Intelligence Artificielle et Cybersécurité : solution ou menace ?

Besma ZEDDINI,

Enseignante-chercheur en intelligence artificielle et cybersécurité, CY Tech, CY Cergy Paris Université,
Responsable de la filière Cybersécurité et du Mastère Spécialisé® Cybersécurité & Smart Systems,
Chargée de mission à l'innovation et transfert des sciences expérimentales

L'intelligence artificielle vise à construire des programmes informatiques qui s'adonnent à des tâches demandant des processus mentaux de haut niveau tels que l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critique [Minsky, 1956]. En d'autres termes, il s'agit de l'ensemble des moyens théoriques et techniques pour faire simuler à des machines des comportements généralement associés à l'intelligence humaine. De nos jours, les travaux sur l'intelligence artificielle visent plutôt à résoudre des problèmes d'une manière plus satisfaisante que l'intelligence humaine. Les travaux et applications se concentrent sur l'intelligence artificielle faible[1], focalisée sur une tâche précise (la recommandation de contenus, la reconnaissance d'images pour l'authentification ou la détection de pathologies, la traduction automatique, etc.).

Divers domaines de recherche s'inscrivent dans la démarche générale de création d'une intelligence artificielle, tels que la représentation de connaissances, le traitement automatique des langues, la robotique, la planification, la modélisation cognitive, etc. Ces domaines de recherche sont actifs depuis des décennies, mais le regain réel d'intérêt pour l'intelligence artificielle, tant académique qu'industriel, a eu lieu au début des années 2010 avec le développement vertigineux du Big data, des sciences de données en général et de l'apprentissage automatique en particulier (supervisé ou non supervisé) et surtout de l'apprentissage profond.

Les deux faces de l'Intelligence artificielle

En décembre 2018, le cabinet d'études McKinsey a recensé les usages réels de l'intelligence artificielle, et plus particulièrement l'apprentissage profond, pour le bien social [McKinsey, 2018]. Diverses actions, telles que la fondation et la plateforme ONU « AI for good » ou le projet « AI for Social Good » de Google[2] mettent en avant les projets en intelligence artificielle au service du bien commun et du progrès de tous les humains. Les avancées en IA sont d'ailleurs généralement accueillies avec enthousiasme, et les usages duaux des nouvelles technologies, leur usage détourné ou leur sensibilité aux attaques sont généralement passés sous silence.

Pourtant, l'intelligence artificielle peut avoir des applications néfastes voire criminelles, en utilisant les mêmes théories, les mêmes technologies et les mêmes avancées. Le système de lecture labiale de Google et Oxford [Chung *et al.*, 2017], permettant de lire sur les lèvres d'une manière souvent plus pertinente que des professionnels humains, peut aider les personnes ayant un trouble de la parole, sourde ou malentendante ; mais il permet également de développer un système de surveillance ou d'« écoute » plus performant par des entités malintentionnées. Un système

de génération automatique de vidéos [Greenmier, 2018] peut optimiser la production cinématographique et documentaire, mais il peut également participer à la diffusion massive de fake news et à la désinformation. Un drone autonome peut prendre des vidéos d'endroits inaccessibles, mais il peut également faciliter le lancement de projectiles sur ces mêmes cibles inaccessibles. Il est donc d'une grande importance de protéger l'intelligence artificielle contre les usages frauduleux et de dispenser des efforts comparables sur la cyber-protection de l'IA que sur l'IA elle-même.

Intelligence artificielle comme cible privilégiée des cyberattaques

Plusieurs failles de sécurité dans les intelligences artificielles commerciales sont régulièrement mises en évidences (e.g. [Zhang *et al.*, 2018]). Le chemin semble balisé pour des attaques de plus grande gravité, comme le déverrouillage d'accès et le transfert d'argent. Suivant des techniques similaires, des cybercriminels peuvent cibler l'intelligence artificielle commandant l'authentification d'une institution financière ou une entreprise peu scrupuleuse peut cibler l'IA définissant la stratégie de détermination du prix de son concurrent. En décembre 2017, la société de sécurité Webroot a effectué une enquête concluant que plus de 90% des professionnels de la cybersécurité aux Etats-Unis et au Japon s'attendent à ce que les attaques utilisent l'intelligence artificielle contre les entreprises pour lesquelles elle est supposée travailler [Webroot, 2017].

