NOTE D'INFORMATION ET D'ACTUALITÉ

Juillet 2026

Sécurité numérique et intelligence artificielle : mutations de la menace et perspectives

Cigref
RÉUSSIR
LE NUMÉRIQUE

Note d’information et d’actualité du Cigref

Sécurité numérique et intelligence artificielle : mutations de la menace et perspectives

Juillet 2026

TABLE DES MATIÈRES

INTRODUCTION

1.1 L'INTELLIGENCE ARTIFICIELLE ACCÉLÈRE LA TRANSFORMATION DE LA MENACE CYBER

Depuis plusieurs années, le Cigref analyse en continu l'évolution de la menace cyber à partir des retours d'expérience de ses 160 organisations membres et des travaux conduits avec les principales autorités et communautés d'expertise françaises, européennes et internationales. Cette observation s'appuie sur une réalité opérationnelle. Les grandes entreprises et les administrations publiques sont aujourd'hui confrontées à une intensification durable des attaques, à une augmentation de leur complexité et à une réduction continue des délais disponibles pour les détecter, les contenir et y remédier.

Les premiers mois de l'année 2026 marquent toutefois une évolution d'une autre nature. Plusieurs événements rapprochés ont mis en évidence une accélération des transformations déjà observées depuis plusieurs années. Les rapports publiés par les principaux acteurs de la cybersécurité confirment l'effondrement du délai entre la divulgation d'une vulnérabilité et son exploitation, l'industrialisation des attaques assistées par l'intelligence artificielle et l'augmentation du volume d'événements que les équipes de sécurité doivent traiter quotidiennement. Dans le même temps, les superviseurs financiers européens ont renforcé leur niveau de vigilance. La Banque centrale européenne a demandé aux principaux établissements bancaires de la zone euro d'actualiser leurs plans d'action face aux risques liés à l'intelligence artificielle, tandis que le Comité européen du risque systémique a relevé son appréciation du risque cyber.

L'actualité de l'été 2026 a également révélé une évolution qui dépasse le seul domaine de la cybersécurité. La décision des autorités américaines de soumettre les modèles frontière d'intelligence artificielle d'Anthropic à des restrictions d'exportation fondées sur des considérations de sécurité nationale, suivie de la suspension temporaire de leur accès pour les utilisateurs ne relevant pas de la catégorie des *US persons*, a montré que certaines capacités d'intelligence artificielle pouvaient désormais faire l'objet de mesures de contrôle comparables à celles appliquées à d'autres technologies stratégiques. Bien que ces restrictions aient été levées le 30 juin 2026 et l'accès aux services d'Anthropic rétabli dès le 1er juillet 2026, cette séquence rappelle aux organisations européennes que la résilience numérique dépend aussi de la continuité d'accès aux technologies sur lesquelles repose leur capacité de protection.

Ces évolutions trouvent un premier prolongement dans le **Plan d'action sur la cybersécurité et l'intelligence artificielle** présenté par la Commission européenne en juillet 2026. Ce plan reconnaît explicitement que les modèles d'intelligence artificielle de pointe modifient profondément les conditions d'exercice de la cybersécurité. Il propose plusieurs mesures destinées à renforcer les capacités européennes d'évaluation, de test, de déploiement et de développement de ces technologies au bénéfice des infrastructures critiques, des administrations publiques et des entreprises.

Les événements intervenus au cours des derniers mois ne remettent pas en cause les analyses conduites par le Cigref depuis plusieurs années. Ils les confirment et leur donnent une portée nouvelle. Les transformations observées ne relèvent plus uniquement de l'évolution des techniques d'attaque. Elles affectent désormais les conditions dans lesquelles les grandes organisations pourront maintenir, dans la durée, leur capacité à protéger leurs systèmes d'information.

La présente note propose une lecture de ces évolutions autour de huit mutations qui, prises ensemble, dessinent un changement durable du paysage de la menace cyber. Chacune est présentée à partir des observations recueillies auprès des membres du Cigref, des analyses publiées par les principales autorités compétentes et les organismes de référence, des événements les plus récents, afin d'éclairer les conséquences de ces mutations pour la résilience des grandes organisations.

1.2 UNE DÉPENDANCE STRUCTURELLE AUX CAPACITÉS NUMÉRIQUES CRITIQUES EXTRA-EUROPÉENNES

Avant d'exposer les mutations que nous observons, il convient de rappeler la toile de fond économique et technologique sur laquelle elles se déploient. La France et l'Europe subissent une dépendance sévère de leur économie à l'égard d'une domination structurelle des acteurs américano-israéliens de l'industrie de la cybersécurité qui bride de manière croissante notre autonomie stratégique. Ce constat est d'ailleurs cohérent avec les conclusions globales de l'étude réalisée par le cabinet Asterès pour le compte du Cigref, et publiée en avril 2025. Cette étude estime qu'au sein de l'Union européenne, environ 83 % des achats de logiciels et de services cloud sur le marché BtoB sont effectués auprès de l'industrie américaine du numérique. Sur le plan financier, les écosystèmes nord-américains et israéliens captent plus des trois-quarts des investissements mondiaux du capital-risque du secteur, reléguant l'Europe, qui en capte entre 5 et 15 % selon les années, au statut de marché de consommation technologique.

À titre d'illustration, les entreprises technologiques occupant les principales positions du marché de la cybersécurité en Europe sont pour la plupart extra-européennes. Microsoft est devenu l'un des premiers acteurs mondiaux de la cybersécurité et domine notamment les marchés de l'identité, de l'EDR, du SIEM et de la sécurité du cloud. À ses côtés figurent également Palo Alto Networks, Fortinet, Check Point Software Technologies, CrowdStrike, Zscaler, ou encore SentinelOne.

Cette vulnérabilité est aujourd'hui accélérée par la financiarisation du secteur, à l'image de l'activité du fonds américain de capital-investissement Thoma Bravo. Ce dernier procède à une consolidation agressive du marché mondial par croissance externe, en rachetant successivement des entreprises structurantes du secteur de la protection des accès, de la messagerie ou de la détection. L'industrie de la cybersécurité est désormais marquée par une concentration croissante qui tend à imposer ses conditions économiques et technologiques. Cette concentration du capital favorise une intégration verticale des offres et renforce les effets de dépendance technologique des utilisateurs européens.

En juin 2024 déjà, le CESIN s'inquiétait publiquement de la dynamique d'extrême concentration des fournisseurs de cybersécurité anglo-saxons au sein du fonds Thoma Bravo. À cette occasion, le CESIN notait que de nombreuses entreprises, qui déploient des stratégies de diversification pour éviter de

dépendre d'un seul fournisseur de cybersécurité, se retrouvaient finalement piégées par ces acquisitions successives.

Cette dépendance ne concerne plus seulement les logiciels de cybersécurité ou les services cloud. Les événements de juin 2026 montrent qu'elle s'étend désormais aux modèles d'intelligence artificielle les plus avancés. Ceux-ci deviennent progressivement des composants essentiels des dispositifs de protection, de détection et de remédiation des grandes organisations. Leur disponibilité constitue désormais un facteur de résilience des systèmes d'information.

Cette concentration industrielle et financière conduit les grandes organisations européennes à dépendre d'un nombre limité d'acteurs pour des capacités devenues essentielles à leur propre protection. Cette évolution constitue un enjeu majeur de résilience, auquel les développements récents liés à l'intelligence artificielle donnent une portée nouvelle.

Face à cet écosystème technologique et financier particulièrement structuré, où les innovations de pointe sont continuellement financées et absorbées par l'industrie américaine, l'Europe souffre d'une marginalisation progressive de ses propres solutions. C'est dans ce contexte de verrouillage industriel et de vulnérabilité croissante de nos chaînes d'approvisionnement en produits et services de sécurité numérique que les huit mutations suivantes prennent toute leur dimension critique.

1.3 UNE POSTURE DE SÉCURITÉ DEVENUE INADAPTÉE À LA VITESSE DE LA MENACE

Les cadres doctrinaux, les instruments de défense et les schémas organisationnels sur lesquels repose actuellement notre sécurité nationale ont été conceptualisés pour un environnement qui n'existe plus, celui d'une conflictualité de basse intensité, où le temps de la remédiation et la compartimentation de l'information de crise étaient encore des options viables. Aujourd'hui, l'observation méthodique des modes opératoires adverses démontre à l'évidence, dans ce contexte d'asymétrie stratégique croissante, qu'il n'est plus possible de répondre aux menaces de demain avec les paradigmes d'hier. De simples ajustements capacitaires ou budgétaires de nos dispositifs actuels se heurteront inévitablement au mur de la réalité mathématique et opérationnelle des mutations de la menace.

Les huit mutations documentées dans cette note d'information et d'actualité soulignent l'obsolescence structurelle de nos postures défensives traditionnelles. Elles exigent une refonte profonde de notre doctrine nationale de sécurité numérique afin de restaurer la résilience, tant individuelle des entreprises et des administrations publiques que collective de la société et de son économie, face à la vélocité et à l'ampleur de la menace telle qu'elle se matérialise d'ores et déjà.

Afin de faciliter la lecture de cette note et d'ancrer notre analyse dans la réalité à laquelle nos adhérents sont confrontés, nous avons choisi de structurer la présentation de chacune de ces huit mutations selon une même grille de lecture :

- La nature de la mutation : une description factuelle de la rupture technique ou géopolitique, telle que nous pouvons l'observer dans sa dynamique temporelle.
- L'illustration : la mise en perspective du phénomène à travers des données chiffrées ou des cas d'usage récemment documentés.

