Intelligence artificielle et cybersécurité

L'intelligence artificielle (IA) modifiera le visage de la cybersécurité au cours des prochaines années. Elle devrait renforcer les cyberopérations offensives et défensives et contribuer à façonner le paysage des cybermenaces. Ces évolutions s'annoncent particulièrement délicates à contrôler, pour les acteurs étatique. Pour relever ces défis, ceux-ci devront adopter des cadres politiques et normatifs adéquats.

Par Matteo E. Bonfanti
et Kevin Kohler

L'apprentissage automatique permet de créer des images à partir de photos de personnes réelles. *Devinez: lequel de ces enfants existe et lequel est généré par l'IA ? (*Solution à la dernière page)*

L'IA est un terme générique créé en 1955 par l'informaticien John McCarthy, qui renvoie à «la science et l'ingénierie des machines intelligentes». Aujourd'hui, l'IA désigne un système habilitant et un domaine de recherche. En tant que telle, l'IA est une discipline scientifique qui vise à rendre des systèmes artificiels capables d'exécuter des tâches dont on estime qu'elles requièrent un certain degré de rationalité ou d'intelligence lorsqu'elles sont réalisées par des humains. Différentes approches permettent d'atteindre cet objectif. L'une d'entre elles est l'apprentissage automatique ou *machine learning*, qui repose sur trois grands piliers: les algorithmes d'apprentissage, les données et la puissance de calcul. Les récents succès de l'IA sont, pour la plupart, issus d'un sous-domaine de l'apprentissage automatique: l'apprentissage profond ou *deep learning*. Ce procédé utilise des réseaux neuronaux profonds constitués de multiples couches de neurones artificiels, chacun transformant les données qu'il reçoit. Ces réseaux neuronaux sont inspirés du cerveau humain. Avec l'accroissement de leurs capacités d'apprentissage et de leur pouvoir de décision, on peut s'attendre à ce que les systèmes artificiels deviennent de plus en plus autonomes.

L'IA est également décrite comme une technologie habilitante car elle peut être déployée dans une multitude de domaines, à des fins civiles (voir l'analyse du CSS no 260) ou militaires (voir l'analyse du CSS no 251), avec de bonnes ou de mauvaises intentions. Sans surprise, elle peut également être utilisée pour atteindre des objectifs liés à la cybersécurité. Dans ces cas-là, ces solutions technologiques consistent à intégrer des approches et des capacités d'apprentissage automatique pour traiter de gros volumes d'informations et en tirer des enseignements susceptibles de façonner un plan d'action d'intérêt cyber.

Les questions de sécurité liées à l'IA

Le milieu de la recherche en IA est globalement favorable à l'informatique libre. Les chercheurs sont convaincus que la connaissance doit être gratuite. Ils sont également soucieux d'inscrire leurs travaux dans la chronologie publique de la recherche afin d'acquérir un certain prestige professionnel

au sein d'un domaine extrêmement dynamique. Par conséquent, ils se contentent rarement de publier de vagues descriptions de leurs réalisations. Au contraire, ils diffusent librement sur Internet leurs codes sources, leurs modèles entraînés, leurs tutoriels et même leurs jeux de données. L'accès au dernier grand pilier de l'IA, la puissance de calcul, s'est également amélioré grâce aux fournisseurs de *cloud computing* (informatique en nuage) à la demande. La recherche et le développement de l'IA connaissent ainsi une diffusion rapide: on parle alors parfois de «démocratisation de l'IA».

Pour limiter les dégâts que pourrait engendrer la prolifération d'outils polyvalents toujours plus puissants aux mains d'acteurs malveillants, les chercheurs en IA étudient de nouvelles approches, telles que les licences d'IA responsables ou la mise à disposition progressive de modèles pleinement entraînés. Cette dernière s'inspire de la divulgation responsable de vulnérabilités *zero-day* dans le domaine de la cybersécurité et a été utilisée pour la première fois lors de la publication du modèle de langage GPT-2 d'OpenAI en 2019. Le côté appliqué de l'IA est moins transparent que celui de la recherche. En effet, les jeux de données commerciaux et les modèles qu'ils ont servi à entraîner sont considérés, du point de vue économique, comme des actifs incorporels dont la confidentialité doit être protégée contre le vol et l'espionnage. Dans certains pays, des jeux de données particulièrement sensibles, tels que des informations sur le génome de la population ou des algorithmes, sont soumis à des restrictions à l'exportation. Des discussions sont également en cours quant à la possibilité d'étendre ces restrictions au matériel d'IA et aux outils connexes.

