Livre blanc

La cybersécurité à l'ère de l'IA

Ondrej Kubovič et René Holt

ESET
Digital Security
Progress. Protected.

Table des matières

Digital Security
Progress. Protected.

Introduction

L'intelligence artificielle (IA) est omniprésente dans les médias et les discussions actuelles. En 2024, une simple recherche en ligne du terme "IA" génère près de 18 milliards de résultats, montrant l'intérêt croissant du public.

Cet intérêt est majoritairement dû aux grands modèles de langage (LLM), des systèmes avancés qui utilisent des algorithmes sophistiqués pour comprendre et générer du texte en langage naturel. Ils peuvent rédiger des articles, répondre à des questions, traduire des langues, et même créer du contenu original, imitant ainsi de manière impressionnante les capacités de communication et de créativité humaines.

L'IA au service du bien

L'IA contribue de manière significative à l'amélioration de la cybersécurité en détectant les menaces, en filtrant et en analysant les données de Threat Intelligence, et en fournissant aux défenseurs des outils plus performants. Les progrès des IA génératives et des grands modèles de langage (LLM) permettent une meilleure compréhension des interactions humaines, offrant ainsi des réponses plus pertinentes et adaptées aux utilisateurs.

Utilisations défensives actuelles de l'IA :

- Traitement de grandes quantités de données
- Corrélation des indicateurs

Utilisations défensives de l'IA en cours de développement ou à venir :

- Création de nouvelles détections basées sur des descriptions de menaces.
- Analyse et explication d'événements passés ou actuels.
- Recherche de vulnérabilités cachées ou inconnues dans un environnement.

Que faut-il savoir de l'IA du point de vue de la sécurité cyber ?

- Les taux élevés de faux positifs réduisent l'efficacité des systèmes de sécurité reposant uniquement sur l'IA.
- Sans corrections et supervision humaine, la performance des systèmes de sécurité basés sur l'IA risque de se détériorer.
- Des modules de formation de haute qualité produisent des modèles d'IA de haute qualité

En voici quelques exemples :

LA THREAT INTELLIGENCE EXPLIQUÉE

La capacité des LLM à distiller des idées simples à partir de documents longs et complexes pourrait s'avérer très utile pour les utilisateurs, les administrateurs informatiques et les hauts responsables. Ces groupes ont souvent des difficultés à comprendre les rapports hautement techniques produits par les professionnels de la sécurité et leurs outils. Avec l'aide de modèles d'IA génératifs, les informations sur les menaces pourraient consister en de courts textes rédigés dans un langage facile à comprendre, mettant en évidence les points clés et les éléments à prendre en compte.

APPREHENDER LA THREAT INTELLIGENCE AVEC L'AIDE DE L'IA

Depuis des années, les organisations de tous les secteurs sont confrontées à des problèmes qui les exposent à des cybercriminels, notamment en raison de la mauvaise configuration des systèmes et des produits de sécurité. Un assistant d'IA intégré qui aide les organisations à optimiser la configuration de leurs environnements et configurer correctement toutes les solutions de sécurité installées pourrait améliorer l'expérience des utilisateurs, les résultats de la détection et la protection en général. Cet assistant pourrait également exécuter des commandes nécessitant plusieurs étapes, telles que l'ajout d'une exemption au pare-feu et la modification des paramètres de la solution de sécurité, ou guider l'administrateur informatique dans le traitement d'une notification du système ou de la sécurité.

FOCUS EXTRÊMEMENT PRÉCIS ASSISTÉ PAR L'IA

Les professionnels de la cybersécurité peuvent grandement bénéficier de l'utilisation de modèles d'IA spécialisés. Actuellement, leurs équipes font face à une surface d'attaque en expansion, à une diversité croissante d'outils de sécurité, et à un flux massif de données sur les menaces. La gestion de toutes les tâches de sécurité est laborieuse, chronophage et nécessite une grande expertise. L'IA peut alléger ce fardeau, réduire la fatigue due aux alertes, et améliorer l'efficacité et la concentration, en hiérarchisant les événements détectés, en les remettant dans leur contexte, en signalant les problèmes les plus urgents et même en résolvant les alertes mineures. Elle pourrait en outre utiliser les événements précédemment recueillis pour générer une cartographie dynamique des éléments et des événements interconnectés, offrant ainsi une vue d'ensemble précieuse pour les équipes d'intervention en cas d'incident et pour les enquêtes qui surviennent ensuite.

CRÉATION DE NOUVELLES DÉTECTIONS

Les chercheurs en sécurité identifient quotidiennement de nouvelles menaces et techniques d'attaque. En intégrant des informations publiques, des indicateurs de compromission, des flux de données sur les dangers et les contributions de l'équipe de sécurité de l'organisation, un modèle d'IA peut générer des détections pertinentes pour analyser les environnements et identifier ces nouvelles menaces. Cependant, afin de minimiser les faux positifs et de mieux trier les incidents, il est essentiel que toutes ces détections soient vérifiées par des experts pour garantir leur fiabilité et leur efficacité.