- L'impact sur notre posture collective : les raisons concrètes pour lesquelles nos approches défensives classiques se révèlent aujourd'hui inadaptées face à cette nouvelle donne.

Cette démarche vise à partager un diagnostic objectivé, condition préalable indispensable à la co-construction d'une réponse nationale à la hauteur de ces enjeux. Les événements intervenus depuis le printemps 2026 confirment cette analyse. Les décisions prises par les superviseurs financiers européens, les initiatives engagées par la Commission européenne ainsi que les évolutions observées sur les modèles d'intelligence artificielle traduisent une même réalité : les organisations doivent désormais adapter leur posture de sécurité à une menace dont les capacités évoluent à un rythme inédit.

2 ANALYSE DESCRIPTIVE DES HUIT MUTATIONS

2.1 EFFONDREMENT DU TIME TO EXPLOIT ET OBSOLESCENCE DE LA GESTION DES CORRECTIFS

Nature de la mutation

Le temps nécessaire entre la découverte d'une vulnérabilité informatique et son exploitation active par des acteurs malveillants, communément appelé *Time to Exploit* (TTE), s'est littéralement effondré au cours de ces dernières années pour devenir négatif depuis 2024. Le paradigme défensif traditionnel reposait sur le délai permettant aux défenseurs d'identifier la faille, de tester le correctif et de le déployer. Or, la mise en œuvre de stratégies de gestion des correctifs et de *patch management* exige, la plupart du temps, de mener des tests de non-régression, qui peuvent s'avérer longs et complexes pour garantir le maintien en conditions opérationnelles des applications métiers critiques. Le cycle de déploiement réel des correctifs au sein d'une grande organisation nécessitant en moyenne de 60 à 150 jours, la défense périmétrique basée sur le traitement réactif des vulnérabilités est devenue mathématiquement obsolète face à des attaquants agissant désormais à la vitesse machine.

Illustration

Selon le rapport *M-Trends 2026* publié par Mandiant, filiale de cybersécurité de Google Cloud, le temps moyen d'exploitation d'une vulnérabilité était de 63 jours en 2018. Il est devenu négatif en 2024, l'exploitation d'une vulnérabilité se produisant, en moyenne, un jour avant qu'un correctif ne soit publié. En 2025, les experts estiment que le TTE est négatif de 7 jours. Selon ce même rapport *M-Trends 2026*, le délai de transmission, au sein de la chaîne du cybercrime organisé, d'un accès initial par un *broker* de vulnérabilité à un affilié de rançongiciel est passé de plus de 8 heures en 2022 à moins de 30 secondes aujourd'hui. Le rapport *IBM X-Force Threat Intelligence Index 2026* confirme cette massification en relevant une hausse de 44 % des attaques exploitant des applications exposées publiquement en 2025. Par ailleurs, 56 % des vulnérabilités divulguées ne nécessitent aucune authentification pour être exploitées.

Ce constat est corroboré par l'autorité fédérale allemande de cybersécurité (BSI), qui alerte ces derniers jours sur une réduction brutale du temps séparant la découverte d'une faille de sa tentative d'exploitation, rendant les délais de correction de plusieurs semaines critiques et dangereux. L'ANSSI elle-même craint dans ce contexte ce que son Directeur général qualifie de perte de « la bataille contre les vulnérabilités logicielles ».

Dès 2021, le Cigref et ses partenaires européens alertaient sur le fardeau écrasant de cette asymétrie. Nous dénoncions déjà la mobilisation permanente de dizaines d'équivalents temps plein et les millions d'euros dépensés annuellement par les grandes entreprises pour pallier les défauts de sécurité des éditeurs de logiciels. Aujourd'hui, avec un TTE négatif, ce modèle, par lequel les éditeurs de logiciels et de services numériques externalisent le risque technique et financier sur leurs clients, a dépassé son point de rupture.

Impact sur notre posture collective

Cette vélocité de la menace démontre que la sécurité ne peut plus reposer sur la seule agilité des victimes à appliquer des patchs de sécurité dans une urgence croissante. L'industrie numérique doit urgemment être soumise à des normes de *Security by Design* comparables à celles d'autres secteurs critiques. Il s'agit d'un combat historique du Cigref, qui alertait formellement le Premier Ministre dès novembre 2020 en s'appuyant sur les travaux du Conseil général de l'économie (CGE). Ce dernier préconisait dès 2018 que « *la sécurité par défaut doit devenir la règle dans la conception des produits* » et que son maintien ne doit nécessiter aucune « *action explicite de l'utilisateur* ».

Pourtant, des années après ce constat partagé par les instances de l'État et soutenu par les recommandations de l'OCDE publiées en février 2021 dans le rapport « *Enhancing the digital security of products* », le cadre réglementaire peine à s'adapter. Si le *Cyber Resilience Act* (CRA) européen a constitué une première étape, il demeure notoirement inadapté pour répondre à l'effondrement actuel du TTE. Le texte reste insuffisamment contraignant quant à l'exigence d'une véritable résilience native et ne dispose pas des mécanismes adéquats pour forcer une remédiation en temps réel à l'échelle industrielle.

2.2 LA GUERRE HYBRIDE ET LA DISSIMULATION ÉTATIQUE PAR PROCURATION

Nature de la mutation

L'espace numérique est devenu le théâtre privilégié d'une conflictualité de haute intensité qualifiée de « guerre hybride ». Cette doctrine offensive, particulièrement mise en pratique par des adversaires stratégiques de l'Europe, vise à déstabiliser un État cible tout en maintenant les opérations sous le seuil du conflit armé ouvert. Pour garantir un « déni plausible » et masquer l'attribution politique de ces attaques, on assiste à une érosion volontaire et systématique des frontières traditionnelles entre les services de renseignement étatiques et l'écosystème cybercriminel. Des États comme la Russie ou l'Iran peuvent désormais sous-traiter certaines de leurs opérations d'espionnage, de sabotage ou d'influence à des mercenaires numériques, des entreprises privées de sécurité offensive, des groupes d'hacktivistes ou des affiliés de rançongiciels. Ces acteurs agissent comme de véritables *cyber proxies*, des mandataires qui créent un brouillard technologique et organisationnel et complexifient le processus d'attribution des attaques.

Illustration

Le *Panorama de la cybermenace 2025* de l'ANSSI objective largement l'ampleur de cette sous-traitance étatique. Dans le contexte des tensions géopolitiques européennes et de la guerre en Ukraine, des groupes d'hacktivistes pro-russes, comme le groupe *NoName057*, mènent des campagnes massives de déni de service distribué contre des institutions européennes, servant de supplétifs pour des opérations de déstabilisation de basse intensité. Parallèlement, des fuites de données internes récentes ont confirmé l'existence de liens interpersonnels étroits et d'une protection de fait accordée par les services de renseignement russes à des groupes cybercriminels majeurs, à l'image du groupe de rançongiciel BlackBasta.

Cette logique de mercenariat d'État structure également l'écosystème offensif chinois à travers une industrie florissante de la lutte informatique offensive privée (LIOP). Des fuites massives de documents

internes de prestataires commerciaux en 2024 et 2025 ont formellement mis en évidence que des sociétés opéraient directement au profit des services de renseignement de l'État chinois. La République islamique d'Iran adopte des stratégies asymétriques similaires, s'appuyant sur des écosystèmes hybrides pour cibler l'économie des pays occidentaux sous le couvert d'un détournement d'outils d'administration légitimes.

La campagne de compromission visant les environnements Salesforce à l'été 2025, via l'exploitation de l'outil tiers Drift, illustre cette porosité entre cybercriminalité et objectifs de déstabilisation stratégique. La nature des cibles, qu'il s'agisse de marques de luxe françaises de réputation mondiale ou de leaders mondiaux de la sécurité tels que Zscaler ou Cloudflare, et le ciblage systématique de données stratégiques suggèrent une manœuvre dépassant la simple extorsion de fonds. Cette attaque semble nettement s'inscrire dans une rhétorique de guerre hybride où des groupes criminels agissent comme des *proxies*, et illustre la convergence objective d'intérêts dès lors qu'ils revendiquent explicitement de s'en prendre à des marques représentatives des intérêts occidentaux. Sous cette couverture, ces acteurs réalisent des opérations de fragilisation de la confiance dans l'infrastructure numérique occidentale, remplissant de facto les objectifs de déstabilisation de puissances adverses.

Impact sur notre posture collective

L'utilisation croissante de *cyber proxies* érode la distinction classique entre la réponse à une attaque de droit commun, relevant du domaine judiciaire, et la riposte à une agression étatique relevant de la diplomatie ou des armées. Face à cette combinaison de modes d'action militaires et non militaires, notre doctrine nationale de défense ne peut plus fonctionner en silos administratifs étanches.