L'IA n'est pas encore très solide et fait souvent des erreurs qu'un humain ne ferait pas. Par exemple, elle peut mal classer des images en raison de corrélations accidentelles de l'arrière-plan dans le jeu de données d'entraînement, d'angles de vue inhabituels ou de manipulations de caractéristiques de très bas niveau que les humains ne perçoivent pas activement. Des acteurs malintentionnés peuvent ainsi exploiter certaines vulnérabilités nouvelles, non encore résolues et souvent inconnues, pour altérer la qualité des décisions prises par des systèmes basés sur l'IA. Il peut s'agir, par exemple, d'attaques visant à «empoisonner les données» (injection dans les données d'entraînement d'éléments poussant l'algorithme d'apprentissage à faire des erreurs) ou d'«exemples contradictoires», c'est-à-dire des entrées numériques et des objets de la vie réelle conçus pour tromper les solutions d'apprentissage automatique. Ces derniers sont plus efficaces si les paramètres du modèle d'IA sont connus. Ce procédé porte le nom d'«attaque en boîte blanche». Les exemples contradictoires peuvent aussi fonctionner sans connaître ces éléments. C'est alors une «attaque en boîte noire».

L'IA au service de la détection des logiciels malveillants

En entraînant un réseau neuronal à partir d'un grand ensemble de fichiers dont le contenu est identifié comme acceptable ou comme malveillant, on le dote d'une intuition relativement fiable pour déterminer si un nouveau fichier est malveillant sans avoir à s'appuyer sur des listes mises à jour manuellement. Cela aidera très probablement à **identifier les logiciels malveillants modernes et émergents**, capables de générer automatiquement de nouvelles variantes pour échapper aux méthodes traditionnelles d'identification basées sur des règles, et à catégoriser ces variantes dans la bonne famille de logiciels malveillants. Cependant, cette tâche de classification binaire est loin d'être facile. La prévalence des logiciels malveillants étant très faible dans la masse des fichiers, **ces classificateurs peuvent générer des faux positifs** et bloquer les fichiers exécutables de logiciels légitimes. Pour contourner ce problème, certaines entreprises ont établi une «liste blanche» qui recense les types de fichiers inoffensifs. Les chercheurs ont toutefois montré qu'il était facile d'associer des fichiers figurant sur la liste blanche à des logiciels malveillants afin que ceux-ci passent inaperçus. Dans un avenir proche, **le recours à l'IA pour détecter des logiciels malveillants est donc plutôt voué à compléter les méthodes traditionnelles, et non à les remplacer**.

L'IA peut accroître les cybermenaces, modifier leur caractère typique et introduire de nouvelles menaces inconnues.

L'IA et les cybermenaces

Le déploiement de composants d'IA dans un intérêt cyber peut avoir trois types de répercussions sur le paysage des cybermenaces. En l'absence de mesures préventives substantielles, l'IA peut: accroître les cybermenaces existantes (quantité); modifier le caractère typique de ces menaces (qualité); et introduire de nouvelles menaces inconnues (quantité et qualité).

L'IA pourrait agrandir le nombre d'acteurs capables de mener des cyberactivités malveillantes, le rythme auquel ces acteurs conduiraient ces activités et les cibles possibles. Cet élargissement est dû à l'efficacité, l'évolutivité et l'adaptabilité de l'IA, ainsi qu'à la «démocratisation» des activités de recherche et développement dans ce domaine. La diffusion de composants d'IA auprès des acteurs traditionnels des cybermenaces (États, criminels, hacktivistes et groupes terroristes) pourrait, en particulier, accroître le nombre d'entités qui auraient les moyens de mener des attaques. Les applications de l'IA étant évolutives, les acteurs possédant les ressources nécessaires pour réaliser des attaques pourraient être en mesure d'intensifier leur fréquence. De nouvelles cibles pourraient également devenir intéressantes pour eux.

D'un point de vue qualitatif, les cyberattaques utilisant l'IA pourraient se traduire par des actions plus efficaces, mieux ciblées et plus sophistiquées. Ce gain découle de l'efficacité, l'évolutivité et l'adaptabilité de ces solutions. Les cibles potentielles seront alors plus faciles à identifier et à analyser.