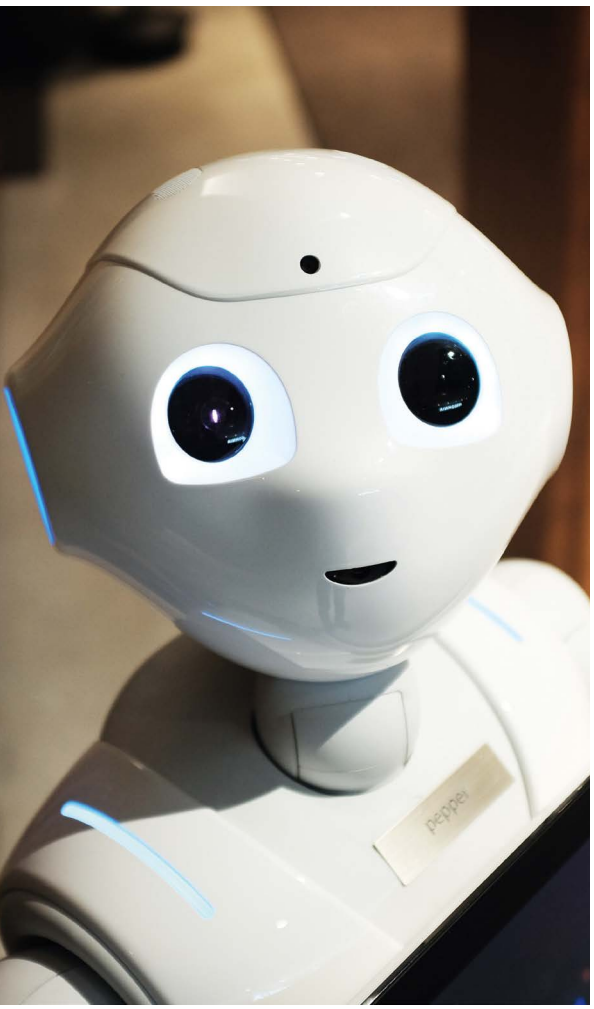
Ces menaces pèsent sur toutes les étapes de mise en place d'une intelligence artificielle. La méthodologie de l'apprentissage automatique est en effet fondée sur trois étapes : l'acquisition de données (e.g. images, vidéos, voix, transactions, etc.), l'apprentissage sur cette base de données (e.g. raisonnement par analogie, apprentissage de compétences comme la conduite, ou la prévision d'états futurs, etc.) et enfin l'action fondée sur de nouvelles données (génération d'images, de textes ou de vidéos, la conduite ou la navigation, etc.). Le système résultat est amélioré et corrigé par un processus itératif durant l'exécution. Chacune de ces étapes présente un risque de compromission.

1. Lors de l'étape d'acquisition des données d'apprentissage, ces données peuvent être corrompues ou manipulées.

2. Lors de l'étape d'apprentissage elle-même, les algorithmes d'apprentissage peuvent être détournés ou corrompus.

1/ Par opposition à l'intelligence artificielle forte, capable de construire des programmes ayant conscience d'eux-mêmes

2/ https://ai.google/social-good



3. Lors de l'étape de mise en place du système, la configuration des composants du système peut être changée et détournée de l'objectif principal.

Cette méthodologie à trois-étapes présente également deux principaux risques en termes de sécurité. Le premier est que les systèmes sont généralement conçus pour s'exécuter en boucle fermée, sans intervention humaine, dans leur travail quotidien. Les attaques vers ces systèmes peuvent ainsi rester longtemps sans être détectées. Le second risque est relatif à la grande masse de données manipulée par les algorithmes d'IA qui font que les raisons guidant une telle décision sont souvent difficilement interprétables. Cela signifie que, quand bien même une attaque serait détectée, ces motivations peuvent demeurer opaques.

Intelligence artificielle comme solution aux cyberattaques

La méthodologie, la technologie et les outils d'intelligence artificielle peuvent également être mis à contribution pour protéger les systèmes des cyberattaques. Une analyse des cyberattaques passées peut permettre de différencier les situations réellement dangereuses de celles qui le sont bien moins (une faute de frappe sur un mot de passe comparée à un usage frauduleux de carte bancaire, par exemple). Plus généralement, l'IA peut aider à renforcer la sécurité des systèmes fondés sur l'IA durant les trois principales étapes de cybersécurité : la prévention, la détection et la réponse aux attaques.

1. **La prévention** : l'apprentissage automatique peut servir à apprendre des attaques précédentes et de pouvoir mettre en place les systèmes pertinents pour chaque menace de sécurité identifiée. Ce système de prévision pourra rapidement s'adapter aux menaces précédemment inconnues.

2. La détection : les méthodes de détection fondée sur la signature des attaques (des règles statiques identifiant les attaques) sont bouleversées avec l'IA. Les algorithmes fondés sur l'IA peuvent maintenant détecter tout changement comparé à une situation définie comme normale du système. Cela offre davantage de potentiel de détection des menaces inconnues jusqu'alors. Par ailleurs, l'apprentissage par renforcement et l'apprentissage profond permettent maintenant de se dispenser des grandes bases de données d'entrainement, et peuvent être bien plus rapidement opérationnels dans ce contexte de détection d'attaques.

3. **La réponse** : l'IA peut participer grandement aux cyberattaques en priorisant le travail des analystes et en l'orientant vers les activités à haute valeur ajoutée. Elle peut également permettre la mise en quarantaine automatique de parties du système ou de ses utilisateurs pendant une attaque.

En guise de conclusion et perspectives, Les questions principales qui se posent : Faut-il se fier à l'IA ? Quid de l'explicabilité et de l'éthique de l'IA ? ■

Références

[Minsky, 1956] M. Minsky, "Heuristic Aspects of the Artificial Intelligence Problem", Lincoln Laboratory, M.I.T., Lexington, Mass. Group Report No. 34–55, 1956

[McKinsey, 2018], McKinsey, "Notes from the AI frontier, applying AI for social good", discussion paper, december 2018, 48 pages, 2018

[Chung et al. 2017] J. S. Chung, A. Senior, O. Vinyals and A. Zisserman, "Lip Reading Sentences in the Wild", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 3444-3453, 2017

[Greenmeier, 2018] L. Greenemeier, "Don't Believe Your Eyes", Scientific American 318(4):12-14, 2018

[Webroot, 2017] Webroot, "Game Changers: AI and Machine Learning in Cybersecurity," 9 pages, 2017

[Zhang et al., 2018] R. Zhang, X. Chen, J. Lu, S. Wen, S. Nepal, Y. Xiang, "Using AI to Hack IA: A *New Stealthy Spyware Against Voice Assistance Functions in Smart Phones*", CoRR abs/1805.06187, 11 pages, 2018