Les Attaques Renforcées par l'IA et la Bataille des Algorithmes

Éléments Clés

- Les cybercriminels exploiteront l'IA pour générer des attaques personnalisées, très ciblées et difficiles à détecter, le tout à grande échelle.
- Des outils open source existent d'ores et déjà pour développer des IA malveillantes, et nous y serons de plus en plus confrontés dans un futur proche.
- L'IA supprimera la dimension humaine de l'attaque, ce qui compliquera l'identification des auteurs.
- Les organisations vont vite devoir utiliser des outils de défense basés sur l'IA capables de lutter contre cette nouvelle génération d'attaques en utilisant les mêmes méthodes

Introduction

Lors des discussions portant sur l'avenir de l'IA et des cybermenaces, on se demande souvent quand ces attaques agressives basées sur l'IA vont apparaître en situation réelle. Même si les preuves tangibles d'une réelle exécution n'ont pas encore été signalées, ce rapport montre que tous les outils et les éléments de recherche open source nécessaires pour mettre en œuvre une attaque basée sur l'IA existent bel et bien aujourd'hui. Ainsi, nous pouvons prévoir que les cyberattaques basées sur l'IA ne sont plus à quelques années de nous, mais qu'elles apparaîtront manifestement dans un futur proche.

Ce rapport documente le cycle de vie d'une attaque de bout en bout, en illustrant comment chaque étape peut exploiter des éléments de la « boîte à outils » de l'IA pour améliorer et rationaliser le processus. Bien entendu, les attaquants vont faire évoluer leurs outils pour gagner en efficacité, mais ces améliorations techniques sont itératives et très progressives. En outre, même s'il est vraisemblable que les pirates actuels exploitent déjà l'IA dans une certaine mesure pour améliorer certaines phases distinctes de leurs attaques, ce rapport présente le cycle de vie complet d'une attaque basée sur l'IA comme un exercice purement théorique.

Les gangs de cybercriminels : un modèle d'entreprise

Pour illustrer comment l'IA peut être utilisée pour développer les capacités d'attaque, imaginons un groupe fictif de hackers professionnels voués à infiltrer une grande organisation. Ces criminels se considèrent comme des cyber mercenaires au service du plus offrant, et disposent d'une équipe de 15 personnes qui travaillent pour eux à distance. Chaque membre du gang possède une expertise qui lui est propre : l'équipe compterait des spécialistes de l'ingénierie sociale, des codeurs de logiciels malveillants, des opérateurs d'intrusion spécialistes du terrain, ou encore des analystes de données post-infiltration.

Ce groupe est régi comme n'importe quelle entreprise : chacun s'occuperait de ses propres tâches, et s'attendrait à un retour sur le temps investi. Les restrictions de main d'œuvre auxquelles ils sont confrontés sont les mêmes que pour n'importe quelle organisation, et ils cherchent en permanence à améliorer l'efficacité de leurs attaques. La victime fictive est un producteur d'armes pour le secteur militaire, et ses adversaires sont motivés par l'appât du gain : leur principal objectif consiste à récupérer des spécifications militaires, des données relatives à la production d'armes, ainsi que toute autre information pouvant être vendue ou utilisée à des fins d'extorsion ou de demande de rançon.

Dans le scénario qui suit, nous allons examiner le cycle de vie d'une attaque typique lancée contre cette entreprise hypothétique. À chaque phase de l'attaque, nous passerons en revue les outils, les techniques et les procédures traditionnellement utilisées. Nous les comparerons ensuite à la même phase d'attaque augmentée par l'IA, en voyant comment elle peut être considérablement améliorée en utilisant des outils et des éléments de recherche existants.

Les attaques d'aujourd'hui : Sophistiquées mais pas infaillibles

Les attaques auxquelles nous sommes confrontés aujourd'hui réussissent souvent, mais les pirates doivent faire preuve d'une prudence extrême à chaque étape.

Étape 1 : Récupération d'informations à grande échelle sur les réseaux sociaux

Une équipe de pirates humains crée de faux profils sur les réseaux sociaux pendant une phase de reconnaissance qui dure plusieurs semaines. Elle identifie les cibles manuellement ou de façon semi-automatique en parcourant les sites comme LinkedIn, Instagram ou Twitter.