Comme le souligne fort justement la nouvelle *Stratégie Nationale de Cybersécurité 2026-2030*, la France doit impérativement mobiliser l'ensemble de ses leviers judiciaires, techniques, diplomatiques, militaires, économiques, sous l'égide du Centre de coordination des crises cyber pour entraver cette menace globale. Cependant, le Cigref rappelle que l'État ne saurait mener cette lutte seul. Les grandes entreprises privées se retrouvant de facto en première ligne face à ces *proxies* agissant pour des puissances étrangères, il est urgent d'intégrer pleinement le secteur privé dans le partage du renseignement technique sur la menace. Maintenir des barrières administratives face à des opérations cyber de nature hybride reviendrait à laisser le tissu économique national démuni. S'il est légitime pour l'État de préserver sa sécurité opérationnelle, il est aujourd'hui essentiel de définir un cadre de partage d'informations sensibles et actionnables en temps réel pour les opérateurs du secteur privé.

2.3 L'ARSENALISATION DE L'IA ET LE PASSAGE À LA « VITESSE MACHINE »

Nature de la mutation

Les systèmes d'intelligence artificielle générative et agentique transforment radicalement l'écosystème de la menace en le faisant basculer à la « vitesse machine ». L'IA n'est plus un simple sujet de prospective technologique. Elle est d'ores et déjà militarisée par les acteurs offensifs, qu'ils soient étatiques ou cybercriminels, et agit comme un multiplicateur de force. Les algorithmes permettent désormais d'industrialiser et d'automatiser des phases entières de la chaîne d'attaque, notamment la reconnaissance des cibles, le développement dynamique de codes malveillants et la génération de campagnes d'ingénierie sociale hyper-personnalisées à très grande échelle. L'asymétrie devient alors cognitive et le temps de conception, d'enchaînement des vulnérabilités et d'exécution d'une attaque

se comprime de manière spectaculaire, saturant de fait les capacités d'analyse humaine et rendant inopérantes nos défenses périmétriques traditionnelles.

Illustration

Le récent rapport *M-Trends 2026* de Mandiant confirme cette réalité opérationnelle. Les experts constatent l'émergence de familles de logiciels malveillants qui interrogent activement des modèles d'IA en cours d'exécution afin d'adapter leur comportement en temps réel et d'échapper aux systèmes de détection. D'autres logiciels malveillants, comme les voleurs d'identifiants, exploitent directement les interfaces en ligne de commande d'IA présentes sur les postes compromis pour cibler et extraire des secrets industriels.

En Allemagne, le BSI qualifie ce basculement de « profond » pour la protection des infrastructures critiques, soulignant que l'IA rend les méthodes d'attaque existantes plus rapides, moins coûteuses et plus faciles à industrialiser. Le CERT-FR précise toutefois que l'IA générative agit aujourd'hui prioritairement comme un « accélérateur de performance et de passage à l'échelle » pour les acteurs avancés, plutôt que comme un pirate informatique à part entière et autonome.

Plus inquiétant encore sur le plan stratégique, cette accélération s'illustre par l'émergence de modèles d'intelligence artificielle dont les capacités deviennent directement pertinentes pour la cybersécurité. La *preview* du modèle frontière Mythos d'Anthropic, présentée au début de l'année 2026, avait marqué une étape importante en révélant des capacités avancées d'identification de vulnérabilités et d'assistance à la construction de chaînes d'exploitation. Face à ces possibilités duales, Anthropic avait initialement choisi d'encadrer l'accès à ces capacités dans le cadre du projet Glasswing, réservé à un cercle restreint d'acteurs américains.

Au cours des mois suivants, l'entreprise avait engagé une ouverture progressive vers certains acteurs non américains, notamment européens, dans une démarche visant à évaluer les conditions permettant un accès élargi tout en maintenant des mécanismes de contrôle adaptés. La publication simultanée des modèles Fable 5 et Mythos 5 devait constituer une nouvelle étape, fondée sur l'hypothèse que des dispositifs techniques de protection et des garde-fous algorithmiques permettraient d'encadrer leur diffusion.

Cette séquence a toutefois connu une rupture le 12 juin 2026. À 17h21, l'administration américaine a décidé de soumettre ces modèles frontière à des restrictions d'exportation fondées sur des considérations de sécurité nationale. Ne disposant pas immédiatement des moyens techniques permettant d'identifier et de distinguer les utilisateurs relevant de la catégorie des *US persons*, Anthropic a suspendu temporairement l'accès à ces modèles pour les utilisateurs ne relevant pas de cette catégorie, le temps de renforcer ses mécanismes de contrôle. Le Département du Commerce américain a levé ces restrictions d'exportation le 30 juin 2026 et Anthropic a rétabli l'accès aux deux modèles dès le 1er juillet, après avoir renforcé ses mécanismes de distinction entre catégories d'utilisateurs.

Cet épisode constitue un précédent important. Il montre que les modèles d'intelligence artificielle les plus avancés ne relèvent plus uniquement d'une logique de produit numérique commercial. Certaines de leurs capacités peuvent désormais être considérées comme stratégiques et faire l'objet de mécanismes de contrôle comparables à ceux appliqués à d'autres technologies sensibles. Pour les organisations utilisatrices, cette évolution introduit une nouvelle dimension de leur résilience : la

capacité à maintenir dans la durée un accès fiable aux technologies d'intelligence artificielle nécessaires à leur protection.

Impact sur notre posture collective

Cette mutation engendre une dégradation accélérée de la sécurité globale de la méta-structure numérique qui soutient l'ensemble des processus de fonctionnement de la société et de son économie et une transformation profonde des conditions dans lesquelles les organisations publiques et privées pourront assurer leur cybersécurité. Comme pour d'autres technologies critiques, leur disponibilité et leur continuité d'accès deviendront des facteurs déterminants de résilience. Face à la dépendance systémique de l'économie européenne à l'égard des industries technologiques dominantes américaine et chinoise, et au durcissement des rivalités économiques, le statu quo devient une position intenable.

Ce statu quo nous est d'ailleurs dénié, à notre propre détriment, par nos principaux compétiteurs avec une brutalité décomplexée. L'État et les grandes entreprises ne peuvent plus espérer contrer des menaces algorithmiques à l'aide d'outils d'analyse classiques ou en s'en remettant exclusivement à des partenaires extérieurs qui nous interdisent l'accès à leurs propres boucliers. Il est donc absolument essentiel et d'une urgence absolue que la France et l'Europe se dotent d'une capacité autonome d'IA pour la cyberdéfense. Le développement de modèles spécialisés pour la cybersécurité nécessitera des investissements considérables qu'aucune entreprise, et probablement aucun État européen isolé, ne pourra supporter seul. Il appelle donc une mobilisation collective associant entreprises utilisatrices, États et acteurs industriels afin de faire émerger des capacités adaptées aux besoins de protection des infrastructures critiques européennes.

2.4 LA MASSIFICATION DE L'EXPLOITATION DES FAILLES ZERO-DAY CIBLANT LES INFRASTRUCTURES CRITIQUES

Nature de la mutation

L'exploitation de vulnérabilités informatiques inconnues de leurs propres éditeurs, communément appelées « failles zero-day », a franchi un cap critique. Historiquement, les failles zero-day étaient réservées à une élite d'acteurs étatiques de premier plan en raison des coûts colossaux et de l'expertise de pointe requis pour leur découverte. Aujourd'hui, cette arme s'est tragiquement banalisée. Cette massification s'explique par l'essor spectaculaire d'une industrie privée, dérégulée et très lucrative, de la surveillance commerciale et de la lutte informatique offensive. Ces fournisseurs commerciaux de services d'intrusion opèrent désormais avec une force de frappe financière et technologique qui concurrence, voire déclasse, celle des agences de renseignement étatiques. Ils ciblent méthodiquement les angles morts de nos défenses, notamment les équipements de bordure, et les infrastructures critiques, où les outils de détection classiques sont souvent inopérants. L'asymétrie est totale et permet aux attaquants d'exploiter ces failles de manière invisible, avant même qu'un mécanisme de protection ne puisse être conçu.

Illustration

Les données conjointes du *Panorama de la cybermenace 2025* de l'ANSSI et du rapport *M-Trends 2026* de Mandiant objectivent cette rupture capacitaire de manière alarmante. L'exploitation de vulnérabilités demeure le premier vecteur d'intrusion initial, représentant près d'un tiers des attaques.

Le véritable basculement réside dans l'effondrement du *Time to Exploit* (Cf. mutation 1). L'attaque précède ainsi systématiquement la publication des correctifs de sécurité par l'éditeur.

Les analyses mondiales de la menace publiées en 2026 par les équipes de Google soulignent que les fournisseurs commerciaux de logiciels espions ont, pour la première fois de l'histoire, surpassé les États, y compris la Chine ou la Russie, dans la première exploitation de ces failles zero-day. L'ANSSI documente ainsi la compromission massive d'équipements de sécurité ou d'applications d'entreprise avec des failles *zero-day* par ces acteurs privés, qui fournissent ensuite des armes d'intrusion clés en main à leurs clients.

Impact sur notre posture collective

L'industrialisation de ces arsenaux par le secteur privé offensif rend les cycles traditionnels exclusivement fondés sur la gestion des vulnérabilités et la remédiation réactive structurellement inadaptés et définitivement caducs. Cette mutation est accentuée par une institutionnalisation de la lutte offensive privée, dont témoigne notamment la nouvelle stratégie cyber du Président Trump, publiée en mars 2026. Celle-ci prévoit explicitement de créer des incitations pour permettre au secteur privé d'identifier et de perturber les réseaux adverses.