AMÉLIORATION DE LA SENSIBILISATION DES UTILISATEURS

La sensibilisation à la sécurité est un autre domaine qui pourrait bénéficier de l'IA. Des modèles de génération de contenu pourraient être utilisés pour convertir les articles de presse sur les dernières menaces en mémos internes, ou les transformer en infographies faciles à assimiler. Ils pourraient rédiger de nouveaux emails d'hameçonnage à des fins de tests internes, et proposer également les explications envoyées par la suite aux collaborateurs pour décrire la menace et les risques qu'elle comporte. Afin de rendre la sensibilisation plus ludique, les modèles génératifs pourraient élaborer des quiz et d'autres supports pédagogiques tenant compte de la culture interne, de la terminologie, des produits, des processus et d'autres spécificités de l'organisation en question.

SANDBOX ENRICHIE PAR L'IA

L'IA générative présente un potentiel significatif pour renforcer la sécurité informatique. Elle peut tester les menaces dans un environnement sandbox, analyser le texte des captures d'écran et examiner des événements détaillés, tels que les modifications du registre Windows. De plus, elle peut rédiger des descriptions claires et compréhensibles des menaces identifiées, facilitant ainsi leur compréhension et leur gestion par les équipes de sécurité. Cette capacité à expliquer les comportements observés dans les menaces aide à réagir plus rapidement et efficacement aux cyberattaques.

ANTISPAM ET ANTI-HAMEÇONNAGE AVANCÉS GRÂCE À L'IA

Même les technologies antispam et anti-hameçonnage pourraient bénéficier des avancées récentes de l'IA. En analysant les communications électroniques d'un utilisateur, les modèles d'IA peuvent apprendre à identifier des éléments inhabituels. Un changement soudain dans le ton ou le contenu d'un message peut ainsi signaler une tentative de piratage. Cela permet de prévenir les attaques dans lesquelles se glissent des emails malveillants au sein d'une conversation existante pour paraître légitimes.

ESET utilise l'IA depuis plus de 25 ans

ESET a intégré pour la première fois l'IA dans ses produits en 1997 dans le but d'améliorer principalement la détection des menaces dans les macros des logiciels bureautiques. Voici un résumé de notre expérience avec l'IA :

1997 : PREMIÈRE EXPÉRIENCE – DÉTECTION DE VIRUS DANS LES MACROS

ESET possède une riche expérience dans le déploiement de systèmes basés sur l'IA et le Machine Learning. Nos ingénieurs expérimentent cette technologie depuis la fondation de l'entreprise et l'ont intégrée pour la première fois dans nos produits en 1997 afin d'améliorer la détection des virus dans les macros.

2005 : DETECTION ADN

L'introduction de la Détection ADN a permis de convertir les échantillons en profils ADN pour la détection. Enrichi soit par des systèmes automatisés, soit par des chercheurs humains experts, ce modèle régulièrement mis à jour fonctionne comme un "modèle de Machine Learning en ligne" depuis sa création en 2005.

2006 : BACK-END DE TRAITEMENT DES ÉCHANTILLONS

Inspirée par l'efficacité de nos précédents systèmes basés sur l'IA contre les menaces connues et émergentes, une série de projets internes d'ESET utilisant cette technologie ont suivi, conduisant à l'introduction de back-ends experts pour le traitement à grande échelle de centaines de milliers d'échantillons chaque jour.

2010 : ESET LIVEGRID®

Ces premières expériences liées à l'IA s'étant révélées très prometteuses, ESET a poursuivi dans cette voie et a introduit en 2010 ESET LiveGrid®, son système de réputation dans le cloud. À l'époque, il fournissait déjà rapidement des informations, en quelques minutes au lieu de quelques heures, grâce à la puissance du Machine Learning.

2017–2019 : ESET LIVESENSE

Dès l'arrivée des premiers algorithmes de Deep Learning, ESET a adopté ces nouvelles technologies en les intégrant à ses produits après une phase de test avancée. Deux nouvelles couches de protection ont ainsi été créées : le Machine Learning Avancé dans le Cloud (2017) et le Machine Learning Avancé sur les endpoints (2019). Ces solutions offrent un taux de détection élevé et un faible taux de faux positifs, et sont intégrées à notre technologie multicouche principale, ESET LiveSense, qui s'ajoute à un large éventail d'autres couches défensives développées en interne et supervisées par nos experts.

2018 : ESET LIVEGUARD

Cette sandbox dans le cloud basée sur l'IA offre essentiellement aux utilisateurs toute la puissance de la technologie de détection d'ESET, à la fois interne et dans le produit. Elle est capable d'effectuer une analyse complète à la demande en quelques minutes, voire quelques secondes, et même d'émuler des fichiers dans le temps et sur divers OS.

2020–2021 : MODÈLES BASÉS SUR DES TRANSFORMEURS POUR LA DÉTECTION

A quoi sert un transformeur ? Lors de la saisie d'un message sur un téléphone et que des suggestions de mots apparaissent après chaque mot que vous tapez (par exemple, après "Bonjour, comment ça", le téléphone peut proposer "va", "se" ou "passe"). Cependant, en choisissant simplement parmi ces suggestions, le message final peut manquer de sens, car le modèle ne prend pas en compte le contexte global du message.

Les transformeurs, en revanche, mémorisent le contexte de l'ensemble du texte, ce qui leur permet de produire des messages cohérents et significatifs.