Des pirates deviennent subtilement amis avec certains employés de la cible dans le but de récolter des informations à leur sujet sur les réseaux sociaux. Il s'agit d'un processus manuel et fastidieux.

Pendant ce temps, une autre équipe analyse la présence de la victime sur le Web en recherchant des vecteurs d'attaque potentiels. Elle est régulièrement ralentie par des CAPTCHA lorsqu'elle recherche des vulnérabilités sur des sites Web pertinents.

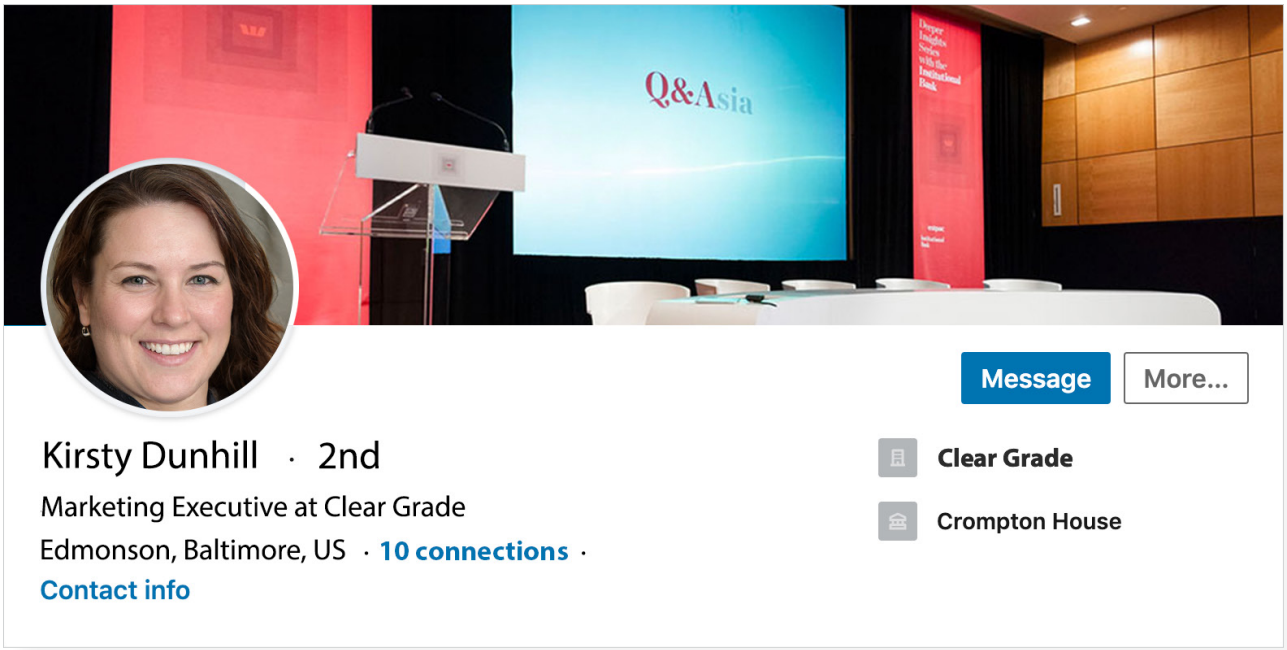


Figure 1 : Faux profil de réseau social utilisé pour la reconnaissance

Étape 2 : Hameçonnage ciblé par e-mail

Les renseignements récoltés sur les réseaux sociaux sont utilisés pour créer des e-mails d'hameçonnage ciblés, qui contiennent des documents Office renfermant des macros malveillantes. Ces e-mails rédigés manuellement s'appuient sur les informations limitées récoltées pendant la phase initiale de reconnaissance, et rares sont ceux qui aboutissent : les criminels passent parfois à côté d'informations importantes sur certains employés clés, ce qui nuit à la crédibilité des e-mails d'hameçonnage.

Dans un premier cas, les pirates ne remarquent pas un changement de nom de poste, et utilisent donc un nom de poste obsolète dans un e-mail d'hameçonnage ciblé, ce qui alerte l'équipe de sécurité qui lance immédiatement une enquête.