Enfin, l'IA pourrait rendre possible une nouvelle gamme d'activités malveillantes qui exploitent les vulnérabilités que ces technologies ont introduites dans les cybersystèmes auxquels elles sont intégrées. La cybersécurité deviendra alors, en soi, l'un des axes de recherche et développement de l'IA. Pour préserver leur bon fonctionnement, leur fiabilité et leur intégrité, ainsi que pour éviter les effets néfastes, les cybersystèmes intégrant de l'IA doivent être protégés contre les cyberincidents ou les cyberattaques. L'adoption de pratiques de cybersécurité associée à la promotion de vastes programmes de cyberhygiène avec des exigences spécifiques concernant la recherche, le développement et l'application de l'IA est regroupée sous la notion de «cybersécurité pour l'IA».

Usage défensif et offensif

De nombreuses caractéristiques qui font de l'IA un outil adapté aux cyberapplications défensives présentent également un intérêt pour les cyberopérations offensives. Au cours des trois à cinq prochaines années, on doit donc s'attendre à ce que des organisa-

tions adoptent et déploient des capacités de cyberdéfense basées sur l'IA pour protéger leurs ressources (réseaux, informations, personnes) contre des ennemis qui pourraient utiliser des outils d'IA ou autres à des fins offensives. De même, certains acteurs utiliseront des capacités cyberoffensives basées sur l'IA pour compromettre des cibles qui pourraient s'engager dans des cyberopérations défensives intégrant ou non de l'IA. Des cybercapacités reposant sur l'IA peuvent notamment soutenir des activités visant à exécuter des opérations sur des réseaux informatiques ou à s'en protéger, qu'il s'agisse d'opérations d'attaque ou d'exploitation. Elles peuvent également appuyer des systèmes conçus pour réaliser des activités de cyberinformation et de cyberinfluence ou pour s'en défendre.

Les opérations défensives et offensives peuvent toutes deux tirer parti du déploiement de l'IA pour produire du cyberrenseignement, c'est-à-dire des connaissances aptes à soutenir la prise de décisions sur les questions liées au cyberespace. En effet, l'IA est intégrable dans plusieurs fonctions du processus de cyberrenseignement, en particulier la collecte, le traitement et l'analyse d'informations. Elle peut renforcer la collecte d'informations et élargir son champ d'action à de multiples sources et à plusieurs points finaux. Elle peut également améliorer la sélection des informations et les corroborer avec des données additionnelles fournies par d'autres sources.

Les médias de synthèse peuvent être utilisés à des fins de chantage, d'arnaque, de sabotage et pour la propagande politique.

L'IA peut aussi aider à l'analyse en trouvant des corrélations et des modèles cachés dans les données traitées. L'intégration de capacités d'IA dans ces fonctions fera certainement progresser le processus de cyberrenseignement sur les plans de la vitesse et de l'automatisation.

Les réseaux informatiques

La capacité des composants d'IA à produire du cyberrenseignement se traduira par des applications défensives spécifiques aux niveaux tactique/technique et opérationnel de la cybersécurité. Sur le plan opérationnel, l'IA pourrait servir à récupérer et traiter des données recueillies par des programmes d'analyse de la sécurité des réseaux, puis à les mettre en corrélation avec d'autres informations disponibles. Sur le plan tactique, l'IA sera de plus en plus intégrée dans la détection, l'analyse, voire la prévention des cybermenaces. Elle améliorera plus particulièrement les systèmes de détection des intrusions (*Intrusion Detection Systems*, IDS) visant à identifier des activités illicites dans un ordinateur ou un réseau. Les systèmes de détection du spam et du *phishing* (hameçonnage) ainsi que les outils d'analyse et de détection des logiciels malveillants en tireront également parti (voir encadré). Des composants d'IA s'intègreront également dans des systèmes d'authentification ou de vérification multifacteur. Ils aideront à déterminer un modèle de comportement caractéristique d'un utilisateur afin d'identifier tout changement dans ce modèle. Une autre cible prometteuse des applications défensives tactiques de l'IA est le test de vulnérabilité automatisé, également appelé *fuzzing*.

Des applications de l'IA seront également utilisées à des fins cyberoffensives, c'est-à-dire pour compromettre une organisation ou un utilisateur cible, ses réseaux et les données qu'ils traitent. Elles permettront des cyberattaques plus nombreuses et plus sophistiquées. Comme pour les opérations de cyberdéfense, les applications de l'IA peuvent produire du cyberrenseignement permettant de préparer et mettre en œuvre ces attaques. Elles peuvent affiner la sélection et la hiérarchisation des cibles pour des cyberattaques faisant appel à l'ingénierie sociale. Ces attaques consistent à manipuler psychologiquement les utilisateurs cibles pour les amener à révéler certaines informations ou à effectuer certaines actions dans un but non légitime. Grâce à l'IA, les auteurs de ces attaques peuvent recueillir et traiter des informations en ligne sur les victimes potentielles afin de créer automatiquement des sites web, des e-mails et des liens malveillants adaptés au profil de la personne.