Pour les ingénieurs d'ESET, l'introduction des transformeurs, capables de contextualiser la réflexion de l'IA, a été une opportunité d'améliorer les outils de détection. Les résultats positifs obtenus avec cette technologie ont conduit à son intégration dans nos systèmes de détection multicouches, à la fois dans le cloud en 2020 et sur les endpoints en 2021.

2023 : CRÉATEUR D'INCIDENTS AUTOMATISÉ DANS ESET INSPECT

L'IA est également intégrée au créateur d'incidents automatisé d'ESET Inspect dans notre offre XDR. Cette technologie aide les défenseurs en corrélant des indicateurs extraits d'un grand nombre d'événements collectés sur les endpoints, afin de trier les événements et de regrouper automatiquement les détections associées dans une représentation visuelle. Cela permet de réduire le délai de réponse nécessaire à chaque incident, ainsi que la lassitude ressentie par les équipes de sécurité en raison des alertes ; des équipes qui sont par ailleurs en sous-effectif.

L'IA générative basée sur les transformeurs

L'idée de créer de nouveaux contenus à l'aide de modèles basés sur l'IA existe depuis des années et a été utilisée dans des domaines spécifiques, tels que la chimie computationnelle. Cependant, pour les tâches liées à la sécurité, une approche différente était nécessaire.

En 2017, Google a publié un document intitulé « Attention Is All You Need ». Il propose une nouvelle architecture pour les modèles de Machine Learning basés sur des mécanismes d'attention. Baptisée "transformation", cette architecture s'est révélée très efficace pour traiter le langage naturel et produire une grande variété de contenus compréhensibles par l'homme. En 2022, des modèles tels que ChatGPT, Midjourney et DALL-E ont attiré l'attention du public en montrant qu'avec une simple requête utilisateur, les modèles basés sur des transformeurs sont capables de rédiger un article complet, générer une photo réaliste et produire de nouvelles vidéos. Bien entendu, il ne s'agit là que de la partie la plus visible de leur utilisation.

Chronologie du développement des technologies d'IA chez ESET

1997

Première utilisation des réseaux neuronaux dans les produits ESET, pour la détection des virus dans les macros.

2005

La Détection ADN d'ESET, synonyme de Machine Learning en ligne, utilise les gènes des malwares pour détecter les menaces actuelles et émergentes.

2010

ESET LiveGrid®, un système de réputation dans le cloud, s'appuie sur ESET DNA Detections pour accélérer considérablement les mises à jour côté utilisateur.

2017

Le Machine Learning Avancé d'ESET dans le Cloud utilise l'IA pour alimenter nos systèmes de détection automatisés.

2018

ESET LiveGuard, la sandbox dans le cloud basée par l'IA, fournit une analyse à la demande pour les clients d'ESET avec un temps de réponse de quelques minutes.

2019

Le Machine Learning Avancé d'ESET dans le Cloud dans les endpoints utilise l'IA pour alimenter nos systèmes de détection automatisés.

2020- 2021

Modèles basés sur des transformeurs déployés dans les solutions cloud et endpoint d'ESET.

2023

Le créateur d'incidents automatisé a été ajouté à ESET Inspect pour réduire le bruit et tirer parti des techniques, y compris l'IA, pour produire une vue d'ensemble claire d'un incident.

2024

ESET AI Advisor : lancement d'un conseiller en sécurité reposant sur l'IA générative capable de générer des descriptions détaillées des incidents à partir d'ESET Threat Intelligence, et de fournir des réponses aux requêtes concernant la portée des incidents dans ESET Inspect.

2024 : ESET AI ADVISOR

L'IA étant déjà utilisée dans ESET Inspect pour son créateur d'incidents, l'intégration d'ESET AI Advisor, un assistant d'IA générative, étendra encore les ressources disponibles pour ses utilisateurs. Le personnel de sécurité peut demander à ESET AI Advisor d'interagir directement avec les données d'incident, en utilisant des requêtes en langage naturel pour obtenir des renseignements supplémentaires sur les menaces, tels que le contexte des artefacts détectés, ainsi que les tactiques, techniques et procédures (TTP observées au cours d'un incident).

Pour le personnel moins expérimenté, ESET AI Advisor peut offrir un soutien en répondant aux requêtes relatives à la sécurité. Pour les cadres supérieurs, ESET AI Advisor peut générer des présentations faciles à comprendre, utiles pour communiquer avec d'autres équipes et auprès de personnes ayant différents niveaux de compréhension technique. ESET AI Advisor peut même produire des guides étape par étape pour des groupes spécifiques de collaborateurs dans l'organisation, afin de leur permettre de participer à la prévention et l'atténuation des incidents de sécurité.

L'IA peut également être utilisée avec ESET Threat Intelligence, qui offre une source extraordinairement riche d'informations sur les menaces. Elle détaille les scénarios d'attaque courants et le contexte général d'un large éventail de groupes de pirates, ainsi que leurs TTP, et fournit des indicateurs de compromission (IoC). Les ingénieurs d'ESET ont donné à ESET AI Advisor les moyens d'analyser cette bibliothèque d'informations pour éviter de rechercher une aiguille dans une botte de foin, et ainsi fournir rapidement aux professionnels de la sécurité des données utilisables.