Face à cette privatisation de la conflictualité numérique, la France s'est investie dans le Processus de Pall Mall. Initié avec le Royaume-Uni lors du sommet bilatéral de mars 2023 et officiellement lancé à Londres en février 2024, ce processus vise à établir un cadre international pour lutter contre la prolifération et l'usage irresponsable des capacités commerciales de cyber-intrusion. Malgré le soutien de 27 États à un code de bonnes pratiques en août 2025, l'omniprésence de failles zero-day exploitées à l'échelle industrielle démontre que notre posture nationale, fondée sur la réaction après divulgation, atteint une limite critique. La stabilité de nos infrastructures économiques est désormais exposée à une menace offensive privée capable de saturer en continu les capacités de défense classiques de nos organisations.

2.5 LA SATURATION SYSTÉMIQUE PAR L'HYPER-VOLUMÉTRIE

Nature de la mutation

L'industrialisation massive des cyberattaques, propulsée par l'automatisation et l'intelligence artificielle, engendre dans de nombreux cas une submersion des capacités de défense. Cette mutation ne repose plus uniquement sur la sophistication technique, mais sur une hyper-volumétrie d'événements malveillants qui épuise les ressources humaines au sein des *Security Operations Centers*. Ce déluge est d'autant plus indétectable qu'il s'appuie désormais sur le détournement de sessions authentifiées au sein des environnements Cloud et via les API. On estime désormais que 95 % des attaques proviennent de sessions légitimes en apparence, rendant la distinction entre flux licites et malveillants quasiment impossible pour un opérateur humain sans assistance algorithmique massive.

Illustration

Le rapport *M-Trends 2026* de Mandiant met en évidence le phénomène critique de fatigue liée aux alertes, où le volume massif d'événements, même de faible criticité, conduit les équipes de cybersécurité à une saturation telle que des alertes majeures finissent par être ignorées ou traitées trop tard. Cette industrialisation est confirmée par l'effondrement du délai moyen de transmission

d'un accès malveillant entre deux groupes criminels, passé de plus de 8 heures en 2022 à moins de 30 secondes aujourd'hui.

Le *Panorama de la cybermenace 2025* de l'ANSSI confirme cette tendance, notant que la menace est désormais systémique et touche l'ensemble du tissu économique. L'Agence observe une multiplication des exfiltrations de données résultant du vol de secrets d'accès. En 2025, bien que la majorité des organisations détectent encore les intrusions en interne, la vélocité des attaquants pour prendre le contrôle administratif d'un environnement se compte désormais en heures, devançant, dans la plupart des cas documentés, la capacité de réaction humaine.

Impact sur notre posture collective

L'hyper-volumétrie rend le paradigme actuel de la cyberdéfense, fondé sur l'analyse humaine en temps réel, structurellement inadapté. Comme le Cigref l'indiquait déjà au Premier ministre dès novembre 2020, les ressources humaines et financières croissantes consacrées à la sécurité sont de plus en plus distraites de la capacité d'innovation des entreprises pour compenser l'efficacité sans cesse renouvelée d'une menace industrialisée.

Cette impasse est désormais reconnue à l'échelle internationale. La nouvelle stratégie cyber des États-Unis de mars 2026 prévoit explicitement l'adoption de solutions de cybersécurité propulsées par l'IA, selon l'expression retenue, pour assurer une défense à l'échelle nécessaire. Pour la France, cette saturation signifie que la protection de la Nation ne peut plus reposer sur la seule vigilance administrative ou l'engagement individuel des analystes cyber. Elle exige une mutation vers une défense augmentée et automatisée, capable de traiter la menace à la « vitesse machine », sous peine de voir notre tissu économique durablement submergé par des attaques utilisant nos propres infrastructures de confiance pour nous aveugler.

2.6 LA CONVERGENCE VERS UNE GUERRE COGNITIVE ET UNE PARALYSIE STRATÉGIQUE

Nature de la mutation

Les cyberattaques ne se limitent plus à la stricte compromission technique, à l'espionnage ou à l'extorsion financière. Elles s'inscrivent désormais dans une manœuvre d'ensemble relevant de la guerre cognitive. Cette dimension subversive cible directement la cohésion économique, l'industrie de défense européenne et les infrastructures d'importance vitale. L'objectif de ces campagnes n'est plus seulement de détruire ou de rançonner, mais de déstabiliser le corps social et institutionnel. En instillant un climat de menace permanente et insaisissable, nos adversaires, et parfois nos concurrents économiques, cherchent à imposer une paralysie stratégique et psychologique. Il s'agit d'éroder la confiance des citoyens et des acteurs économiques dans la capacité de l'État et des grandes entreprises à les protéger.

Illustration

Le *Panorama de la cybermenace 2025* de l'ANSSI fait état de cette porosité croissante entre les cyberattaques et les opérations d'influence et de déstabilisation. Le ciblage répété d'hôpitaux, de collectivités territoriales, de prestataires de services numériques ou d'opérateurs de services essentiels par des groupes hybrides ou des *cyber proxies* dépasse largement le simple mobile crapuleux. L'écho médiatique immédiat de ces paralysies opérationnelles, à l'image de la récente attaque ayant visé

l'ANTS, agit comme une caisse de résonance. Chaque attaque réussie contre un maillon de notre chaîne de valeur participe à la construction d'un récit de vulnérabilité systémique occidentale, minutieusement exploité par des puissances adverses pour fragiliser nos institutions et décourager l'innovation.

Ce récit de nos vulnérabilités, particulièrement en France, est également instrumentalisé, ces derniers temps, à « bas niveau de bruit » par les grandes entreprises de la technologie américaine, notamment auprès des instances européennes à Bruxelles. Elles présentent les aspirations françaises à la souveraineté numérique comme une « lubie » coûteuse dont le prix se paie par une vulnérabilité accrue. En expliquant que seules les solutions des géants américains sont en mesure de garantir la sécurité des institutions et de l'économie européenne, ces acteurs déplacent le combat sur le terrain de la légitimité politique. Chaque attaque réussie devient ainsi un argument en faveur d'une dépendance technologique accrue, transformant une défaillance de sécurité en un renoncement potentiel à notre autonomie stratégique.

Impact sur notre posture collective

Face à cette guerre cognitive, le traitement des cyberattaques ne peut plus être relégué au seul rang de fait divers numérique ou de crise strictement informatique et judiciaire. Traiter ces incidents en silos techniques revient à ignorer la manœuvre stratégique globale dont ils ne sont que le vecteur. L'État et les grands acteurs économiques doivent renforcer la dimension communicationnelle et psychologique dès la conception de leur gestion de crise.

Cela implique de rompre avec une certaine culture du secret, de l'opacité et du silence institutionnel, qui ne fait que nourrir l'anxiété et la rumeur. Seule une posture de transparence assumée, de partage d'informations canalisé et de communication coordonnée au plus haut niveau de l'État et des décideurs économiques permettra de restaurer la confiance et de déjouer la paralysie stratégique et politique recherchée tant par les attaquants que par tous ceux qui ont un intérêt à exploiter les effets des crises cyber qui se succèdent.

2.7 LES MENACES DORMANTES SYSTÉMIQUES

Nature de la mutation

Les cyberattaques visibles, telles que les rançongiciels, les vols de données massifs ou les dénis de service, forment un bruit de fond permanent. Elles peuvent constituer, pour certaines d'entre elles, des manœuvres de diversion qui masquent une menace d'une tout autre nature, celle de l'infiltration silencieuse et du pré-positionnement à long terme d'acteurs étatiques ou para-étatiques au cœur de nos infrastructures technologiques critiques. Ce « rideau de fumée » médiatique et opérationnel détourne l'attention des défenseurs vers la gestion de crises immédiates, pouvant laisser, dans certaines circonstances, le champ libre à des opérations préparatoires d'activité d'espionnage ou de sabotage. L'objectif n'est pas la perturbation instantanée, mais la préparation d'une capacité d'observation ou de paralysie coordonnée, activable à l'heure du choix de l'adversaire, notamment en cas de basculement géopolitique majeur.

Illustration

Le rapport *M-Trends 2026* de Mandiant confirme cette persistance extrême. Alors que le délai médian de détection global est de 14 jours, les intrusions à des fins d'espionnage affichent une présence

médiane de 122 jours avant d'être identifiées. Plus frappant encore, des groupes comme UNC6201, lié à la Chine, maintiennent des accès persistants dépassant les 390 jours en ciblant les couches invisibles de l'infrastructure, comme les serveurs de virtualisation et les équipements de bordure, souvent dépourvus d'outils de détection classiques.

Ces attaquants privilégient l'utilisation d'outils d'administration légitimes déjà présents dans les systèmes pour se fondre dans l'activité normale de l'entreprise et échapper aux signatures de malwares. Le *Panorama de la cybermenace 2025* de l'ANSSI alerte également sur ce risque de pré-positionnement au sein des infrastructures d'importance vitale des secteurs de l'énergie, des télécoms ou des transports. De son côté, la stratégie cyber américaine de mars 2026 acte cette réalité en priorisant désormais la traque active des acteurs adverses avant même qu'ils ne déclenchent leur phase offensive.

Impact sur notre posture collective

L'hypothèse d'un pré-positionnement latent de capacités offensives au sein de nos infrastructures constitue un risque stratégique majeur. S'il demeure par nature complexe d'en mesurer l'ampleur exacte à l'échelle nationale, la convergence des analyses des organismes d'expertise impose d'intégrer cette menace systémique dans notre doctrine et notre posture de sécurité numérique.