Les composants d'IA amélioreront également la découverte et l'exploitation de vulnérabilités à des fins malintentionnées. Ils aideront à perfectionner la conception et le fonctionnement des logiciels malveillants et faciliteront leur dissimulation. Les logiciels malveillants utilisant l'IA peuvent échapper aux systèmes de détection et réagir de manière créative aux changements de comportement des cibles. Ils fonctionneront comme un implant autonome et adaptatif qui apprend de son hôte afin de rester inaperçu et de pouvoir alors chercher et classer les contenus intéressants à exfiltrer, détecter et infecter de nouvelles cibles et découvrir de nouvelles voies ou méthodes pour se déplacer dans un réseau et trouver les données clés qui constituent la cible ultime de l'attaque. Les chercheurs d'IBM ont développé dès 2018 un logiciel malveillant de ce type appelé Deeplocker. Enfin, l'IA sera également déployée pour tromper des systèmes d'authentification ou de vérification, comme ceux intégrant des identifiants biométriques.

Cyberinformation et cyberinfluence

L'IA va probablement améliorer la planification et l'exécution des opérations de cyberinformation et de cyberinfluence. En renforçant l'automatisation, elle favorisera la collecte numérique d'informations et la surveillance du comportement des cibles en ligne. Elle élargira la palette d'outils permettant d'informer et d'influencer des adversaires dans et par le biais du cyberespace, en s'appuyant notamment sur les plateformes de réseaux sociaux. Dans le domaine des réseaux sociaux, justement, l'IA peut améliorer la gestion des bots et bots sociaux et permettre la production de messages ciblant spécifiquement les personnes les plus susceptibles d'y être réceptives. En phase avec la tendance actuelle, des solutions basées sur l'IA, notamment celles qui intègrent des réseaux antagonistes génératifs profonds, aideront à créer des contenus numériques manipulés. Ces contenus, appelés «hypertrucages» ou *deepfakes*, sont des vidéos, des enregistrements audio, des images ou des textes hyperréalistes dont le

Pour aller plus loin

Matteo E. Bonfanti, «Artificial Intelligence and the Offence-Defence Balance in Cyber Security», dans: Dunn Cavelty, M. & Wenger A. (dir.), *Cyber Security Politics: Socio-Technological Transformations and Political Fragmentation,* Routledge, 2020, à paraître.

Matteo E. Bonfanti, «Cyber Intelligence: In Pursuit of a Better Understanding for an Emerging Practice», dans: *INSS Cyber, Intelligence, and Security* vol. 2 no 1, 2018, p. 105–121.

Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, février 2018.

Ben Buchanan, A National Security Research Agenda for Cybersecurity and Artificial Intelligence. Center for Security and Emerging Technology, CSET Issue Brief, mai 2020.

caractère falsifié n'est pas facilement identifiable par des méthodes manuelles ou d'autres techniques d'analyse conventionnelles. Une fois générés, les hypertrucages peuvent être utilisés à des fins abusives. Les exemples sont déjà nombreux et documentés dans les médias. Il s'agit le plus souvent de vidéos falsifiées grâce à l'IA qui sont diffusées dans le but de diffamer, de harceler ou d'intimider des cibles en ligne. Les médias de synthèse peuvent être de véritables armes dans le cyberespace. À court terme, on observera probablement une hausse de leur utilisation à des fins de chantage, d'arnaque, de sabotage d'entreprises via le marché ou d'autres types de manipulations, ainsi que pour la propagande politique.

Si l'IA s'intègre dans toutes ces activités et les rend possibles, elle pourra aussi aider à les contrer. D'un point de vue défensif, l'IA peut soutenir la détection des opérations de cyberinformation et cyberinfluence, et contribuer à y répondre. Elle peut aider à surveiller l'environnement en ligne, notamment les plateformes de réseaux sociaux, à identifier les premiers signes d'opérations malveillantes, tels qu'une hausse de l'activité des bots ou bots sociaux, et à mettre en évidence les contenus numériques modifiés, y compris les médias de synthèse.