En passant au crible les IoC, les TTP, et les données spécifiques de chronologie, de secteur et de localisation, et en les reliant à des groupes de pirates spécifiques, ESET AI Advisor est capable de fournir des résumés complets, mais toutefois digestes. ESET AI Advisor peut également être utilisé pour créer des rapports destinés à des publics cibles désignés, tels que le personnel informatique, les RSSI et les équipes de sécurité, ainsi que d'autres cadres dirigeants. Pour faire face au risque potentiel d'hallucinations (phénomène où les modèles d'IA générative produisent des informations incorrectes ou inventées) ESET AI Advisor fournit systématiquement des références aux documents source.

ESET AI Advisor utilise également la génération augmentée par extraction pour accéder à la puissance des outils et des données internes développés par ESET, et peut ainsi produire des rapports complets sur les menaces et les incidents.

Cette utilisation intégrée de l'IA permet non seulement de créer des défenses proactives robustes, mais également de fournir une interface intuitive et facile à gérer entre le personnel de sécurité et ses outils.

Même l'IA a des limites

L'IA présente plusieurs limitations qui peuvent affecter la cybersécurité :

Limites de l'IA :

- Les faux positifs et les alertes de faible priorité peuvent entraîner une lassitude et une mauvaise configuration des produits de sécurité.
- En l'absence de mises à jour et de supervision par des experts, les modèles peuvent se dégrader.
- La fiabilité à long terme est essentielle pour les solutions de sécurité, mais elle n'est pas garantie pour les modèles d'IA. Les modèles d'IA générative risquent d'halluciner
  (phénomène où les modèles d'IA générative produisent des informations incorrectes ou inventées)
  Une protection adéquate nécessite d'autres couches et outils de sécurité en plus de l'IA.

LES FAUX POSITIFS COMPTENT TOUJOURS

Lorsque les outils basés sur l'IA identifient à tort un fichier bénin comme malveillant (faux positif), cela peut avoir de graves conséquences, parfois pires qu'un faux négatif. Par exemple, dans le secteur de la fabrication, cela peut entraîner l'interruption de la production, l'endommagement de la chaîne de production, des retards et des pertes financières.

Les faux positifs peuvent aussi fatiguer le personnel de sécurité. Si les fausses alertes sont trop nombreuses, les responsables devront soit passer trop de temps à les gérer, soit relâcher la protection, diminuant ainsi la capacité de détection. Ces scénarios peuvent affaiblir la sécurité de l'organisation et ouvrir de nouvelles voies aux cybercriminels.

LES MODÈLES DE MACHINE LEARNING DÉGÉNÈRENT ET L'IA GENERATIVE DOIT TENIR LE RYTHME

Dans les années 2010, lorsque le Machine Learning est devenu standard dans les produits de sécurité, de nombreux éditeurs ont affirmé que leurs modèles pouvaient traiter les menaces sans mises à jour. Cependant, cette approche a conduit à un taux élevé de faux positifs et à une dégradation des performances. Notre expérience montre que la supervision et l'ajustement continus des modèles de Machine Learning sont essentiels pour leur efficacité.

Pour les grands modèles linguistiques, le langage évolue lentement, donc ils n'ont pas besoin d'être formés aussi fréquemment. Cependant, pour fournir des réponses récentes et détaillées, un LLM doit utiliser la génération augmentée par extraction pour obtenir les informations en ligne ou à partir de sources propriétaires. Sans cela, les modèles peuvent produire des résultats biaisés ou manquer de contexte pour répondre correctement aux requêtes.

QUALITÉ ET FIABILITÉ À LONG TERME

Dans la cybersécurité, la constance des performances et de la fiabilité est essentielle. Si une solution IA détecte bien les menaces une semaine, mais échoue ou génère trop de fausses alertes la suivante, cela augmente la charge de travail des équipes. La supervision humaine est donc cruciale pour maintenir l'efficacité de l'IA sur le long terme. Une formation préalable au déploiement, adaptée aux spécificités d'une organisation, est préférable à un flot de faux positifs ou négatifs.

HALLUCINATIONS DANS LES MODÈLES GÉNÉRATIFS

Ne croyez pas tout ce que vous voyez en ligne, surtout le contenu généré par l'IA. Les modèles actuels peuvent produire des résultats crédibles mais parfois incorrects, incluant des références et des données fabriquées, appelées hallucinations. Une vérification humaine est donc essentielle. Les hallucinations posent des problèmes pour le déploiement de l'IA générative, notamment en cybersécurité, où elles peuvent fausser l'analyse des échantillons et mener à des décisions erronées et dangereuses, compromettant la sécurité.

REMARQUE : Les hallucinations de l'IA générative peuvent être avantageuses dans certains cas d'utilisation. Lorsqu'il s'agit de créer du contenu audio, vidéo ou visuel, l'algorithme doit avoir une "latitude créative" pour produire des idées innovantes ou répondre à une requête d'une manière inédite pour les humains.

L'IA GÉNÉRATIVE SEULE NE SUFFIT PAS

Le déploiement de l'IA générative pour des tâches spécifiques peut être complexe, notamment en raison des exigences précises de l'ensemble de formation. Sans étiquetage et garde-fous appropriés, les modèles peuvent devenir biaisés. En cybersécurité, un ensemble de formation inadéquat peut entraîner des faux positifs ou ignorer des attributs malveillants.

