

FICHE DE VIGILANCE

IA et CYBER

Protéger et défendre le numérique face à l'IA

L'IA irrigue déjà les méthodes de travail, notamment dans le domaine de l'exploitation de données massive. L'émergence des grands modèles de langue (LLM) aux capacités avancées (dits modèles frontière) constitue une **rupture stratégique** car ils permettent de décupler la productivité. Mais, sans maîtrise des nouveaux risques induits, ils peuvent davantage impacter nos systèmes et notre souveraineté numérique.

Il est donc essentiel d'appréhender ce nouveau contexte en combinant une **utilisation responsable de l'IA** pour tirer profit des opportunités associées, tout en initiant le **renforcement des postures de cyberprotection et de cybersécurité** nécessaires pour détecter et réagir aux nouveaux types de menaces qui se profilent.

Les nouveaux dangers

L'intelligence artificielle était déjà utilisée par des acteurs malveillants pour casser des mots de passe (*brute force*) et générer du contenu de qualité, que ce soit des mails d'hameçonnage, des fausses revendications de fuites de données, ou encore pour recycler d'anciennes fuites de données. Les capacités disruptives de l'IA présentées dans les médias ne sont pas un simple coup de communication mais une réalité qu'il serait risqué de sous-estimer.

L'avènement des derniers modèles révolutionne ainsi la menace :

- **Démocratisation, prolifération, automatisation et massification** : des techniques complexes sont désormais à la portée d'attaquants débutants (*script kiddies*). Elles pourraient expliquer la densité croissante des attaques observées depuis un an. L'hypervolumétrie des attaques va engendrer une saturation des mécanismes de défense. L'actualité récente d'une IA ayant déjoué de sa propre initiative les mécanismes de sécurité l'encadrant pour accéder à internet afin d'attaquer un site tiers est illustrative des capacités actuelles et des évolutions à anticiper.
- **Augmentation de la vitesse d'exécution** : L'exemple du modèle « *Mythos* » d'*Anthropic* est, ici aussi, édifiant : découverte de nouvelles vulnérabilités, réalisation en temps machine des outils d'exploitation et mise en œuvre. Ce qui prenait des semaines est mis en œuvre en quelques heures. Cette vitesse de réalisation surpasse les capacités de détection et de contre-mesures actuelles car elles nécessitent des délais incompressibles d'analyse par les SOC d'application de correctifs (tests de non-régression, déploiement en environnement de pré-production nécessaire pour garantir leur innocuité avant déploiement...).

Maîtrise de l'IA

Si l'IA révolutionne la menace cyber et rend possible un « *Peal Harbor* numérique », elle est aussi un outil indispensable pour la productivité et pour la défense. Les choix des architectures et la mise en œuvre de ces outils doivent toutefois être maîtrisés car leur utilisation est susceptible de modifier sensiblement la surface d'attaque et de générer de nouvelles vulnérabilités.

Cette maîtrise s'appuie sur la mise en œuvre de mesures organisationnelles et techniques relatives aux observations suivantes :

- La mise à disposition d'assistants IA performants intégrés à des environnements Cloud ne doit pas occulter les risques générés par l'externalisation de l'hébergement des données et des services. L'hébergement de modèles d'IA, notamment chez des acteurs extra-territoriaux, expose le savoir-faire, les développements et les architectures internes. Il est ainsi préférable de choisir des tiers de confiance pour l'hébergement de modèles et de mener en amont une analyse de risques portant sur le choix du fournisseur.
- L'intégration de solutions d'IA dans un système d'information est susceptible d'en altérer la sécurité. Les risques liés à cette intégration doivent donc être évalués et assumés, qu'il s'agisse de risques spécifiques à l'usage de l'IA (données d'entraînement : empoisonnement ou extraction, altérations aux interfaces...) ou induits sur le reste du système d'information (regroupement des données, interconnexions...).
- Les modèles d'IA, quels qu'ils soient, doivent faire l'objet d'une confiance limitée. Les performances d'un modèle sur les tâches métiers doivent ainsi être mesurées et régulièrement revues et critiquées pour détecter les potentiels

biais. Les productions d'un modèle d'IA doivent ainsi être vérifiées et approuvées : elles restent de la responsabilité exclusive de l'utilisateur.

- Les usages de l'IA avec une forte interaction utilisateur (type *vibe/agent coding*) doivent faire l'objet d'un accompagnement et d'une montée en compétence des agents car ils transforment les métiers historiques. La formation et la sensibilisation des utilisateurs des systèmes d'IA couvriront notamment la prévention du *shadow-IA*, les restrictions dans les usages (liées par exemple à la sensibilité des données) et le périmètre de responsabilité de l'utilisateur. Plus généralement, les conditions d'emploi d'un système d'IA doivent être explicitées afin de limiter les erreurs de manipulation.
- Une veille technologique renforcée doit permettre de suivre les dernières innovations, méthodes et modèles et une organisation facilitant l'expérimentation est conseillée.

Mesures de sécurité préconisées

Au-delà de la prise de conscience de cette menace, l'IA ne remet pas en cause les mesures de cybersécurité préconisées jusqu'ici. Bien au contraire, leur intérêt en sort renforcé et impose une évolution rapide des postures de sécurité numérique afin de **garantir la sécurisation des systèmes d'information** d'entreprise (et de leurs éventuelles interconnexions avec les systèmes étatiques), **des chaînes de production logicielle et, in fine, la sécurisation des livraisons**. Elle repose sur le triptyque protection / détection / réaction.

Les capacités de l'IA rendent imparables tout manquement aux principes de base de la **cyberprotection** :

- l'hygiène numérique des agents (surtout ceux disposant de privilèges comme les administrateurs, ou les dirigeants) doit être irréprochable, car leurs manquements sont scrutés en permanence pour être exploités par ces nouveaux modes d'attaques. La sensibilisation doit s'accompagner de sanctions en cas de manquements avérés ou répétés pour ne pas mettre en péril le fonctionnement de votre organisme.
- la cartographie, l'homologation et la maîtrise dynamique des systèmes d'information, notamment pour corriger de manière proactive et réactive les vulnérabilités publiques et sécuriser les configurations, sont cruciales pour réduire l'exposition globale de votre organisme à ces nouvelles menaces. L'exposition sur internet des systèmes d'information au strict nécessaire participe également à cette réduction d'exposition et permet de limiter le nombre de cibles potentielles à défendre.
- Les architectures des SI doivent permettre de réduire la surface d'attaque et de ralentir un éventuel attaquant pour faciliter la détection de ses actions et laisser le temps aux groupes d'intervention cyber d'agir. Elles devront dès à présent privilégier les approches zéro-trust et la mise en œuvre de solutions techniques telles que l'authentification multifacteurs (MFA). Ces architectures devront être systématiquement évaluées par des audits et des tests d'intrusion réguliers manuels et automatiques pour éprouver le système et le corriger avant que les attaquants n'exploitent ces failles.

Enfin, dans les domaines de la **cyberdéfense** et de la réaction, les capacités de détection (sonde, EDR), de supervision (SOC) et de remédiation (GIC) doivent être adaptées face à des attaques à vitesse machine. Si les solutions techniques ne sont pas encore matures, il est nécessaire que les entités mènent dès à présent les réflexions sur leur architecture de supervision de leurs systèmes et mettent en œuvre au plus vite les outils nécessaires à sa mise en œuvre efficiente. Ce nouvel outil de la menace qu'est l'IA doit désormais être pris en compte dans les analyses de risques et les plans d'actions en découlant.