Une question de gouvernance

L'IA aura une incidence sur la cybersécurité dans les années à venir. Elle étoffera le paysage des cybermenaces, aussi bien sur le plan quantitatif que qualitatif. Elle augmentera probablement le nombre d'acteurs des cybermenaces en leur offrant davantage de vulnérabilités et de cibles à exploiter et en donnant un nouveau souffle à leurs actions malintentionnées. À l'inverse, l'IA aidera aussi à se défendre contre ces menaces en permettant la découverte de vulnérabilités inconnues, la détection de cyberactivités malveillantes et la mise en œuvre de contre-mesures. Elle soutiendra aussi bien les cyberopérations défensives qu'offensives. Lesquelles en tireront le plus de bénéfices? Cette question est difficile à trancher. Cela dépendra probablement de la capacité des acteurs publics ou privés de la cybersécurité à maîtriser et à exploiter l'IA. Cela dépendra également de leur aptitude générale à identifier, comprendre et gérer les risques, les menaces et les opportunités associés au déploiement de ces technologies.

> Les gouvernements peuvent jouer un rôle important en pilotant la transformation de la cybersécurité induite par l'IA.

Face à ces risques et opportunités, les gouvernements peuvent jouer un rôle important en pilotant la transformation de la cybersécurité induite par l'IA. Jusqu'à présent, ils ont soutenu l'innovation en matière d'IA par divers mécanismes politiques. Ils ont investi dans des infrastructures d'IA, encouragé l'enseignement universitaire et la formation professionnelle, financé la recherche scientifique, favorisé les partenariats et la collaboration public-privé et promu des normes à travers leurs politiques d'achat. En consultation avec le secteur privé et la société civile, ils ont appuyé l'adoption de principes directeurs ou de normes de base (droits fondamentaux, confidentialité des données) dans le but de favoriser une innovation responsable et fiable dans ce domaine technologique.

Les gouvernements de nombreux pays orientent leurs actions vers l'acquisition de capacités d'IA dans le cadre de stratégies nationales de grande envergure, dont la plupart considèrent la cybersécurité comme un domaine d'application prometteur. Ces stratégies sont ensuite complétées par des instruments politiques sectoriels ou d'autres documents techniques. En général, elles ont pour objectif de mettre des capacités d'IA à la disposition des acteurs nationaux de la cybersécurité et de faire en sorte que ces derniers s'appuient sur l'IA pour obtenir un avantage sur leurs concurrents.

Pour influer sur la transformation de la cybersécurité induite par l'IA, les gouvernements peuvent également définir des normes d'essai dynamique, de validation et de certification des outils d'IA pour les applications cyber. Au niveau international, ils peuvent travailler à l'élaboration de normes communes pour la recherche et le développement en matière d'IA et réfléchir à des contraintes intelligentes pour maîtriser la prolifération des connaissances et des capacités dans ce domaine technologique. Les gouvernements peuvent également favoriser une gouvernance positive et inclusive de l'IA en mettant en place des principes de haut niveau, à l'image de ceux adoptés par l'UE et l'Organisation de coopération et de développement économiques (OCDE) pour une IA fiable.

Voir le site thématique du CSS pour en savoir plus sur la cybersécurité.

Matteo E. Bonfanti est Senior Researcher de l'équipe «Risk and Resilience» au CSS.

Kevin Kohler est Researcher au sein de l'équipe «Risk and Resilience» au CSS.

* *Solution du puzzle de la page 1: En fait, les deux photos sont générées par l'IA. Ces «enfants» n'existent pas.*

Les **analyses de politique de sécurité** du CSS sont publiées par le Center for Security Studies (CSS) de l'EPF de Zurich. Le CSS est un centre de compétence en matière de politique de sécurité suisse et internationale. Deux analyses paraissent chaque mois en allemand, français et anglais.

Editeur: Benno Zogg
Révision linguistique: Névine Schepers
Layout: Rosa Guggenheim

Feedback et commentaires: analysen@sipo.gess.ethz.ch
Plus d'éditions et abonnement: www.css.ethz.ch/cssanalysen

Parus précédemment:

Technologies numériques et la crise du coronavirus No 264
Les Balkans occidentaux entre nouvelles dépendances No 263
La candidature suisse au Conseil de sécurité de l'ONU No 262
Ukraine: la dimension religieuse du conflit No 261
L'intégration de l'IA dans la protection de la population No 260
Ukraine: la dimension religieuse du conflit No 259

ISSN: 2296-0228; DOI: 10.3929/ethz-b-000417543