Les cybercriminels et adversaires étatiques ajoutent des couches de compression, d'obscurcissement et de chiffrement pour rendre leur produit invisible. Un modèle d'IA sans outils supplémentaires et supervision humaine ne peut pas surmonter ces obstacles, produisant des analyses de faible valeur. Les attaquants peuvent aussi fractionner leurs malwares, rendant le comportement malveillant visible seulement lorsqu'ils fonctionnent ensemble, ce qui peut tromper même une IA bien formée.

ADVERSAIRES INTELLIGENTS ET ADAPTATIFS

Les machines d'aujourd'hui peuvent vaincre les humains aux échecs et deviennent rapidement plus efficaces dans d'autres tâches. Cependant, la plupart de celles-ci existent dans des environnements avec des règles strictes. Les pirates, quant à eux, n'obéissent pas aux règles et n'acceptent pas les limites. Ils trichent, manipulent ou changent les règles du jeu sans avertissement.

Les voitures autonomes en sont un bon exemple. Malgré d'énormes investissements dans leur développement, ces véhicules dépendent fortement des marquages environnementaux tels que les panneaux de signalisation et les feux de circulation. Un adversaire pourrait attaquer les véhicules sans conducteur en masquant les panneaux de signalisation ou en faisant clignoter les feux de circulation à un rythme incompréhensible pour l'œil humain. Ces déformations pourraient conduire les véhicules à prendre de mauvaises décisions et provoquer des accidents mortels.

Cette nature changeante du paysage des menaces empêche la création d'une solution de protection universelle capable de contrer tous les cyberdangers actuels et futurs. Même les tout derniers modèles d'IA n'y changeront rien.

Menaces basées sur l'IA attendues en 2018 :

- Génération de campagnes d'ingénierie sociale réalistes, y compris d'hameçonnage.
- Optimisation des malwares pour les adapter aux environnements sélectionnés.
- Reproduction et mise en œuvre de fausses preuves.
- Amélioration du ciblage et de la sélection des victimes.
- Recherche de nouvelles vulnérabilités dans les logiciels et les microprogrammes des objets connectés.
- Génération de nouveaux malwares ou réécriture dans différents langages de programmation.
- Déclenchement de mécanismes d'autodestruction dans les malwares en dernier recours pour déjouer les enquêtes et les analyses.
- Réduction de la durée d'une attaque pour raccourcir la fenêtre d'intervention des défenseurs.
- Apprentissage collectif des botnets (IoT).

Menaces basées sur l'IA attendues en 2024 et dans les prochaines années :

- Génération d'une grande quantité de campagnes de spam, d'escroquerie et d'hameçonnage de haute qualité.
- Génération d'une grande quantité de fausses informations, d'images et de vidéos « deepfake » plus vraies que nature pour les utiliser dans des escroqueries, des extorsions et des opérations d'influence.
- Analyse du trafic réseau et des entrées provenant d'appareils compromis, et utilisation ultérieure de ces informations pour dissimuler et protéger l'infrastructure, le code, les opérations et les acteurs malveillants.
- Extraction d'informations légalement protégées, propriétaires ou autrement sensibles connues des modèles génératifs via des requêtes malveillantes spécialement conçues.
- Campagnes d'ingénierie sociale réalistes exploitant la communication humaine générée par les LLM.

L'IA au service du mal

MENACES ACTUELLES EXPLOITANT POTENTIELLEMENT L'IA

En sous-estimant l'usage de l'IA que peuvent faire les cybercriminels et les groupes malveillants les plus sophistiqués, les organisations et leurs experts en sécurité se trouveraient dangereusement pris au dépourvu.

ESCROQUERIES ET SPAM MALVEILLANT

La création de nouveaux contenus malveillants occupe une place de choix dans cette liste. En 2018, c'est surtout la traduction en ligne assistée par l'IA qui alimentait ce type d'activité ; aujourd'hui, armés de LLM capables d'imiter le style d'écriture de n'importe quelle personne, des attaquants peuvent concevoir des campagnes de spam et d'escroqueries avancées qu'il est de plus en plus difficile d'identifier uniquement à partir du contenu de leur message.

CAMPAGNES DE DÉSINFORMATION

Les campagnes de désinformation sont désormais facilitées par les modèles d'IA générative. Il est plus simple de réécrire des articles en ligne et d'y implanter de fausses informations, photos ou vidéos. Ces contenus peuvent avoir un impact majeur sur les réseaux sociaux, où les gens parcourent souvent rapidement les titres et les images.

CONTOURNEMENT DE LA DÉTECTION

Certaines formes d'IA et de Machine Learning pourraient être utilisées pour protéger les infrastructures malveillantes. Emotet était l'exemple même d'un cheval de Troie collectant des données et déjouant la détection. Il tenait les chercheurs en sécurité à distance en analysant chaque victime potentielle pour y déceler des signes de surveillance. L'utilisation d'un modèle de Machine Learning aurait facilité cette tâche autrement gargantuesque pour ces criminels.

DÉTOURNEMENT DE FILS DE DISCUSSION PAR EMAIL

Le retour sur investissement des campagnes d'hameçonnage augmentera considérablement grâce à l'utilisation de LLM formés sur les emails volés, permettant aux attaquants de rédiger des messages crédibles. L'injection de ces messages dans des conversations existantes, appelée attaque sur la chaîne de réponse, accroît significativement leurs chances de succès.