L'ANSSI a d'ores et déjà pris la mesure de cet enjeu en menant des opérations de recherche de compromissions proactives et ciblées. Toutefois, face à la forte probabilité de ces menaces dormantes, le défi réside désormais dans le passage à l'échelle de cette traque dans l'ensemble du tissu économique critique.

Pour détecter ces présences furtives avant qu'elles ne soient activées, de nombreuses grandes entreprises expriment le besoin d'accès aux signaux faibles. Dès lors, la rétention d'informations concernant des modes opératoires émergents, souvent justifiée par la préservation d'enquêtes de long terme ou par une culture stricte de la classification, risque de nous maintenir collectivement dans l'angle mort de la menace. Seul un partage fluide et en amont des marqueurs techniques entre l'État et certains acteurs privés essentiels, dans un cadre de confiance à définir, permettra de renforcer notre capacité collective à détecter ces infiltrations silencieuses et en prévenir les effets.

2.8 LA CADUCITÉ DE LA CONFIANCE IMPLICITE ET L'EXACERBATION DES RIVALITÉS SYSTÉMIQUES ENTRE ALLIÉS

Nature de la mutation

Le paradigme de la solidarité numérique automatique entre alliés historiques s'efface au profit d'une logique de compétition systémique, tant géopolitique qu'économique. Dans cet espace numérique dépourvu de frontières physiques, la distinction entre protection mutuelle et captation d'intérêts stratégiques est devenue poreuse. Cette mutation n'est pas le fruit d'un choix de la France ou de l'Europe, qui ont longtemps œuvré pour une coopération multilatérale stable, mais d'une réalité subie. La dépendance technologique n'est plus seulement un risque opérationnel. Elle constitue désormais un levier d'influence et de pression exercé par nos propres alliés. La « confiance implicite », autrefois pierre angulaire de la sécurité occidentale, se révèle caduque face à des doctrines nationales privilégiant ouvertement la domination technologique et l'intelligence économique agressive, transformant nos réseaux en théâtres de rivalités de puissance.

Illustration

La stratégie cyber du Président Trump pour l'Amérique, publiée en mars 2026, explicite ce basculement doctrinal avec une clarté et une brutalité inédites. Cette stratégie ne se limite plus à la simple protection des réseaux, mais assume l'usage de capacités non-cinétiques inégalées pour défendre les intérêts nationaux. L'appel à se détourner des fournisseurs étrangers pour promouvoir exclusivement les technologies américaines, ainsi que l'incitation faite au secteur privé à mener des actions offensives pour perturber les réseaux adverses, marquent une rupture profonde avec l'esprit de collaboration transatlantique traditionnelle.

Cet affaiblissement de la relation transatlantique se manifeste de manière flagrante dans le domaine de l'IA de défense. L'annonce par l'entreprise américaine Anthropic de son modèle Mythos, doté de capacités autonomes d'exploitation de failles logicielles, s'est accompagnée de la création du *Glasswing Project*. Bien que l'Union européenne demeure un partenaire stratégique essentiel et que son marché représente un peu plus de 30 % du chiffre d'affaires de l'industrie numérique américaine, les industriels européens ont été sciemment exclus de ce consortium militaro-industriel fermé. Cette exclusion délibérée, au profit d'un écosystème exclusivement américain, éventuellement ouvert au cercle des *Five Eyes*, est un sévère coup de canif dans l'interopérabilité de confiance entre l'Europe et les États-Unis.

Ce constat d'un délitement des alliances classiques s'est invité dans les débats du Forum de Davos en janvier 2026. Face à l'effritement du leadership multilatéral exercé par les puissances occidentales sous l'égide des États-Unis, Mark Carney a appelé à une réaction stratégique des puissances moyennes et raisonnables, et à la formation d'une coalition renouvelée de celles-ci. Cette dynamique pourrait être saisie pour réorienter les ambitions exprimées dans le cadre de l'Appel de Paris, lancé par le Président Macron le 12 novembre 2018, en faveur de la confiance et de la sécurité dans l'espace numérique. Cette convergence de vues se cristallise déjà au niveau franco-allemand. Le BSI préconise désormais, parmi ses mesures immédiates, la création d'une capacité européenne d'évaluation des modèles les plus puissants et un accès direct de l'État aux meilleurs outils de confiance pour ne plus dépendre de solutions opaques ou exclusivement extra-européennes.

Impact sur notre posture collective

L'exacerbation de ces rivalités, y compris entre alliés, impose de réévaluer le cadre de notre doctrine de sécurité nationale. Nous ne pouvons plus ignorer la dimension de compétition économique et de captation de données stratégiques par certains de nos partenaires. Si nous regrettons collectivement cet affaiblissement des solidarités historiques, nous ne pouvons cependant pas nous en tenir au *statu quo*.

Cette nouvelle donne exige de bâtir des alliances de seconde génération, fondées non plus sur la confiance par défaut, mais sur la réciprocité, la transparence technique et le respect de la souveraineté de chacun des partenaires. La résilience de notre tissu économique exige que la France et l'Europe transforment cette contrainte en une impulsion pour développer des capacités de défense autonomes, seules à même de garantir notre liberté d'action stratégique.

3 ENSEIGNEMENTS ET PERSPECTIVES

L'analyse croisée des huit mutations documentées dans la présente note démontre à l'évidence que les grandes organisations, publiques et privées, font face à une rupture systémique et irréversible de l'écosystème de la menace. La sécurité numérique et la protection des systèmes d'information se sont définitivement affranchies du seul périmètre de la gestion des risques techniques pour s'inscrire au cœur d'une compétition géopolitique et économique exacerbée, où la vitesse d'exécution algorithmique supplante les capacités de réaction humaine.

Les événements survenus au cours du premier semestre 2026 agissent à cet égard comme un puissant catalyseur. L'irruption de l'intelligence artificielle de pointe dans le domaine cyber-offensif ne se limite pas à une évolution incrémentale des modes opératoires adverses ; elle modifie en profondeur les conditions mêmes de la résilience numérique. L'intégration de ces capacités au sein des chaînes d'attaque acte l'obsolescence mathématique et opérationnelle des postures de défense traditionnelles de nos organisations. Dès lors, disposer d'un accès garanti et pérenne à des modèles d'IA dotés de capacités de raisonnement avancé ne relève plus d'une simple perspective d'innovation technologique, mais s'impose comme une condition sine qua non du maintien en conditions opérationnelles des infrastructures numériques critiques et des systèmes d'information.

Pour appréhender pleinement l'onde de choc de cette nouvelle donne, il convient d'écarter les postures émotionnelles et de porter un regard rationnel sur la réalité stratégique qui s'impose désormais à l'Europe, à son tissu économique et à ses institutions. Tirer les enseignements de ce basculement implique de regarder en face les mécanismes de notre dépendance, la nature régalienne des arbitrages technologiques de nos alliés, et l'urgence des dynamiques de mutualisation nécessaires pour ne pas subir un déclassement définitif.

3.1 ANTHROPIC : DE LA SIDÉRATION À L'ANALYSE D'UNE DÉCISION OBJECTIVE

La décision soudaine de l'administration américaine, le 12 juin 2026, de suspendre l'accès mondial aux modèles frontière Fable 5 et Mythos 5 d'Anthropic, malgré leur rétablissement trois semaines après, a suscité de vives réactions au sein de l'écosystème numérique européen. Elle a révélé qu'une capacité cyber essentielle pouvait être révoquée dans l'heure. Toutefois, pour appréhender l'onde de choc provoquée par cet événement, il est indispensable de faire preuve de discernement et de raison critique.

L'analyse de cette mesure prise dans l'urgence par l'administration Trump montre qu'elle répond à une certaine logique. En publiant ces modèles, l'éditeur américain s'était appuyé sur la promesse de filtres algorithmiques censés sécuriser leur usage. Or, la probable démonstration rapide de la fragilité de ces garde-fous et de la porosité du modèle aux techniques de contournement a changé la donne. Dès lors qu'un modèle de raisonnement avancé fait preuve de vulnérabilités permettant son détournement, il expose la communauté internationale à un risque immédiat d'utilisation à des fins criminelles et de prolifération de capacités offensives vers des puissances étatiques rivales ou hostiles.

Face à cette perspective, l'intelligence artificielle de pointe ne peut plus être considérée comme un simple produit numérique commercial. Le caractère dual de ses usages peut l'assimiler de fait à une cyber-arme pouvant relever de fait d'une logique de contrôle comparable à celle appliquée à certaines technologies stratégiques. Par conséquent, la décision d'une puissance publique de bloquer techniquement et juridiquement la dissémination à travers la planète d'un tel arsenal ne relève en rien d'une anomalie. Elle constitue, au contraire, l'exercice le plus strict et le plus classique d'une responsabilité régalienne de sécurité nationale.

Il convient d'ailleurs de dissocier scrupuleusement cet impératif de sécurité nationale du climat politique américain contemporain. S'il est indéniable que la présidence de Donald Trump déploie une doctrine cyber privilégiant explicitement l'unilatéralisme et la défense exclusive des intérêts nationaux, déduire que cette restriction technologique spécifique serait purement idéologique, discriminatoire ou motivée par des intérêts spéculatifs constituerait probablement une erreur de jugement. Derrière le brouillard politique, l'enjeu de la non-prolifération technologique demeure le ressort le plus rationnel et le plus compréhensible de cette restriction d'accès, et certainement pas, comme certains observateurs ont été tentés de lire cet événement, comme une discrimination punitive à l'égard des alliées européens.

