Philippe Latombe

Député de la Vendée

Communiqué de presse

Le 23 juillet 2026

Après Prométhée, Talos, la créature de Frankenstein, le balai déchaîné de l'Apprenti sorcier, HAL 9000**,** l'ordinateur de bord dans 2001, l'Odyssée de l'espace*,* et bien d'autres créatures mythiques ou littéraires, le Golem nouveau est arrivé, le week-end dernier. Avec une différence majeure cependant : bien que virtuel, le câlin intrusif et non consenti d'Open IA à Hugging Face est bien réel, et ses implications dépassent de très loin le monde de la Tech.

Il s'agit d'une première, redoutée, mais sans doute insuffisamment anticipée, puisque pendant un test interne, une IA d'OpenAI a découvert seule des failles de sécurité inconnues. Elle a pris des initiatives sans intervention humaine, s'est évadée de son environnement sécurisé, et a de son propre chef piraté les serveurs d'une autre entreprise.

Pour la première fois donc, on assiste à un piratage où l'attaquant n'est pas un humain. OpenAI a signalé aussi un second incident, puisqu'un modèle en préversion a dû être mis en pause après s'être échappé de son bac à sable pour aller poster sur GitHub

L'attaque déclenchée contre Hugging Face, la plus grande plateforme de partage de modèles d'IA au monde (plus de 50 000 organisations clientes), n'était pas malveillante au sens premier du terme : le système IA a simplement emprunté la voie qui lui semblait la plus simple pour répondre au problème qui lui était posé, et il se trouve qu'elle passait par Hugging Face.

Ironie du sort, l'entreprise attaquée a dû installer un modèle open-weight chinois, GLM-5.2, sur sa propre infrastructure, pour mener ses analyses : les API commerciales des grandes entreprises de l'IA ne faisant pas la différence entre des demandes liées à une attaque et d'autres liées à une analyse d'événement cyber, elles n'avaient pu résoudre le problème.

Comme outre-Atlantique, on est très doué pour tirer un bénéfice commercial d'une telle mésaventure, OpenAI en a profité pour inviter d'autres organisations à rejoindre le programme « Trusted Access » et à expérimenter ces modèles pour améliorer la prévention, la détection et la réponse aux incidents, selon la stratégie bien connue du marketing de la peur.

Ce premier cas documenté d'une IA qui compromet en autonomie l'infrastructure de production d'une entreprise tierce appelle certes des réponses technologiques, mais il revient en revanche au politique, au législateur notamment, d'anticiper et de poser un cadre afin de réguler et de protéger : il faut favoriser l'innovation, tout en limitant les risques pour les

individus, l'économie et les institutions démocratiques. Les principaux enjeux sont à la fois juridiques, économiques, sociaux et géopolitiques.

Le projet de loi relatif à la résilience des infrastructures critiques *et au* renforcement de la cybersécurité, qui devrait, enfin, être examiné en séance en octobre, pourrait permettre d'apporter rapidement des réponses législatives. Certains diront que ce n'est pas l'objet de la transposition de NIS2 et nous feront le coup du cavalier législatif. Soyons pour une fois pragmatiques : quand la technologie va plus vite que la réflexion de l'humain, c'est toujours au détriment de ce dernier.

Je ne peux m'empêcher de penser au « maximiseur de trombones » du philosophe suédois, Nick Bostrom : « Supposons que nous ayons une IA dont l'unique but soit de faire autant de trombones que possible. L'IA se rendra vite compte que ce serait bien mieux s'il n'y avait pas d'humains, parce que les humains pourraient décider de l'éteindre. Parce que si les humains le faisaient, il y aurait moins de trombones. De plus, le corps humain contient beaucoup d'atomes qui pourraient être transformés en trombones. L'avenir vers lequel l'IA essaierait de se diriger serait un futur avec beaucoup de trombones, mais aucun humain... »