ÉCRITURE DE NOUVEAUX MALWARES

Toutes les menaces ne sont pas aussi sérieuses que certains titres le laissent entendre, comme l'idée que l'IA pourrait écrire des malwares entièrement à partir de rien. Les compétences en programmation de l'IA sont encore limitées. Les modèles génératifs actuels peuvent assister dans la réécriture de bibliothèques, le débogage et l'optimisation de code, mais ils ne sont pas encore capables de créer des logiciels complexes, y compris des malwares, de manière optimale.

Même si l'IA pouvait créer des malwares de qualité, cela n'est qu'une étape. Les attaquants doivent concevoir une stratégie de diffusion, échapper à la détection, monétiser l'accès et parfois continuer à communiquer avec la victime. L'IA peut aider, mais ne peut pas remplacer un adversaire humain intelligent, du moins ce n'est pas encore le cas aujourd'hui.

SCIENCE-FICTION OU FUTUR PROCHE ?

Il convient de souligner qu'au rythme actuel des progrès de l'IA, nous verrons probablement les modèles d'IA s'améliorer dans toutes les activités susmentionnées au cours des prochaines années, voire des prochains mois. Cela nous amène à des scénarios de science-fiction qui ne se sont pas encore concrétisés mais qui pourraient devenir réalité dans un avenir proche.

Mise en place de fausses preuves

Des groupes de pirates pourraient former des modèles d'IA générative à partir des études publiées sur l'activité d'autres groupes afin de mener des campagnes sous une fausse bannière. Cela rendrait encore plus difficile l'attribution des cyberattaques à des groupes spécifiques.

Amélioration de la sélection des victimes

L'IA pourrait devenir le crible capable d'analyser les ensembles de données recueillis lors de la phase de reconnaissance d'une attaque afin d'identifier les cibles les plus intéressantes, qu'il s'agisse d'un salarié négligent ou crédule disposant de larges privilèges sur le système ou d'un sous-traitant dont les systèmes sont mal protégés ou mal configurés.

Recherche de vulnérabilités

Ces dernières années ont montré que les vulnérabilités zero-day constituent une activité lucrative tant pour les groupes qui se consacrent à l'intrusion et l'espionnage que pour les cybercriminels à la recherche d'un gain financier. La formation de modèles basés sur l'IA pour rechercher des failles inconnues et exploitables pourrait ouvrir des portes (dérobées) à presque tous les environnements informatiques de la planète, en particulier si des objets connectés (IoT), notoirement peu sûrs et souvent impossibles à corriger, s'avéraient présents.

Botnets intelligents

Concernant l'IoT, l'IA pourrait aider les auteurs de menaces à développer de nouveaux botnets, plus efficaces et capables d'un apprentissage collectif. Cela signifie que ces grands organismes numériques pourraient mener des opérations plus vastes et plus sophistiquées, telles que la recherche de vulnérabilités ou la collecte d'informations, au lieu d'être simplement utilisés pour leur force de frappe dans des attaques par déni de service distribué (DDoS).

Conclusion

Comme en témoigne ce livre blanc, **l'IA représente une technologie extrêmement bénéfique pour la cybersécurité**. Intégrée aux solutions de sécurité, l'IA peut renforcer les capacités de détection et de réponse aux menaces. Mais elle peut aussi améliorer les connaissances et l'accessibilité des services de renseignements et de recherche de menaces, contribuer à une meilleure protection, et ainsi prévenir les attaques avancées. En réduisant le bruit des alertes et la lassitude associée, elle permet également aux experts en cybersécurité d'identifier les activités malveillantes, et d'y répondre plus rapidement et plus efficacement.

L'adoption de l'IA peut avoir et aura probablement un effet transformateur dans divers domaines de la cybersécurité, comme la création de nouvelles détections, la recherche de vulnérabilités inconnues et l'aide à la configuration correcte des outils de protection. Malgré ces avantages, l'IA présente également des limites et des défis, notamment la nécessité de disposer d'ensembles de formation de haute qualité, le risque de taux élevés de faux positifs et le besoin d'une supervision humaine experte.

Il ne faut pas négliger **les scénarios potentiels de détournement de l'IA par des acteurs malveillants**. Les groupes de pirates peuvent, et certains le font déjà, exploiter cette technologie pour créer des campagnes de spam et d'escroqueries convaincantes, améliorer leur ingénierie sociale, échapper à la détection et la surveillance, et même déboguer et optimiser leurs malwares. Bien que ces menaces soient préoccupantes, ce document précise que **l'IA n'est pas en mesure de remplacer totalement un adversaire humain intelligent**, en particulier pour effectuer des tâches d'accompagnement complexes, par exemple imaginer une chaîne d'attaque efficace ou générer de nouveaux malwares sophistiqués.

Nous avons rédigé ce document pour souligner l'importance de comprendre et d'utiliser l'IA dans le domaine de la cybersécurité, bien que nous soyons conscients des limites et des risques potentiels liés à cette technologie. En bref, nous plaidons en faveur **d'une approche équilibrée qui combine l'IA avec une supervision humaine experte afin de garantir le développement de solutions de cybersécurité efficaces et fiables**.