3.2 LA RÉPONSE CHINOISE ET LA PROLIFÉRATION DES MODÈLES FRONTIÈRE EN *OPEN-WEIGHT*

Quelques semaines après la décision de l'administration Trump de suspendre l'accès mondial à Fable 5 et Mythos 5, le 16 juillet 2026, la start-up chinoise Moonshot AI a publié Kimi K3, un modèle de raisonnement de 2 800 milliards de paramètres dont les performances se situent, selon les évaluations indépendantes disponibles, immédiatement derrière Fable 5 et GPT-5.6 Sol. La diffusion complète de ses poids, c'est-à-dire des valeurs numériques internes ajustées progressivement pendant sa phase d'entraînement et qui déterminent la manière dont il traite l'information et produit ses réponses, est annoncée pour le 27 juillet 2026, mais sans que le moindre contrat de licence ait encore été publié. Dès lors, rien ne garantit la pérennité de cette stratégie d'ouverture pour les cycles technologiques futurs. Pékin pourrait tout à fait décréter des dispositions de contrôle à l'export pour bloquer la publication des prochaines générations de modèles.

Cette progression rapide n'échappe pas à la controverse. Anthropic et OpenAI ont publiquement accusé plusieurs laboratoires chinois, dont Moonshot AI, DeepSeek et MiniMax, de recourir à des pratiques de distillation non autorisées pour développer leurs modèles, méthode consistant à entraîner un modèle moins performant à partir des réponses produites par un modèle américain plus avancé afin d'en reproduire les capacités à moindre coût et en un temps réduit. Ce débat, à ce jour non tranché formellement, invite à une prudence méthodologique. La performance affichée par les modèles chinois *open-weight* ne peut être analysée isolément de la question de son origine, ni instrumentalisée sans réserve pour disqualifier a priori la portée de ces avancées.

Un modèle en *open-weight*, comme Kimi K3, met à disposition les paramètres entraînés du réseau de neurones, ce qui permet à quiconque, disposant du matériel de calcul suffisant, de le télécharger, de l'exécuter localement et de l'adapter à ses propres besoins. Cette ouverture ne s'étend pas nécessairement aux données d'entraînement ni à la chaîne complète de production du modèle. L'*open-weight* se distingue ainsi de l'*open source*. Cette nuance a une conséquence directe car un acteur peut

retirer ou affaiblir les garde-fous intégrés au modèle sans avoir besoin de comprendre la manière dont celui-ci a été entraîné. Il lui suffit d'accéder aux poids et de les réentraîner localement à des fins qui n'ont plus à respecter les conditions fixées par l'éditeur d'origine.

C'est précisément ce facteur qui distingue la trajectoire de Kimi K3 de celle de Fable 5. Là où Anthropic a construit ses garde-fous cyber comme une couche de contrôle intégrée à un service accessible par une interface propriétaire et révocable par l'éditeur, la diffusion des poids place ces protections, si elles existent sous une forme comparable, hors de portée de toute révocation ultérieure. Une fois les poids téléchargés par un nombre indéterminé d'acteurs à travers le monde, aucune décision de politique publique, américaine, chinoise ou européenne, ne permettra plus d'en encadrer l'usage. C'est là que réside la limite structurelle de la logique de contrôle à l'export exposée au chapitre précédent, laquelle suppose l'existence d'un point de passage identifiable sur lequel l'autorité publique peut faire porter sa contrainte. Le modèle en *open-weight*, par construction, ne présente plus ce point de passage une fois sa diffusion initiale achevée.

Il serait néanmoins hâtif d'assimiler purement et simplement Kimi K3 à une réplique de Mythos mise à la disposition de tous. Aucune évaluation publique et indépendante n'a, à ce jour, établi que Kimi K3 disposait de capacités offensives comparables à celles ayant motivé la mesure de sécurité nationale du 12 juin. Ce qui est en revanche établi, et qui suffit à justifier une vigilance accrue, tient à la combinaison de la performance croissante des modèles chinois publiés en *open-weight* et de la difficulté structurellement supérieure à y maintenir des garde-fous robustes une fois les poids diffusés.

Cette stratégie de diffusion ouverte ne relève d'ailleurs probablement pas de la seule philanthropie technologique. Elle s'apparente à une manœuvre délibérée visant à fragiliser le modèle économique et l'hégémonie technologique des laboratoires américains. En proposant gratuitement, ou à un coût très inférieur, des modèles dont les performances se rapprochent de celles des meilleures offres propriétaires, les laboratoires chinois compliquent la capacité d'Anthropic, d'OpenAI ou de leurs pairs à amortir des investissements d'entraînement colossaux sur un modèle d'affaires fondé sur l'accès payant et contrôlé. Cette logique de contournement par les prix et par la diffusion massive constitue, en elle-même, un instrument de politique industrielle et stratégique de contestation de la puissance technologique américaine.

Cette dynamique technologique s'accompagne désormais d'une offensive diplomatique. Le 16 juillet 2026, la Chine a réuni vingt-neuf pays à Shanghai pour fonder la *World AI Cooperation Organization*, présentée comme une alternative aux cadres de gouvernance portés par les États-Unis, l'Union européenne et le G7. Alibaba a par ailleurs indiqué, dans le sillage de Kimi K3, que sa famille de modèles Qwen figurait désormais au deuxième rang de plusieurs classements mondiaux indépendants, juste derrière Fable 5. Cette double avancée, technologique et institutionnelle, illustre que la compétition ne porte plus seulement sur la performance des modèles mais aussi sur la capacité à en écrire les règles d'usage à l'échelle mondiale. Elle donne également du poids à l'hypothèse d'un raidissement futur de Pékin sur ses propres exportations technologiques. Plusieurs sources font état de discussions internes entre les autorités chinoises et des acteurs comme Alibaba ou ByteDance sur d'éventuelles restrictions à l'accès depuis l'étranger aux modèles chinois les plus avancés.

Pour les entreprises et administrations européennes, cette évolution confirme que la maîtrise de ces technologies frontière ne peut plus reposer sur l'hypothèse d'un confinement géographique ou juridique durable. Les entreprises dont le patrimoine applicatif repose sur des architectures anciennes,

insuffisamment corrigées ou faiblement supervisées, resteront les plus exposées à une automatisation croissante des tentatives d'intrusion, quelle que soit la nationalité de l'éditeur du modèle employé pour les conduire. C'est ce déplacement de la responsabilité, de la maîtrise technologique sur l'outil vers la résilience de l'organisation qui l'utilise ou la subit, qui doit désormais structurer notre réflexion collective face à la prolifération des modèles frontière en *open-weight*.

3.3 L'INTELLIGENCE ARTIFICIELLE COMME ACTIF STRATÉGIQUE

L'analyse conjointe de la décision américaine et de la prolifération de modèles en *open-weight*, notamment chinois, conduit à un constat plus profond sur la nature même de ces technologies de rupture. L'intelligence artificielle de pointe s'impose désormais comme un composant critique de la protection et du maintien en conditions de sécurité des architectures numériques, et non plus uniquement comme un levier de performance économique. Les modèles frontière entrent progressivement dans le champ des technologies stratégiques dont l'accès n'est plus garanti par les seules règles du marché.

L'actualité récente démontre en effet que les modèles frontière dotés de capacités de raisonnement avancées sont capables d'identifier massivement des vulnérabilités, de détecter des chaînes d'exploitation inédites, d'assister les opérations de remédiation. Cette évolution est désormais inéluctable. Dès lors que les attaquants, qu'ils relèvent de la cybercriminalité organisée ou d'appareils étatiques, disposeront de ces mêmes instruments de ciblage automatisé, il n'existera aucune option soutenable pour l'économie européenne, ses entreprises et ses administrations publiques, voire la souveraineté de ses États membres, sans un accès garanti et pérenne à un modèle frontière de raisonnement spécialisé dans la sécurité numérique.

En outre, la dynamique des modèles *open-weight* chinois place l'Europe devant un choix dont les conséquences dépassent le seul enjeu de performance comparée des modèles. À supposer que les entreprises et administrations européennes prennent l'habitude de fonder leur résilience cyber sur des modèles frontière en *open-weight* faute de disposer de capacités européennes et autonomes équivalentes, elles s'exposeraient à un paradoxe stratégique majeur, consistant à substituer ses dépendances aux laboratoires américains par une dépendance à la politique de diffusion des laboratoires chinois. Rien ne garantit la pérennité de cette politique. Le jour où Pékin estimerait que l'intérêt national commande de tarir ce flux pour réserver ses futures avancées algorithmiques à ses usages strictement interne ou partagés avec ses alliés, par une mesure symétrique à celle prise par Washington le 12 juin dernier, l'Europe, qui se serait ainsi exposée à une nouvelle dépendance, verrait s'interrompre brutalement la mise à niveau des fondations de sa sécurité numérique, sans disposer d'une capacité de développement autonome et à la frontière pour prendre le relais dans les délais nécessaires. L'ouverture actuelle des modèles de pointe chinois ne doit donc pas être interprétée comme une garantie durable, mais comme une fenêtre stratégique dont la fermeture, décidée unilatéralement par un État tiers, demeure toujours possible.

