L'intelligence artificielle au sein de l'espace cybernétique

Colonel Patrick Perrot, PhD
Coordonnateur pour l'intelligence artificielle
Chargé de la stratégie de la donnée
Service de la Transformation
Gendarmerie nationale

Il n'est plus un domaine qui échappe à l'intelligence artificielle tant elle est omniprésente au sein de notre monde, qu'il soit physique ou cybernétique. En matière de cybercriminalité, la question est de savoir si l'intelligence artificielle constitue le mal ou l'antidote. Comme bien souvent, c'est un peu des deux mais peut-être plus encore dans ce milieu spécifique où la célérité des actions comme des réactions nécessite une forte capacité d'anticipation. Il est difficile de définir précisément la cybercriminalité, nous prendrons donc en considération la proposition de l'Organisation des Nations Unies qui la définit comme étant « tous faits illégaux commis au moyen d'un système, d'un réseau informatique ou en relation avec un système informatique ».

Par son imprégnation dans nos vies quotidiennes (smartphone, chatbot, voiture autonome, médecine prédictive, systèmes de recommandations, réseaux sociaux et influence politique), l'intelligence artificielle représente pour les organisations criminelles comme pour l'individu, un moyen d'acquisition de profits considérables. Elle est une discipline qui démultiplie les opportunités d'attaques et accroît les capacités de nuisance. Et il est un terrain particulièrement fertile au développement de la cybercriminalité qui est encore en devenir : les territoires connectés. Demain, la connexion apportera une solution à la circulation de l'information, à la gestion de l'énergie, à la régulation des flux et de la mobilité, à la gestion des déchets comme à la protection des biens et des services, sans compter le développement des objets connectés individuels qui ne manqueront pas d'équiper nos poignets ou nos vêtements. L'exploitation de toutes ces données reposera sur les méthodes d'intelligence artificielle afin d'améliorer les processus par un apprentissage continu.

L'intelligence artificielle, de nouvelles opportunités pour les cyberdélinquants

L'intelligence artificielle qui est à la source de plus en plus d'applications offre de belles opportunités aux délinquants pour accroître la quantité comme l'efficacité des attaques cyber par des actions répétitives auto-apprenantes. Les formes d'attaques peuvent revêtir différentes formes à partir de la phase d'acquisition des données, de la phase de traitement des données ou encore de la sortie du système.
Parmi les méthodes les plus communes, la technique du « data Poisoning », ou empoisonnement des données en français,

consiste à polluer les jeux de données et donc à totalement perturber les systèmes automatiques. Elle s'attaque aux données utilisées pour entraîner les modèles d'apprentissage et permet ainsi, soit de fausser totalement les résultats, soit de contrôler le comportement prédictif du modèle entraîné. Il est alors possible de totalement corrompre le modèle et les capacités de classification, de détection ou de prédiction. L'objectif est de détourner le centre de gravité de l'application en modifiant les jeux de données d'entrée.

Les hypertrucages ou « deepfakes » permettent de constituer de faux corpus à l'imitation des véritables données. Dès lors, il est possible, par des jeux de données artificielles, de fausser la distribution des données réelles en déséquilibrant les catégories sur lesquelles se fondent les résultats. Mais les hypertrucages peuvent aussi être exploités pour réaliser des impostures dans le monde réel. C'est le cas par exemple pour l'infraction connue sous le nom d'« arnaque au Président ». Cette infraction reposait sur l'étude de l'ingénierie sociale d'une entreprise pour ensuite tenter de se faire passer pour le Président en appelant un secrétariat ou service comptable de l'entreprise et solliciter un virement conséquent. Désormais par les hypertrucages, l'imposture s'améliore considérablement parce qu'il est possible d'imiter la voix du Président, voire de diffuser un message à partir d'une vidéo truquée et ainsi de rendre l'arnaque encore plus crédible.

Les systèmes d'intelligence artificielle peuvent également être perturbés durant la phase de traitement par des méthodes d'inférence, c'est-à-dire des méthodes ayant pour objectif de comprendre le fonctionnement des systèmes par des requêtes successives. Les systèmes sont souvent considérés comme des boîtes noires mais en leur adressant un nombre illimité de requêtes et en analysant la réponse à chacune de ces requêtes, progressivement, la boîte noire révèle ses secrets à l'attaquant. Il s'agit en quelque sorte de méthodes de « reverse engineering » où l'étude du couple entrée-sortie du système permet de reconstruire le mécanisme interne du système. À titre d'illustration, de plus en plus d'études s'intéressent à l'estimation des poids des réseaux de neurones profonds à partir de ces méthodes. Ces approches sont exploitées par les hackers qui utilisent l'intelligence artificielle en détournant les algorithmes afin de les inciter à prendre de mauvaises décisions. Une attaque de ce type sur un système de trading de cryptomonnaies a été conduite de la sorte. Les criminels ont tenté de comprendre comment les robots effectuaient le trading, puis les ont utilisés pour tromper l'algorithme.

Autre technique utilisée pour altérer totalement la sortie des systèmes est celle des attaques adverses. Cela consiste à optimiser un bruit de façon à générer une perturbation indétectable mais qui altère totalement la réponse d'un système. C'est ainsi que, dans le domaine de la reconnaissance d'images, il est possible, en modifiant quelques pixels, de confondre un chien avec une voiture, un avion avec un gorille. En matière de développement des véhicules autonomes, les attaques adverses constituent une véritable menace considérant qu'un panneau « STOP » peut ne pas être reconnu par la simple apposition d'un autocollant bien placé de faible taille.