Annexe A : Terminologie

INTELLIGENCE ARTIFICIELLE GÉNÉRALE (IAG)

Idéal encore inachevé, il s'agit d'un agent artificiel généralement intelligent et autonome, capable de prendre des décisions et d'apprendre de manière indépendante en se basant uniquement sur des données provenant d'environnements réels ou virtuels, sans intervention ni surveillance humaine. L'IAG se concentre sur le développement d'agents capables d'accomplir un large éventail de tâches, par opposition à l'IA « restreinte » qui conçoit des agents pour un ensemble limité de tâches.

INTELLIGENCE ARTIFICIELLE (IA)

L'intelligence artificielle désigne des agents informatiques intégrés dans des logiciels et du matériel, créés pour opérer de manière intelligente dans des environnements spécifiques. Ces environnements limitent les actions possibles, le temps de réaction et les données observables. L'IA démontre des capacités d'apprentissage, s'adapte aux changements de son environnement, évalue les conséquences de ses décisions et choisit des approches appropriées en fonction des objectifs, des connaissances disponibles et des contraintes du moment.

MACHINE LEARNING (ML)

Le Machine Learning concerne principalement la conception et l'utilisation de modèles capables d'analyser de grands ensembles de données et d'apprendre des fonctions capables de prédire les résultats pour de nouvelles entrées. Les modèles inspirés du fonctionnement des neurones dans le cerveau humain sont appelés réseaux neuronaux. Ils se sont révélés très efficaces pour l'apprentissage de combinaisons de fonctions et constituent un puissant outil de prédiction.

IA GÉNÉRATIVE (GenAI)

Les progrès réalisés dans le domaine du traitement du langage naturel et des réseaux neuronaux basés sur des transformations ont conduit à l'essor de l'IA générative. Généralement formés sur de vastes ensembles de données non étiquetées, ces modèles d'IA tirent parti d'interfaces homme-machine simples qui acceptent des requêtes en langage naturel pour générer de nouveaux résultats en peu de temps. Le contenu produit comprend des données statistiques, du texte, des images, du son, de la vidéo et du code source.

Annexe B : Où s'arrête la réalité de l'IA et où commencent les mythes

Lorsqu'un sujet atteint le niveau d'engouement que connaît actuellement l'IA, des mythes commencent inévitablement à apparaître. La cybersécurité n'échappe pas à cette tendance et il existe un certain nombre d'affirmations farfelues qui tentent de tirer profit des cycles médiatiques. Pour ceux qui s'intéressent à la situation réelle dans ce domaine, voici quelques affirmations actuelles sur l'IA qui ont été démenties.

AFFIRMATION : L'IA PEUT ANALYSER DU CODE ET IDENTIFIER SON COMPORTEMENT MALVEILLANT

La réalité : L'analyse de code par l'IA peut être bien structurée mais souvent incomplète, incorrecte ou hors contexte. Seuls des experts peuvent comprendre ces analyses. Si des non-experts les utilisent, cela peut avoir des conséquences graves. Les adversaires peuvent aussi empoisonner ou obscurcir leur code pour tromper l'IA et produire des résultats erronés.

AFFIRMATION : L'IA PEUT ÉCRIRE DE NOUVEAUX LOGICIELS ET DONC DES MALWARES

La réalité : Certains services en ligne utilisent l'IA générative pour créer du code. Cette méthode est utile et efficace lorsqu'elle est appliquée à des tâches ennuyeuses ou moins complexes qui, autrement, occuperaient le temps précieux de développeurs qualifiés. Toutefois, les tests montrent que l'écriture d'un logiciel à partir de zéro est une tâche bien plus ardue et semble trop avancée pour l'IA contemporaine.