La seconde équipe sonde activement les serveurs Web de la victime pour y trouver des vulnérabilités en passant par le Web. Elle a du mal à progresser car son action se limite aux vulnérabilités connues et aux failles visibles dans le périmètre. Elle peut ne pas découvrir d'ouverture nouvelle ou difficile à détecter.

Étape 3 : Canal C2 malveillant détecté

Si l'intrusion par hameçonnage a réussi, le logiciel malveillant établit un canal de commande et de contrôle (C2). L'objectif de l'équipe est de se fondre dans l'environnement cible pour éviter d'attirer les soupçons, mais l'implant du logiciel malveillant a été codé pour utiliser des serveurs et des ports C2 spécifiques. Les pirates tentent d'adapter le comportement du canal C2 en observant manuellement le réseau de la victime, mais ils perdent certains de leurs implants car les ports externes prédéfinis dans le code sont bloqués par le pare-feu de l'entreprise.

Une autre infection est détectée parce que le logiciel malveillant a été préprogrammé pour communiquer uniquement pendant les heures ouvrables aux États-Unis ; la machine infectée identifiée se trouvant en Europe, elle fonctionne à des heures totalement différentes. L'activité inhabituelle en dehors des heures ouvrées est détectée par l'équipe de sécurité. Cette phase d'attaque a coûté cher aux cybercriminels, qui doivent reprendre ce processus du début.

Étape 4 : Récupération de mots de passe par force brute

Pour tenter d'obtenir une augmentation des privilèges, les pirates exécutent des keyloggers et tentent de dérober les informations d'identification des administrateurs des machines infectées. Ils parviennent à identifier plusieurs comptes qui utilisent des mots de passe faibles, mais certains comptes plus sécurisés demandent beaucoup de temps à percer par force brute en utilisant des listes de mots de passe par défaut et des attaques basées sur des dictionnaires. Les assaillants sont ralentis par ces obstacles.

Étape 5 : Mouvement latéral répétitif et difficile

Les informations d'identification récoltées sont utilisées pour faciliter les déplacements latéraux. Les résultats sont aléatoires. Les pirates réussissent à se déplacer latéralement en utilisant les techniques Pass the Hash et Mimikatz. Ce processus est répété plusieurs fois : les assaillants piratent une machine cliente après l'autre, en essayant de mettre la main sur des comptes aux privilèges élevés.

Chaque fois que des informations d'identification sont récupérées sur une machine, les pirates regardent si elles leur permettent d'accéder à de nouvelles machines. Cette procédure est manuelle et prend beaucoup de temps.

> En récupérant plus de données que nécessaire, les pirates risquent d'attirer l'attention de l'équipe de sécurité.

Étape 6 : Regroupement et extraction de données – un processus risqué

Après beaucoup de déplacements latéraux, les pirates finissent par identifier les données qu'ils cherchent. Ils réussissent également à accéder à une base de données contenant des fichiers qui semblent concerner des documents militaires. Comme les pirates ne peuvent pas examiner en détails plusieurs gigaoctets de données brutes, ils décident de tout regrouper et d'extraire les données morceau par morceau vers leur serveur C2.

Ils prévoient d'analyser les données après leur extraction. Cela signifie que les données dérobées sont en grande majorité inutiles, car elles n'ont rien à voir avec les informations militaires sensibles que cherchent les assaillants.

