risquer de le voir tomber entre les mauvaises mains. L'entreprise américaine s'est contentée de partager le modèle auprès de 40 acteurs, quasi-exclusivement américains, comme par exemple J.P. Morgan, Nvidia, Amazon, Apple ou Chase Bank. La réplique fut immédiate côté OpenAI, qui a lancé le 14 avril GPT 5.4 Cyber, *a priori* moins performant que Mythos. D'autres modèles suivront dans les prochains mois mais l'enseignement est ailleurs ; l'interdépendance entre l'IA et la cybersécurité ne cesse de se renforcer. S'il est encore tôt pour dire si cette technologie bénéficiera plus aux attaquants ou aux défenseurs, certaines remarques de bon sens méritent d'être formulées :

a. **Les acteurs, publics comme privés, qui ne feront pas l'effort permanent de comprendre et d'intégrer ces nouveaux usages seront rapidement déclassés**. Pour s'en prémunir, l'enjeu n'est pas d'attendre la sortie des modèles d'IA les plus avancés, mais il est urgent de se mettre à niveau en matière de conformité et d'appliquer les principes fondamentaux de la cybersécurité tels que les mises à jour régulières de sécurité ou la mise en place de contrôles d'accès rigoureux (en se référant par exemple aux recommandations de l'ANSSI ici).

b. Si l'humain est aujourd'hui de moins en moins dans la boucle et si l'IA constitue une aide précieuse pour développer, corriger et tester les logiciels, **se passer dès maintenant et totalement de l'expertise humaine semble la garantie d'échecs**, l'IA générant également des vulnérabilités que d'autres systèmes d'IA ne sont pas certains de détecter à temps. De la même manière, l'IA agentique appliquée au développement, à la sécurité et à l'exploitation logicielle (« DevSecOps ») a déjà démontré son intérêt mais également d'importants risques si elle n'est pas suffisamment encadrée.

c. Une certitude néanmoins : le rythme auquel les vulnérabilités logicielles sont découvertes et corrigées ne cessera de croître, au risque de devenir totalement ingérable avec les méthodes actuelles. En première intention, il s'agît plutôt d'une bonne nouvelle pour la sécurité des systèmes d'information puisqu'ils seront de plus en plus exempts de failles. Cette vision optimiste pourrait toutefois s'avérer naïve et trompeuse si les services chargés des systèmes d'information ne parviennent pas à suivre le rythme effréné du « *patching* » ou, pire, si de tels mécanismes ne sont pas encore systématisés comme c'est encore trop souvent le cas dans le domaine de l'informatique industrielle. Beaucoup d'acteurs sont aujourd'hui déjà saturés et incapables de conserver leurs systèmes d'information à jour. Or, un correctif publié révèle de fait la vulnérabilité dont il est la conséquence, ce qui poussera les attaquants, dopés à l'IA, à exploiter ces patchs et à cibler tous ceux qui n'auront pas encore eu le temps de se mettre à niveau (et ils seront probablement nombreux).

*

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

4. L'IA, comme tout domaine technologique au développement rapide, est l'objet de beaucoup de communication, d'effets d'annonce et de stratégies *marketing*, chaque acteur cherchant à convaincre qu'il est en place idéale pour remporter la course folle à l'IA et à lever des sommes colossales (pour mémoire, la dernière levée de fonds d'OpenAI en mars 2026 s'élevait à 122 milliards de dollars). **Vanter la « dangerosité » de ses modèles s'ils tombent entre de mauvaises mains est une manière habile de mettre en avant leurs performances et de susciter un vif intérêt, y compris du côté des investisseurs**. La méthode n'est pas nouvelle ; en février 2019[3], OpenAI affirmait que son modèle GPT 2.0 était trop dangereux pour être sorti, avant de finalement le rendre public six mois plus tard. De la même manière, faire appel à des dignitaires religieux[4] pour « régler » l'IA est une façon tout aussi habile de suggérer qu'en plus d'être « intelligents », ces modèles seraient sur le point d'accéder à une forme de « spiritualité ».

5. Si l'usage de l'IA devient indispensable, il risque de créer de nouvelles dépendances technologiques particulièrement importantes, voire même un « **dilemme de souveraineté** » pour les acteurs qui, soucieux de rester à la pointe des usages, fragiliseraient leur maîtrise des risques et l'immunité de l'architecture de leurs systèmes d'information en étendant considérablement la surface du risque cyber. Les conséquences peuvent être financières pour les entreprises, que ce soit directement à cause de violations de données, de sabotage industriel, de coûts de remédiation élevés pour restaurer des systèmes compromis ou par le coût de la dépendance technologique (*lock-in* par des solutions propriétaires). Les risques de fuites massives de données et de propriété intellectuelle peuvent rapidement croître, directement ou bien du fait du « *shadow AI* », encore très répandu. Toutefois, **la question se pose en termes plus sérieux encore pour les acteurs dont les activités sont en lien avec la souveraineté nationale et qui risquent d'être confrontés à des arbitrages délicats entre performance et autonomie.**

6. En outre, ces développements mettent en lumière les besoins croissants et impératif de renforcer les capacités d'évaluation des modèles d'IA en France et en Europe. Dans la Silicon Valley, l'évaluation est désormais plus qu'un sujet d'intérêt ; certains grands laboratoires de recherche, à l'image du *Stanford Data Science Institute*, dirigé par la pointure mondiale des statistiques Emmanuel Candès, qui a pivoté vers l'évaluation des modèles d'IA. Du côté de Londres, le *AI Security Institute* britannique, créé en 2023, a mené un audit indépendant[5] des capacités de Mythos, en tempérant à cette occasion l'emballement collectif : « *our testing shows that Mythos can exploit systems with weak security posture* […] *This highlights the importance of cybersecurity basics* ». Pour autant, **cela pose de façon aigue la question des capacités françaises et européennes en matière d'évaluation des modèles d'IA**. La France souffre ainsi d'une forte asymétrie d'informations par rapport aux écosystèmes britannique, américain et chinois. À moyen terme, les conséquences dépasseront les enjeux de sûreté et de sécurité des modèles, elles se porteront sur la qualité de l'écosystème et des talents, qui graviteront nécessairement autour des laboratoires et entreprises « à la frontière ».

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

*

7. En conclusion, l'IA est agnostique en matière de cybersécurité, au sens où elle aide à la fois les attaquants et les défenseurs. Dans le contexte actuel, **le Conseil de l'IA et du numérique rappelle avec force que la cybersécurité est un sujet éminemment important, vital pour les entités publiques comme privées et qu'un réel sérieux dans ce domaine s'avère plus que jamais indispensable**. À ce titre, dans le but d'encourager dès maintenant la réflexion et l'action dans un champ en pleine expansion, nous ne pouvons qu'appeler :

    - les acteurs à ne pas céder à la panique ambiante, mais à penser les cadres de gouvernance et de sécurité de l'IA dès l'adoption des solutions. À court terme, les entreprises pourraient envisager la création d'une fonction permanente dédiée à la découverte, la qualification et la remédiation autonomes de vulnérabilités, dans le prolongement de la méthode « DevOps » ;

    - à la mise en œuvre de référentiels généraux qui, bien que non spécifiquement pensés pour l'IA, restent totalement à propos, à l'image de celui associé à la directive européenne NIS2 ;

    - à la structuration accélérée d'un écosystème public-privé européen d'évaluation des modèles d'IA autour de l'Institut national pour l'évaluation et la sécurité de l'IA (INESIA) et de laboratoires IA de pointe en Europe.