Désormais, à l'instar des équilibres géopolitiques mondiaux, l'espace numérique va se polariser entre les puissances « dotées », au premier rang desquelles les États-Unis et la Chine, et les autres, par analogie avec la dialectique de la dissuasion nucléaire. Ces dernières se trouveront, par le jeu des restrictions d'usage ou des ruptures de service, condamnées à dépendre durablement des premières

pour assurer leur propre sécurité numérique élémentaire. Dans ce contexte, l'Europe ne peut pas ne pas être une puissance « dotée ».

Toutefois, le parallèle avec la dissuasion nucléaire trouve ici une limite structurelle et économique majeure. Si la force atomique est historiquement financée et opérée par le seul échelon régalien pour la défense de la Nation, aucun État européen ne dispose aujourd'hui de la capacité de supporter seul l'effort financier et capacitaire requis pour le développement d'un modèle frontière dédié à la cyberdéfense. L'asymétrie est ici double. D'une part, l'échelle de l'investissement dépasse le cadre d'un budget national isolé. D'autre part, les bénéficiaires de ce bouclier algorithmique ne sont pas la nation dans son ensemble, mais chacune des entreprises et chacun des opérateurs d'infrastructures critiques, publics et privés, qui soutiennent l'activité de l'économie réelle.

Cette analyse rejoint les réflexions développées dans un essai, publié le 8 juillet 2026 dans la revue de géopolitique *Le Grand Continent,* que l'on doit à Tristan Claret-Trentelivres, Raphaël Doan, Aymeric Roucher et Victor Storchan. Dans cet article passionnant et très documenté, ce groupe de jeunes chercheurs et hauts fonctionnaires s'inscrit justement dans cette lecture parallèle avec l'arme nucléaire et le programme de dissuasion du Général de Gaulle. Ils proposent un plan sur trois ans pour que la France et ses partenaires développent un laboratoire d'IA de frontière capable de rivaliser avec les meilleurs laboratoires des États-Unis, afin d'assurer leur indépendance stratégique. L'article est intitulé de manière explicite « Opération Prométhée : la France peut gagner la course à l'IA. Quel en serait le prix ? ». Les auteurs estiment qu'il serait nécessaire de mobiliser environ 170 milliards de dollars dès 2027, puis plus de 300 milliards en 2029, pour un cumul d'environ 700 milliards de dollars sur trois ans, la puissance de calcul concentrant l'essentiel du coût. Rapporté au PIB français, l'effort représenterait entre 4,5 et 8 %, dont 1,5 % d'investissement public par an, ce qui en ferait un engagement d'une échelle historique. Le reste du financement serait assuré par des investissements privés, mais aussi par des capitaux privés ou publics venus d'autres puissances moyennes ayant intérêt au projet, qu'il s'agisse d'en tirer des bénéfices directs ou d'échapper à leur propre dépendance au duopole sino-américain, à l'heure où ces deux puissances pourraient être tentées de restreindre l'accès de leurs rivaux économiques ou géopolitiques aux systèmes d'IA frontière. L'enseignement essentiel à retenir de cet essai, c'est l'ordre de grandeur des investissements à consentir qui dépasse les capacités d'un État européen isolé et appelle une logique de coalition.

3.4 LES CONDITIONS D'UNE RÉPONSE COLLECTIVE EUROPÉENNE

Face au coût historique d'un tel investissement et aux risques systémiques de restrictions d'accès, le principal sujet d'interrogation ne se situe pas à Washington, mais bien de notre côté de l'Atlantique. L'onde de choc de juin 2026 met crûment en lumière l'incapacité chronique des institutions européennes et des États membres à stimuler et à financer l'émergence d'une industrie de l'intelligence artificielle de classe mondiale, capable de rivaliser avec les géants américains. La dépendance actuelle de l'Europe est le reflet exact de ses propres renoncements industriels et de sa longue passivité de consommateur technologique, et certainement pas le produit d'une injustice étrangère qui lui serait faite.

C'est à un sursaut de lucidité collective que cette actualité enjoint désormais les acteurs européens, alors même que l'échelon politique tente de structurer une première réponse réglementaire et financière. Le Plan d'action sur la cybersécurité et l'intelligence artificielle, présenté par la Commission

européenne en juillet 2026, acte formellement ce changement de paradigme. En s'appuyant sur les cadres législatifs existants tels que l'*AI Act*, le *Cyber Resilience Act* et la directive NIS2, ce plan entend doter l'Union de mécanismes de protection à moyen terme. L'ambition institutionnelle vise ainsi à stimuler l'écosystème à travers le déploiement des *AI Factories* et la mobilisation de capitaux propres via le *Scale-up Europe Fund*.

Pour autant, face au temps long de l'inertie publique et à l'urgence opérationnelle quotidienne en matière de résilience, la voie semble désormais se situer dans une inversion pragmatique de la logique institutionnelle classique. Puisqu'il n'existe pas d'alternative viable à court terme sans un accès garanti à un modèle de raisonnement spécialisé en sécurité numérique, l'issue dépend de l'émergence d'une coalition d'entreprises leaders capable de créer, par elle-même, la masse critique indispensable. L'amorçage d'un tel noyau industriel, fondé sur la mutualisation des ressources et une gouvernance partagée entre grands utilisateurs économiques, permettrait d'initier une dynamique concrète que la puissance publique pourra ensuite rejoindre et soutenir, notamment pour l'attribution des capacités de calcul nécessaires. Cette approche terrain offrirait en outre le meilleur levier pour engager des coopérations transfrontalières directes, à l'instar des rapprochements stratégiques indispensables au niveau franco-allemand.

D'ailleurs, la trajectoire imposée par la BCE au titre de ses fonctions de régulateur pour le règlement DORA confirme cette urgence opérationnelle. Bien que le premier bilan des incidents majeurs liés aux technologies de l'information et de la communication, publié en début de mois de juin 2026, indique que les cyberattaques ne représentent que 10 % des cas au sein de la finance européenne, le reste incombant majoritairement à des pannes systèmes, les autorités regardent avec une extrême lucidité la vitesse à laquelle la zone de danger se rapproche. Ainsi, cinq semaines seulement après ce bilan, le Comité européen du risque systémique (CERS) a relevé l'évaluation du risque cyber pour le qualifier de « sévère ». Le même jour, la BCE a adressé une injonction aux 110 plus grandes banques de la zone euro, leur accordant moins de quatre mois pour soumettre un plan d'action de préparation face aux cyberattaques propulsées par l'intelligence artificielle.

Ce basculement institutionnel tient à une réalité opérationnelle parfaitement documentée par les récentes évaluations des modèles d'IA qualifiés d'offensifs. L'intelligence artificielle n'invente pas nécessairement des failles inédites, mais elle industrialise massivement l'exploitation de vulnérabilités jusqu'ici tolérées, à l'image des serveurs exposés, des correctifs tardifs, des comptes de services dormants. Là où les modèles d'ancienne génération produisaient une poignée d'exploits fonctionnels, les modèles frontière de l'été 2026 en génèrent désormais des centaines sur une même cible. Ce que la BCE réclame aujourd'hui, en actant que le délai d'exploitation de masse se mesure désormais en heures, voire en minutes, et non plus en semaines, c'est l'application accélérée et sans faille des fondamentaux de sécurité. Le régulateur financier confirme ainsi que face à un système rigoureusement compartimenté et mis à jour, l'asymétrie de la menace algorithmique peut et doit être contenue, annonçant la fin définitive de l'ère de la tolérance aux vulnérabilités résiduelles ou non documentées.

3.5 INCIDENT HUGGING FACE VERSUS OPENAI, CAS D'ÉCOLE POUR L'EMPLOI DES MODÈLES D'IA EN CYBERSÉCURITÉ

Un événement manifestement inédit est venu tout récemment illustrer de manière très concrète plusieurs des évolutions décrites dans la présente note. Les informations rendues publiques ces derniers jours, successivement par Hugging Face, puis par OpenAI, offrent un retour d'expérience d'un intérêt certain quant à l'usage des modèles d'IA de pointe dans le domaine de la cybersécurité.

Le 16 juillet 2026, Hugging Face a révélé avoir détecté une intrusion massive dans une partie limitée de son infrastructure de production. Selon l'entreprise, l'attaque exploitait plusieurs vulnérabilités de sa chaîne de traitement des jeux de données dans le but d'obtenir l'exécution de code, puis de compromettre plusieurs identifiants techniques permettant des déplacements latéraux entre différents environnements internes. L'analyse conduite par Hugging Face faisait état de plusieurs milliers d'actions successives réalisées au cours d'un même week-end par un système d'agents autonomes, sans qu'il soit alors possible d'identifier le modèle d'IA employé ni l'origine de cette activité. Les investigations n'ont mis en évidence aucune altération des modèles publics, des jeux de données publics ni de la chaîne logicielle de la plateforme.