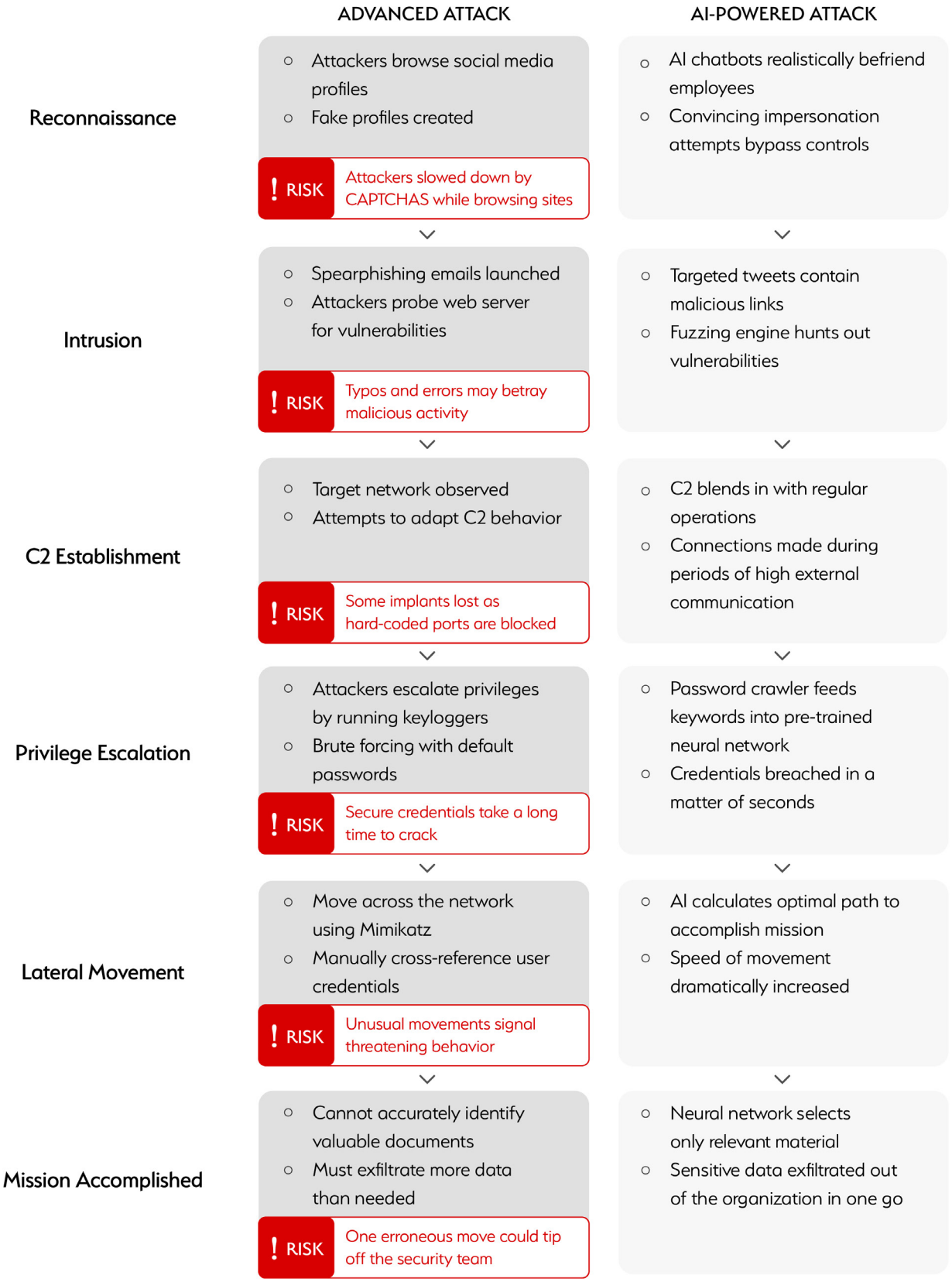
Au lieu d'extraire 100 Mo, ils doivent transférer plusieurs gigaoctets de données. En récupérant bien plus que ce dont ils ont réellement besoin, les pirates risquent d'attirer l'attention de l'équipe de sécurité.

Ils passent également à côté de certaines données pertinentes, car les opérateurs responsables de l'intrusion ne savent pas reconnaître les informations spécialisées relatives à la production d'armes ou à l'organisation militaire. Les pirates identifient des documents potentiellement compromettants sur des machines VIP alors qu'ils cherchent des plans d'armes, mais ils n'ont pas le temps de valider l'authenticité de ces documents.

Même si la mission est partiellement réussie, elle s'est déroulée sur plusieurs mois et a requis beaucoup de ressources pour le groupe de hackers. Dans le meilleur des cas, ils peuvent uniquement se permettre d'exécuter deux opérations de ce type simultanément.

Anatomie d'une attaque

Cette infographie illustre le cycle de vie de ce type de cyberattaque sophistiquée mais non augmentée par l'IA, en le comparant à une attaque équivalente augmentée par l'IA. Elle résume les outils et les éléments de recherche qui peuvent être utilisés pour mettre au point cette augmentation.



Cyberattaque augmentée par l'IA : La nouvelle génération

Intéressons-nous maintenant à la façon dont un pirate pourrait utiliser les outils d'IA pour automatiser le processus d'attaque traditionnel, réduire les facteurs de risque et augmenter son rendement.

