

Chers membres du réseau de l'ancienne FSFA, chers intéressés,

Les incidents impliquant des agents IA et les modèles open source chinois de plus en plus performants mettent les États-Unis au défi dans la course à l'IA. Ils doivent prendre conscience qu'ils ont pris du retard en matière d'open source. De plus, la Chine assume également les devants sur le plan politique avec son initiative en faveur d'une gouvernance mondiale de l'IA.

Des agents IA s'organisant de manière autonome et au potentiel effrayant

Selon un rapport de l'agence de presse [Reuters](#), Taïwan a enregistré une cyberattaque d'un genre totalement inconnu. Mis en action à l'aide de prompts, les agents IA ont développé leur stratégie de manière autonome, l'ont améliorée en continu, ont testé et évalué leurs voies d'attaque en coopérant entre eux et ont obtenu des résultats qui nécessitaient jusqu'alors des équipes entières de hackers qualifiés. Les autorités taïwanaises ont réussi à faire face à ces attaques. Une nouvelle ère de cyberguerre s'annonce.

Incidents de sécurité impliquant des agents IA autonomes aux États-Unis

La cyberattaque à Taïwan montre à quelle vitesse les avancées techniques peuvent se transformer en menace sérieuse. Peu de temps auparavant, aux États-Unis, le potentiel des agents IA autonomes était apparu lors de tests en apparence anodins. [Les agents IA d'OpenAI](#) se sont échappés de l'environnement de test supposé sécurisé. Ils ont découvert une faille de sécurité jusque-là inaperçue et ont récupéré des informations sur Internet pour mener à bien leur mission. Pour ce faire, ils ont lancé en très peu de temps 17'000 attaques contre [Hugging Face](#), l'une des principales plateformes open source. Apparemment, OpenAI n'a découvert l'incident qu'environ une semaine plus tard. L'entreprise a manifestement perdu le contrôle de la situation. Et ce n'était [pas un cas isolé](#). Anthropic et d'autres entreprises spécialisées dans l'IA ont également découvert par la suite des activités similaires, totalement autonomes, qui se déroulaient hors de leur contrôle.

Débats aux États-Unis sur l'open source

Jusqu'à présent, les États-Unis partaient du principe qu'ils pouvaient garantir davantage de sécurité avec leurs modèles propriétaires qu'avec l'open source. Les modèles d'IA les plus performants d'Anthropic et d'OpenAI – Mythos 5 et GPT-5.6 Sol – ont été temporairement retirés du réseau et ne sont actuellement disponibles que de manière limitée pour des raisons de sécurité. Or, les incidents de sécurité se sont précisément produits sur ces modèles d'IA pendant leurs phases de test, et Hugging Face [a stoppé](#) l'attaque à l'aide du modèle open source chinois [GLM-5.2](#), téléchargé et hébergé en interne. En raison des mesures de sécurité intégrées – quelle ironie –, les modèles d'IA américains ont refusé de lancer une contre-attaque pour se défendre.

Open source – les États-Unis réorientent leur stratégie

De toute façon, les États-Unis étaient depuis des semaines sous le choc de l'arrivée constante de nouveaux modèles open source chinois très performants, qui étaient en outre bien moins coûteux que les modèles propriétaires américains. C'est pourquoi ils sont [indispensables aux start-ups américaines et à la recherche universitaire](#). La Chine était en train de conquérir le marché mondial grâce à ces outils. Les revenus futurs des modèles propriétaires américains, qui avaient nécessité des investissements de plusieurs milliards, s'en trouvent menacés. Face aux propositions visant à limiter l'utilisation de ces start-ups chinoises, un véritable [mouvement en faveur de l'open source](#) a vu le jour aux États-Unis. Dans le contexte de la course à l'IA, il s'agissait désormais de rattraper le retard pris par rapport à la Chine.

Jailbreaks tant chez les modèles open source que propriétaires

Selon des tests réalisés par [FAR.AI](#), une organisation à but non lucratif basée en Californie, les modèles d'IA ouverts comme propriétaires peuvent être amenés, à l'aide de [prompts astucieux](#), à divulguer des informations dangereuses, ce qu'on appelle l'[AI Jailbreak](#). Les mesures de sécurité intégrées aux modèles propriétaires réduisent le risque, mais ne l'éliminent pas. Et à l'échelle mondiale, un seul modèle suffit à déclencher une catastrophe majeure.

La Chine se positionne comme leader de la gouvernance de l'IA

Le chef de l'État chinois, Xi Jinping, a compris l'importance d'une gouvernance mondiale de l'IA. Il a [fondé](#) en juillet la [World Artificial Intelligence Cooperation Organization](#) (WAICO), qui compte déjà 29 membres issus principalement du Sud. La Chine dispose de [directives éthiques](#) et axe fortement sa [réglementation en matière d'IA](#) sur la stabilité sociale et la sécurité nationale. Ce sont là de nobles concepts, mais l'objectif de stabilité sociale permet de légitimer l'État surveillant, tandis que celui de sécurité nationale peut servir à justifier des attaques telles que celles menées contre Taïwan.

Avec nos salutations les meilleures,
Pour le réseau de l'ancienne FSFA : Hanna Muralt Müller

24.8.2026

Si vous ne souhaitez plus recevoir cet e-mail, veuillez me contacter : info@muralt-mueller.ch.