Les Botnets, qui sont des robots en réseaux permettant d'infecter les systèmes, voient aussi leurs capacités démultipliées par l'intelligence artificielle. L'objectif est encore d'enrichir les corpus et d'extraire de l'information des systèmes pour mieux en comprendre les failles et vulnérabilités. L'intelligence artificielle peut ainsi sélectionner ou adapter les logiciels malveillants

et contrer activement les efforts de sécurité pour les rendre inefficaces.
Ainsi les méthodes d'exploitation des failles de l'intelligence artificielle tout au long du cycle de vie de la donnée sont légion et ne manquent pas d'ingéniosité. L'action malveillante a pour objectif d'analyser le comportement des systèmes, de récupérer des données ou de s'emparer du modèle d'apprentissage et dès lors de s'approprier la capacité de décision.

L'intelligence artificielle offre donc de belles perspectives aux cyberdélinquants pour détourner ou s'approprier à distance le fonctionnement des systèmes. Fort heureusement, pour protéger les systèmes contre les failles de l'intelligence artificielle, il est une arme capable de réagir avec efficacité et célérité : l'intelligence artificielle.

L'intelligence artificielle, une arme efficace contre les cyberdélinquants

En effet, face aux nombreuses possibilités d'attaques des cyberdélinquants, la seule défense efficace qui permettra de réagir de façon ciblée et dans le temps de l'attaque sera l'intelligence artificielle. Celle-ci permet de détecter les failles comme les menaces, d'adapter les dispositifs pour réagir et de le faire dans un temps suffisamment rapide pour minimiser l'impact des attaques. Grâce à sa capacité d'apprentissage, l'intelligence artificielle s'améliore en continu et est en mesure de proposer une défense ciblée, adaptative et évolutive contre l'action des cybercriminels. Comme pour l'attaque, la défense peut se concentrer sur trois volets : la maîtrise des données entrantes, la fiabilité du traitement et le contrôle des résultats.

L'intelligence artificielle permet de détecter les tentatives d'intrusion et les comportements atypiques et anticiper d'éventuelles attaques. Il est ainsi possible de mesurer le nombre de requêtes de même nature qui pourrait apparaître comme caractéristique d'un signal faible d'attaques de type inférence. Cette action de prévention d'actions suspectes est connue dans la littérature sous l'acronyme UEBA (User and Entity Behavior Analytics). L'UEBA consiste à surveiller les comportements des utilisateurs mais aussi des entités que sont les applications mobiles, en nuage ou en réseaux comme les serveurs. Par apprentissage, chaque entité ou utilisateur se voit modéliser et toute nouvelle action est comparée à ces modèles pour détecter les situations atypiques ou non conformes au modèle.

Pour contrer les attaques ayant pour objet la prise en compte des capacités comme des données, la phase d'apprentissage des modèles doit faire l'objet d'une attention particulière. Il convient en prévention d'allonger la durée puis d'élargir le jeu de données de façon à diminuer la surface d'exposition aux modifications partielles des jeux d'apprentissage et donc d'accroître la robustesse. Une surveillance doit ensuite être mise en œuvre : il s'agit alors d'associer l'expertise métier (cohérence des données), le contrôle de l'intégrité des données (ingénierie des données) et de détecter les dérives de la modélisation (science des données) au fur et à mesure de l'entraînement.

Pour prévenir les attaques qui consistent à tester les systèmes en vue de connaître leur mode de fonctionnement, il est possible de multiplier les modèles de prédictions en fondant l'analyse sur des méthodes différentes. L'objectif est d'insérer des aléas afin de rendre inexploitables les failles de confidentialité.

Cela complique la tâche de l'attaquant soit en multipliant les solutions possibles, soit en proposant des modèles différents. L'inconvénient est que l'injection des aléas peut également altérer le niveau de performance du système, il

s'agira alors d'évaluer le compromis entre la performance et la protection.

Face aux hypertrucages, l'intelligence artificielle constitue également un remède efficace. Il est, par exemple, possible d'insérer durant l'apprentissage des perturbations adverses. Ainsi contrariés, les réseaux génératifs adverses à l'origine des hypertrucages, ne parviendront pas à correctement construire leur modèle d'imposture. La détection des impostures de type deepfakes est assimilable à une course à l'armement, à un conflit continu attaque-défense. Cette détection, pour être efficace, doit entraîner un modèle d'apprentissage à partir d'un maximum de méthodes afin de disposer de la plus grande capacité généralisatrice possible. À défaut, la détection sera trop spécialisée et perdra en performance. Or, dans cette discipline, la capacité de généralisation est le point d'orgue de l'adaptation des défenses aux attaques.

Ainsi que ce soit dans le monde physique ou virtuel, l'intelligence artificielle se révèle en double face. Elle constitue à la fois le moyen de réaliser et d'optimiser des attaques contre les systèmes d'information mais aussi de proposer les solutions pour s'en protéger. Il est indispensable pour les forces de sécurité intérieure de s'investir en partenariat avec les organismes de recherche dans la sécurité des modèles d'intelligence artificielle, au risque de voir cette discipline perdre en confiance chez le citoyen, usager des systèmes, et gagner en opportunité chez le délinquant.