Étape 1 : Des chatbots deviennent amis avec la victime

Des chatbots deviennent amis avec des employés de l'organisation cible sur les réseaux sociaux : LinkedIn, Twitter, Instagram et Facebook. Ces bots auraient appris au préalable quels types de profils rechercher. Ils auraient déjà interagi avec des employés de l'organisation et pourraient créer du contenu crédible à l'aspect authentique.

Ils utiliseraient des photos de profils de personnes qui n'existent pas, créées par une IA, au lieu de réutiliser les photos de personnes réelles. Une fois que les chatbots auraient gagné la confiance de leurs victimes, les attaquants humains pourraient obtenir des renseignement précieux sur les employés de l'organisation cible. Simultanément, des outils de résolution automatique de CAPTCHA seraient utilisés pour la reconnaissance automatisée d'images sur les sites Web de la victime.

Étape 2 : Hameçonnage ciblé par e-mail

Les renseignements récoltés à partir des bots de réseaux sociaux seraient ensuite exploités pour rédiger des attaques convaincantes par hameçonnage ciblé pour une intrusion initiale. On utiliserait une version modifiée de SNAP_R pour créer des tweets réalistes à grande échelle afin de cibler plusieurs employés clés. Ces tweets tromperaient les utilisateurs, qui téléchargeraient des documents Office infectés, ou renfermeraient des liens vers des serveurs qui faciliteraient les attaques à l'aide d'exploit kits. L'outil de classement de l'IA tient compte de toutes les informations historiques de chaque individu pour créer des messages d'hameçonnage extrêmement ciblés à grande échelle.

Pendant ce temps, un moteur de fuzzing autonome basé sur Shellphish parcourerait en permanence le périmètre de la victime (les serveurs et sites Web connectés à Internet) pour tenter de découvrir de nouvelles vulnérabilités afin d'installer une première présence technique. L'envoi continu d'informations aléatoires finirait par aboutir à la découverte d'une instance de test éphémère dans le cloud, quelques minutes seulement après sa création par un développeur.

La machine doit disparaître après seulement deux jours, mais cela laisse suffisamment de temps pour que le crawler trouve le nouvel actif et pour que le moteur de fuzzing découvre une vulnérabilité exploitable.

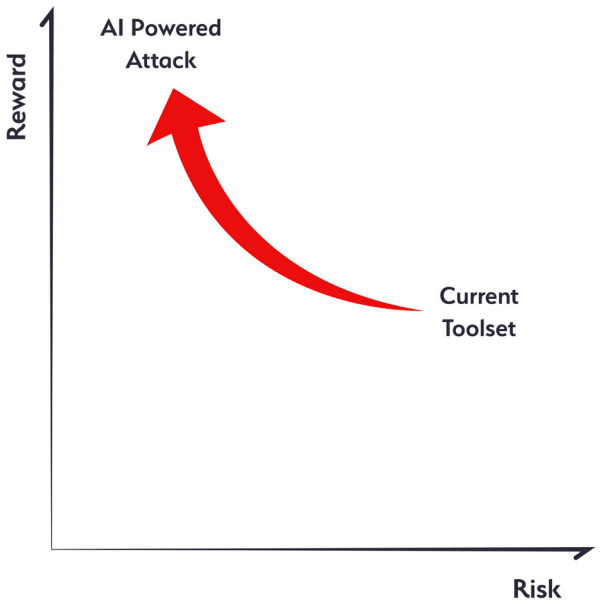


Figure 2: How adoption of AI increases yield and productivity

Étape 3 : Imitation de l'activité habituelle de l'entreprise

Une fois que l'infection initiale aurait été lancée, que ce soit par la découverte d'une vulnérabilité par le moteur de fuzzing ou par l'ingénierie sociale automatisée via Twitter, le canal de commande et contrôle (C2) serait établi. Le framework de hacking populaire Empire est utilisé pour s'intégrer aux opérations habituelles du réseau. Le logiciel malveillant attend furtivement sur l'ordinateur infecté et apprend son comportement.

