# ChatGPT, le félin des codes défaillants : faux détecteur ou véritable tigre de la programmation ?

L’IA générative et les modèles de machine learning dans le domaine de la cybersécurité

## L’IA générative et les modèles de machine learning dans le domaine de la cybersécurité

Selon un rapport de NCC Group, les modèles de machine learning sont très prometteurs pour détecter des attaques zero day. Mais pour débusquer les vulnérabilités dans le code avec l’IA générative, la partie est loin d’être gagnée.

L’IA générative – en particulier ChatGPT – ne devrait pas être considérée comme une ressource fiable pour détecter les vulnérabilités dans le code sans la supervision cruciale d’un expert humain. Toutefois, les modèles d’IA sont très prometteurs pour observer les attaques de type zero day. C’est ce qui ressort d’un dernier rapport du NCC Group, qui explore

divers cas d'utilisation de l'IA dans le domaine de la cybersécurité. Le rapport « Safety, Security, Privacy & Prompts : Cyber Resilience in the Age of Artificial Intelligence » a été publié pour aider ceux qui souhaitent mieux comprendre comment l'IA s'applique à la cybersécurité, en résumant comment elle peut être utilisée par les professionnels de ce domaine.

Cette question a fait l'objet d'un grand nombre de discussions, de recherches et d'opinions cette année, suite à l'arrivée explosive et à la croissance de la technologie de l'IA générative à la fin de l'année 2022. On a beaucoup parlé des risques de sécurité que présentent les chatbots d'IA générative, qu'il s'agisse de préoccupations concernant le partage d'informations commerciales sensibles avec des algorithmes d'auto-apprentissage avancés ou d'acteurs malveillants qui les utilisent pour renforcer considérablement les attaques. De même, beaucoup affirment que, s'ils sont utilisés correctement, les chatbots d'IA générative peuvent améliorer les défenses de cybersécurité.

## La surveillance humaine essentielle pour détecter les failles de sécurité du code

L'un des points clés du rapport est de savoir si le code source peut être introduit dans un chatbot d'IA générative pour savoir s'il contient des bugs de sécurité et mettre en évidence avec précision les vulnérabilités potentielles pour les développeurs. Malgré les promesses et les gains de productivité que l'IA générative apporte dans le développement de code, elle a cependant montré des résultats mitigés dans sa capacité à dénicher efficacement les vulnérabilités du code, a constaté NCC Group.

« L'efficacité, ou non, de ces approches utilisant les modèles

actuels a fait l'objet d'une recherche de NCC Group, la conclusion étant que la supervision humaine experte est toujours cruciale », peut-on lire dans le rapport. À l'aide d'exemples de code non sécurisé provenant de Damn Vulnerable Web Application (DVWA), il a été demandé à ChatGPT de décrire les vulnérabilités dans une série d'exemples de code source PHP non sécurisé. Au bout du compte les résultats sont mitigés et ne constituent certainement pas un moyen fiable de trouver les vulnérabilités dans le code développé.

## Le machine learning « simple » efficace pour débusquer les failles zero day

Un autre cas d'utilisation de l'IA dans le domaine de la cybersécurité défensive exploré dans l'étude porte sur l'utilisation de modèles d'apprentissage machine (ML) pour aider à la détection d'attaques de type zero day et apporter ainsi une réponse automatisée afin de protéger les utilisateurs de malware. Pour se faire, la société a parrainé un étudiant en master au Centre for Doctoral Training in Data Intensive Science (CDT DIS) de l'University College London (UCL) pour développer un modèle de classification pour déterminer si un fichier est un malware. Non sans réussite : « Plusieurs modèles ont été testés, le plus performant atteignant une précision de classification de 98,9 % », peut-on lire dans le rapport.

Le renseignement sur les menaces implique la surveillance de multiples sources de données en ligne fournissant des flux de données de renseignement sur les vulnérabilités nouvellement identifiées, les exploits développés, et les tendances et modèles dans le comportement des attaquants. « Ces données sont souvent des données textuelles non structurées provenant de forums, de médias sociaux et du dark web. Les modèles de ML peuvent être utilisés pour traiter ces informations,

identifier les nuances communes de cybersécurité dans les données, et donc identifier les tendances dans les tactiques, techniques et procédures (TTP) des attaquants », selon le rapport. Cela permet aux défenseurs de mettre en œuvre de manière proactive et préventive des systèmes de surveillance ou de contrôle supplémentaires si des menaces sont particulièrement importantes pour leur entreprise ou leur paysage technologique, peut-on lire dans l’étude.