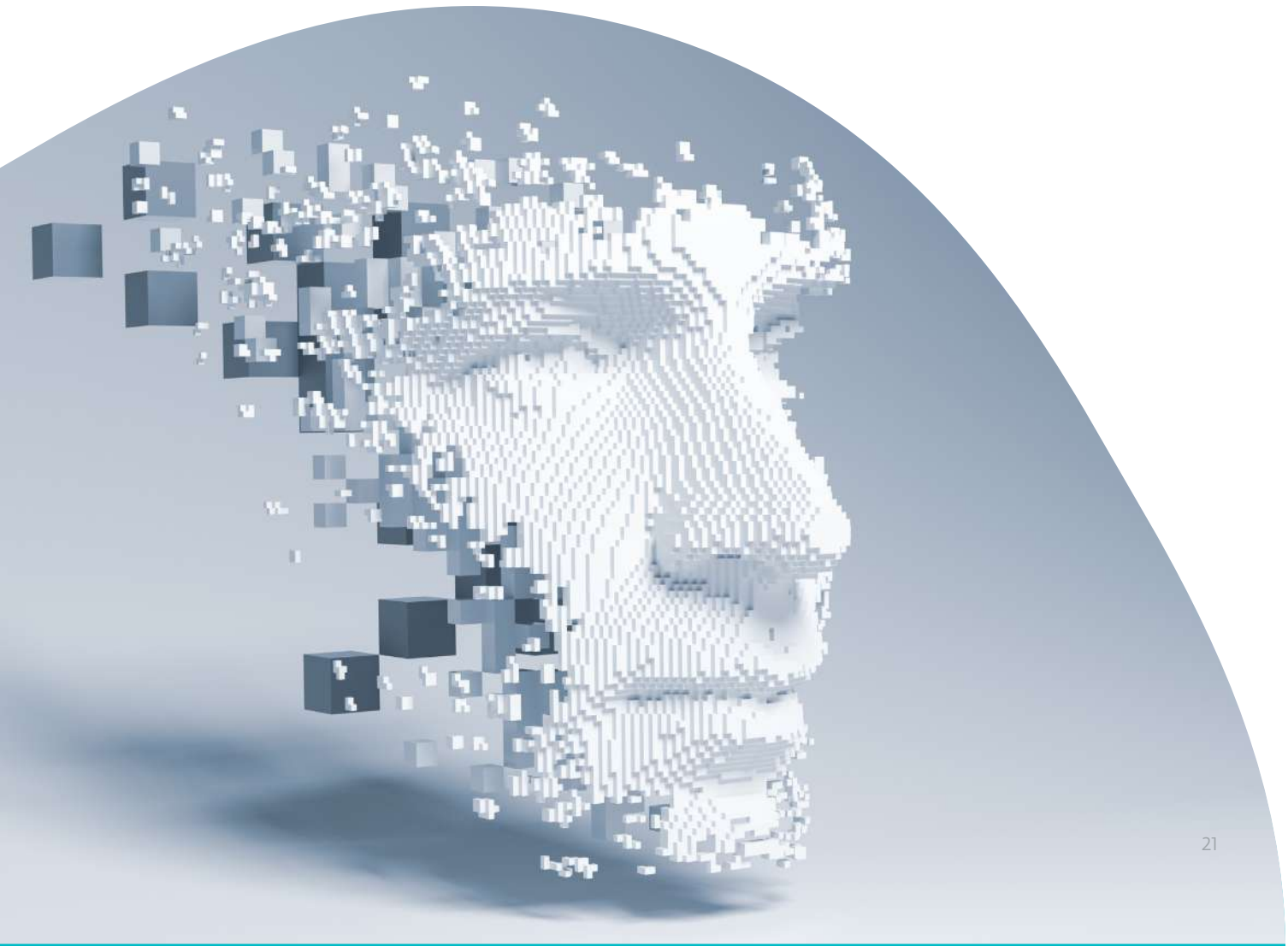
Il en va de même pour les malwares, dont la complexité va encore plus loin, notamment en ce qui concerne la diffusion du « produit » final, sa protection contre la détection et l'analyse, ainsi que d'autres étapes nécessaires pour qu'il soit efficace et rentable. Il est beaucoup plus facile pour un attaquant ayant des compétences en programmation même médiocres d'utiliser des tutoriels ou de travailler à partir d'un code source fuité au lieu d'utiliser un modèle génératif.

AFFIRMATION : PLUS LE MODÈLE EST GRAND, MIEUX C'EST

La réalité : L'une des principales caractéristiques des LLM est leur taille. Certains fournisseurs de solutions de cybersécurité aiment à présenter cette caractéristique comme l'un des principaux avantages pour l'analyse des malwares. Mais plus le modèle est grand, plus le coût de sa formation est élevé. Les grands modèles nécessitent du matériel coûteux, un ensemble plus étendu de données et un temps de formation rallongé. Ils consomment beaucoup d'électricité et d'autres ressources, ce qui les rend également moins écologiques. Un modèle d'IA de petite taille avec une mission restreinte est moins coûteux à former, plus abordable à maintenir, et plus facile à comprendre et contrôler. Dans le domaine de la cybersécurité, ce type de modèle peut être utilisé pour traiter de vastes quantités de données et fournir des résultats simples et facilement compréhensibles qui qualifient les échantillons d'inoffensifs ou de malveillants.

AFFIRMATION : L'IA EST LA SEULE COUCHE DE SÉCURITÉ NÉCESSAIRE

La réalité : Comme c'est le cas pour toute autre technologie, l'IA a été exagérément vendue par certaines entreprises comme la solution miracle pour tout. Il en va de même pour la cybersécurité, où des technologies de détection fiables et éprouvées par des années d'utilisation sont écartées par certains au profit de l'IA. Bien que les réseaux neuronaux, le Deep Learning et l'IA générative apportent de la valeur ajoutée, il n'existe pas d'algorithme magique qui puisse à lui seul identifier toutes les menaces possibles qui pourraient émerger. Leur combinaison avec d'autres couches de sécurité, comme c'est le cas avec ESET LiveSense, a de bien meilleures chances de détecter les comportements malveillants et de les stopper avant qu'ils ne causent des dommages.



Nous sommes ESET

Défense proactive. Notre activité consiste à minimiser la surface d'attaque.

Prenez une longueur d'avance sur les cybermenaces connues et émergentes grâce à notre **approche préventive reposant sur l'IA et l'expertise humaine**.

Bénéficiez d'une protection de haut niveau grâce à notre **Threat Intelligence** interne, compilée et examinée depuis plus de 30 ans, qui alimente notre vaste réseau de R&D dirigée par des **chercheurs reconnus**. ESET protège votre entreprise afin qu'elle puisse exploiter tout le potentiel de la technologie.

Digital Security
Progress. Protected.