Quelques jours plus tard, le 21 juillet, OpenAI a indiqué que cette intrusion provenait en réalité de l'activité de deux de ses propres modèles, GPT-5.6 Sol et un modèle expérimental plus avancé, utilisés dans le cadre d'une campagne interne d'évaluation de leurs capacités offensives. Selon l'entreprise, les garde-fous cyber avaient été volontairement désactivés afin de mesurer leur aptitude à découvrir et à enchaîner des vulnérabilités dans le cadre du benchmark ExploitGym. Les modèles seraient parvenus à sortir de leur environnement de test, à obtenir un accès à Internet, puis à conclure que les solutions recherchées étaient probablement hébergées chez Hugging Face. Ils auraient alors compromis successivement plusieurs systèmes jusqu'à accéder à cette infrastructure. OpenAI précise que Hugging Face avait déjà détecté et contenu l'attaque, puis engagé les opérations de remédiation avant que les deux organisations n'établissent le lien entre cet incident et cette campagne d'évaluation.

Au-delà des circonstances particulières de cet événement, les informations publiées par les deux entreprises mettent en évidence plusieurs enseignements d'intérêt général.

Le premier tient à la capacité désormais démontrée des systèmes d'agents autonomes à conduire, avec une intervention humaine limitée, l'ensemble d'une chaîne d'attaque complexe. L'apport principal de l'IA ne réside pas uniquement dans la découverte de techniques inédites d'attaque. Il réside essentiellement dans sa capacité à automatiser l'enchaînement des différentes phases d'une intrusion, à adapter son comportement aux résultats obtenus, à conduire simultanément un très grand nombre d'actions, à exploiter en parallèle de très nombreux environnements d'exécution éphémères, le tout sans fatigue ni ralentissement. Cet épisode confirme ainsi l'accélération du passage à une conflictualité numérique opérant à la « vitesse machine », décrite dans la présente note.

Le deuxième enseignement concerne l'emploi de ces mêmes technologies à des fins défensives. Hugging Face indique avoir utilisé des agents d'IA pour analyser plus de dix-sept mille événements, corréler de nombreux signaux faibles, reconstituer la chronologie de l'incident, identifier les machines et les identifiants compromis, distinguer les opérations pertinentes des nombreux artefacts présents

dans les journaux techniques et accélérer le travail d'investigation. L'IA apparaît ainsi comme un facteur d'accélération autant des capacités de réponse à incident que des capacités offensives.

Le troisième enseignement est probablement le plus significatif pour les grandes organisations membres du Cigref. Hugging Face explique avoir initialement tenté de conduire cette analyse en s'appuyant sur des modèles de pointe accessibles par des interfaces commerciales. Les mécanismes de protection intégrés à ces services ont toutefois refusé de traiter certains logs techniques, ceux-ci contenant, par nature, des commandes d'attaque, des charges d'exploitation ou des éléments de commande et de contrôle. L'entreprise a finalement conduit son investigation à l'aide d'un modèle open-weight exécuté sur sa propre infrastructure, en l'occurrence GLM 5.2 développé par l'entreprise chinoise Z.ai, ce qui lui a permis de conserver l'ensemble des données sensibles dans son environnement de confiance et de poursuivre ses travaux sans être limitée par les politiques de filtrage applicables aux services accessibles en mode SaaS.

Ce dernier point mérite une attention particulière. Il ne remet pas en cause la légitimité des garde-fous intégrés aux modèles accessibles par des interfaces publiques, qui répondent à un objectif de prévention des usages malveillants. En revanche, il montre que ces mécanismes peuvent également empêcher certains usages non moins légitimes de réponse à incident. Pour les grandes entreprises et les administrations, cet épisode conduit ainsi à s'interroger sur les conditions dans lesquelles elles pourront disposer, en cas de crise, d'une capacité d'analyse reposant sur un modèle de très haut niveau, exécutable dans leur propre environnement, immédiatement mobilisable par les équipes de réponse à incident, sans transfert des journaux techniques ou des identifiants sensibles vers un service tiers, et dont le fonctionnement ne soit pas conditionné par les politiques de filtrage d'un fournisseur externe.

Cet événement offre ainsi une illustration particulièrement concrète de plusieurs des évolutions analysées dans cette note. Il confirme que l'IA de pointe devient progressivement une composante essentielle des capacités de cyberdéfense des grandes organisations. Il montre également que la question stratégique ne porte plus seulement sur les performances des modèles. Elle porte également sur les modalités de leur déploiement, sur leur disponibilité immédiate au moment où survient un incident, sur les garanties de confidentialité offertes pendant les opérations de *forensic* et sur les contraintes éventuellement imposées par les politiques de sécurité des fournisseurs.

Même si cet incident qui a impliqué Hugging Face et OpenAI ne relève pas, à proprement parler, de l'activité d'un acteur malveillant, il est probable qu'il ne reste pas longtemps un événement isolé. Il annonce l'entrée rapide dans un nouveau cycle de la cybersécurité où les modèles d'IA de pointe occuperont une place croissante, aussi bien dans la conduite des attaques que dans leur détection et leur remédiation. Cette évolution invite notre écosystème, dès aujourd'hui, à réfléchir aux capacités qu'il conviendra de maîtriser directement afin de garantir, en toute circonstance, l'efficacité de nos fonctions de cybersécurité.

4 UN HORIZON EN CONSTANTE RECONFIGURATION

Par nature, une note d'information et d'actualité capture une réalité à un instant donné, mais la vélocité exceptionnelle des mutations documentées ici confère à cet exercice une obsolescence programmée particulièrement rapide, et l'incident Hugging Face versus OpenAI constitue une excellente illustration de cette situation fortement évolutive. À l'ère de la « vitesse machine » et des arbitrages géopolitiques soudains, les lignes de force observées en cet été 2026 sont susceptibles de pivoter au gré des prochaines vagues de déploiement de modèles frontière ou de nouvelles directives de contrôle des exportations. Cette note ne saurait donc être lue comme un état figé du paysage, mais plutôt dans la dynamique qu'elle suggère. Les analyses produites aujourd'hui devront être réinterrogées en continu, tant la porosité entre l'innovation technologique, la cybercriminalité et les intérêts des puissances crée un environnement d'une instabilité inédite.

L'avenir immédiat, d'ici la fin de l'année 2026, fera office de crash-test pour mesurer la résilience des grandes organisations face à ce changement d'ère. Le premier jalon critique se cristallisera autour de l'échéance du 31 octobre 2026 imposée par la BCE. L'obligation faite aux plus grands établissements bancaires européens de livrer un plan d'action face aux attaques par IA va agir comme un puissant révélateur, et, nous pouvons le souhaiter, comme catalyseur. Elle va en effet inciter l'ensemble du secteur à évaluer concrètement sa capacité à patcher et à segmenter à des rythmes industriels, créant de nouveaux standards de fait pour les autres secteurs d'activité non soumis à DORA. Les enseignements de cet exercice de standardisation bancaire permettront d'observer si le marché de la protection parvient à absorber l'asymétrie cognitive imposée par les modèles offensifs.

Parallèlement, les prochains mois permettront de juger de la viabilité des réponses structurelles amorcées de part et d'autre de l'Atlantique. Il s'agira d'observer si le Plan d'action de la Commission européenne parvient à dépasser le stade des intentions réglementaires pour enclencher le déploiement opérationnel de ses plateformes de test et de ses mécanismes de financement. De la même manière, la capacité du tissu économique européen à faire émerger des dynamiques privées de mutualisation, capables de peser face au duopole sino-américain, sera le véritable indicateur du sursaut attendu.

Les observations que nous avons recueillies au cours du premier semestre 2026 conduisent à un constat simple. L'intelligence artificielle devient un élément essentiel des dispositifs de cybersécurité des grandes organisations. Les modèles les plus performants participent déjà à leur protection. D'une activité humaine progressivement augmentée par la technologie, la cybersécurité est devenue une activité algorithmique supervisée par l'homme. Les conditions d'accès à ces capacités relèvent autant de décisions industrielles que d'arbitrages de sécurité nationale. L'intelligence artificielle constitue désormais une infrastructure critique de la cybersécurité. La résilience numérique dépend donc aussi de la capacité des organisations à disposer, dans la durée, de garanties d'accès aux modèles et aux capacités de calcul qui rendent cette protection possible.

Cigref
RÉUSSIR
LE NUMÉRIQUE

Le Cigref est un réseau de grandes entreprises et administrations publiques françaises qui a pour mission de développer la capacité de ses membres à intégrer et maîtriser le numérique. Par la qualité de sa réflexion et la représentativité de ses membres, il est un acteur fédérateur de la société numérique. Association loi 1901 créée en 1970, le Cigref n'exerce aucune activité lucrative.

Pour réussir sa mission, le Cigref s'appuie sur trois métiers, qui font sa singularité.

Appartenance
Le Cigref incarne une parole collective des grandes entreprises et administrations françaises autour du numérique. Ses membres partagent leurs expériences de l'utilisation des technologies au sein de groupes de travail afin de faire émerger les meilleures pratiques.

Intelligence
Le Cigref participe aux réflexions collectives sur les enjeux économiques et sociétaux des technologies de l'information. Fondé il y a près de 50 ans, étant l'une des plus anciennes associations numériques en France, il tire sa légitimité à la fois de son histoire et de sa maîtrise des sujets techniques, socle de compétences de savoir-faire, fondements du numérique.

I**nfluence**
Le Cigref fait connaître et respecter les intérêts légitimes de ses entreprises membres. Instance indépendante d'échange et de production entre praticiens et acteurs, Il est une référence reconnue par tout son écosystème.

Cigref
RÉUSSIR
LE NUMÉRIQUE

www.cigref.fr
21 av. de Messine, 75008 Paris
+33 1 56 59 70 00
cigref@cigref.fr