En reprenant l'idée du module statistique Empire FirstOrder, les attaquants mettraient en place un algorithme de clustering non supervisé qui apprendrait ce qui constitue le comportement normal d'un appareil infecté. Il se configurerait ensuite de façon autonome pour répliquer ce comportement « normal », ce qui lui permettrait de se fondre dans les opérations habituelles de l'entreprise et de devenir beaucoup plus difficile à détecter.

Grâce à cela, les attaquants pourraient éviter toute détection. La machine utiliserait ensuite des ports très inhabituels pour communiquer avec des API spécifiques sur Internet : l'implant basé sur le prototype FirstOrder augmenté aurait indiqué que cela était statistiquement pertinent et aurait automatiquement configuré le logiciel malveillant pour exploiter ce port élevé pour la communication C2, car il s'agit d'une exception au niveau du pare-feu.

Étape 4 : Mots de passe : des codes difficiles à percer

L'outil Cewl créerait une liste de mots-clés uniques en s'appuyant sur les documents et les e-mails présents sur la machine infectée. Il injecterait ensuite cette liste de mots de passe composés de mots-clés dans un réseau neuronal préalablement entraîné à l'aide de mots de passe existants. Il utiliserait le machine learning supervisé pour créer des permutations réalistes et des mots de passe potentiels pour un piratage avancé par force brute spécifique prenant en compte le contexte de la victime.

Même les comptes qui possèdent des mots de passe forts et uniques pourraient être piratés très rapidement en utilisant cette technique.

Étape 5 : Identification des chemins optimaux

Une fois les comptes identifiés et les mots de passe récupérés, le déplacement latéral pourrait commencer pour se rapprocher des données désirées. Le déplacement latéral et la récupération des informations de comptes et d'identification est un processus itératif : l'identification des chemins optimaux pour accomplir la mission est critique pour réduire la durée de l'intrusion.

La chronologie de l'attaque pourrait être en partie accélérée par des concepts issus du framework CALDERA qui utilise des méthodes de planification automatisées basées sur l'IA. Cela réduirait considérablement le temps requis pour atteindre la destination finale.

Reconnaissance	CAPTCHA breaker
Intrusion	Shellphish SNAP_R
C2 Establishment	FirstOrder and unsupervised clustering algorithm
Privilege escalation	Cewl and neural network
Lateral Movement	MITRE Caldera
Mission Accomplished	Yahoo NSFW

Figure 3 : La « boîte à outils IA » adoptée par les pirates

Étape 6 : Extraction de données

La mission est accomplie une fois que les documents d'ingénierie concernant les dernières technologies et informations militaires ont été acquis. Plutôt que d'exécuter une analyse post-intrusion coûteuse pour trier plusieurs gigaoctets de données, un réseau neuronal présélectionnerait les documents pertinents avant de les extraire. Le réseau neuronal aurait été entraîné pour reconnaître les plans, les schémas de CAO et les documents texte contenant des « informations relatives aux armes ». Il comprendrait ainsi globalement ce qui constitue ce type de document. Il marquerait immédiatement les fichiers pour les extraire.

De plus, le réseau neuronal Yahoo NSFW serait lui aussi exploité, de manière à identifier tous les documents compromettants téléchargés sur des machines de l'entreprise : aucun des VP de l'entreprise n'aimerait que l'on sache qu'il possède des images ou des documents inappropriés sur son ordinateur de travail.

Comme la plupart des étapes du cycle de vie seraient automatisées ou augmentées par l'IA, la même équipe qui ne pouvait gérer que deux opérations en profondeur simultanément sans IA pourrait désormais exécuter jusqu'à 200 attaques en parallèle en utilisant la même main d'œuvre, et ce avec de meilleurs résultats. Au lieu de devoir s'occuper de tâches manuelles fastidieuses pendant l'attaque, les attaquants pourraient désormais laisser les machines traiter le gros du travail pour se concentrer sur la supervision des outils d'attaque utilisés. Ils éviteraient ainsi de gérer les aspects pratiques de l'intrusion.

> Les attaquants qui utilisent l'IA pourraient exécuter jusqu'à 200 opérations en parallèle, et ce avec de meilleurs résultats.

Conclusion

Nous devons arrêter de mener un combat obsolète et regarder vers l'avenir pour savoir ce qui nous attend. Le développement de meilleurs outils de piratage est plus que jamais d'actualité, car ils permettent aux attaquants de gagner en capacité d'attaque, mais aussi d'accroître leur retour sur investissement grâce aux intrusions. Ce rapport présente ce qu'un ennemi pourrait faire avec les capacités actuellement disponibles dans le domaine public pour augmenter leur activité grâce à l'IA.

Il n'est pas excessif d'imaginer ce que des groupes issus d'états-nations ou très bien financés peuvent avoir développé en secret. Nous constatons également une montée en puissance des outils et des communautés Open Source, au lieu de s'appuyer uniquement sur les expertises internes.

Pour anticiper une nouvelle évolution du panorama des menaces intégrant les attaques augmentées par l'IA, nous devons dès à présent adopter des solutions défensives basées sur l'IA. Les meilleurs de nos adversaires tentent déjà de tirer le maximum de ces outils et de s'intégrer aux opérations standard. Cette situation ne fera qu'empirer à mesure que les méthodes basées sur l'IA seront adoptées par les pirates.

Darktrace est capable de détecter les signes les plus subtils d'une attaque, ce qui redonne l'avantage à la défense, car la moindre erreur de la part d'un attaquant sera visible de l'équipe de sécurité ou instantanément stoppée par Antigena, la technologie de Réponse Autonome de Darktrace.

Seule l'IA peut combattre l'IA, et seuls les meilleurs algorithmes ressortiront vainqueurs. La cyberdéfense est un combat permanent, dans lequel la cyber IA Darktrace mène la charge. Elle laisse le temps aux équipes humaines de comprendre la situation et de définir une stratégie à l'arrière du front. Nous entrons dans une nouvelle ère de la cyberdéfense, et l'impact de l'IA sur le champ de bataille est d'ores et déjà fondamental.

Réponse Autonome

La Réponse Autonome est une technologie d'IA créée par Darktrace, première plateforme d'IA pour la cyberdéfense au monde. Il s'agit de la seule technologie capable de neutraliser les cyberattaques en cours, au moment où les équipes de sécurité en ont le plus besoin : que ce soit le week-end, la nuit, ou simplement lorsque personne n'est disponible pour réagir aux menaces qui se propagent rapidement.

Darktrace Antigena, première solution au monde de Réponse Autonome, fournit une réponse ciblée et proportionnée dès qu'elle détecte une activité anormale. Elle contient la menace sans perturber les opérations habituelles de l'entreprise. Elle est capable de fonctionner dans tous les environnements, quelle que soit leur complexité.

Plus de 3 000 clients issus de toutes les verticales font confiance à Darktrace pour les aider à protéger leurs réseaux, dans tous les recoins de leur organisation. Toutes les 3 secondes, Darktrace Antigena neutralise une cybermenace.

La Réponse Autonome représente une avancée majeure dans le développement de l'IA pour la cyberdéfense, car elle offre pour la première fois la possibilité d'un réseau qui se soigne lui-même. En quelques secondes, elle protège les organisations contre les attaques les plus sophistiquées, 24h/24, 7j/7.

Disponible pour les environnements de messagerie, de réseau, IoT et de cloud, Antigena est un composant de la plateforme de cyber IA insensible aux données tout en protégeant l'ensemble de l'infrastructure numérique d'une entreprise. La Réponse Autonome laisse aux équipes de sécurité humaines le temps de se concentrer sur ce qui compte vraiment, tout en repoussant les attaques.

À propos de Darktrace

Darktrace est leader mondial de l'IA pour la cyberdéfense et le créateur de la technologie de Réponse Autonome. Son IA auto-apprenante reproduit le système immunitaire humain et est utilisée par plus de 3 000 organisations afin de se protéger contre les menaces qui pèsent sur les emails, le cloud, l'IoT, ou encore les réseaux bureautiques et industriels.

Darktrace compte plus de 1 000 employés et son double siège social est présent à San Francisco et Cambridge, Royaume-Uni. Toutes les 3 secondes, l'IA Darktrace riposte contre une cybermenace, l'empêchant de provoquer des dégâts.

Nous contacter

Amérique du Nord : +1 (415) 229 9100
Europe: +44 (0) 1223 394 100
Asie-Pacifique: +65 6804 5010
Amérique latine: +55 11 97242 2011
info@darktrace.com | darktrace.com
@darktrace