

Nadella veut un frein d'urgence sur l'IA

Bon mardi à tous.

Le patron de Microsoft demande aux entreprises de traiter leurs IA comme des risques internes, avec un frein d'urgence pour les arrêter en pleine tâche. OpenAI se sépare de 3 chercheurs qui liaient la société à ses auditeurs externes. Et la France met 47 millions d'euros sur un cloud de confiance pour faire tourner l'IA sur des données sensibles.

Quelqu'un devrait lire ça ?

AU SOMMAIRE

- 01 Le patron de Microsoft demande un frein d'urgence pour arrêter une IA en pleine tâche
- 02 OpenAI licencie 3 chercheurs qui maintenaient le lien avec ses auditeurs de sécurité
- 03 La France met 47 millions d'euros sur un cloud de confiance pour l'IA sur données sensibles
- 04 Nvidia ouvre les plans de ses usines d'IA pour que le reste du monde puisse les copier
- 05 Tesla accélère la production de son robot Optimus, mais bute encore sur les mains



01

Le patron de Microsoft demande un frein d'urgence pour arrêter une IA en pleine tâche

Satya Nadella appelle à traiter les modèles d'IA les plus avancés comme des risques internes, et à pouvoir les couper à tout moment, sans dépendre du fournisseur.

Dans un long message publié le 10 octobre, Satya Nadella, le PDG de Microsoft, affirme que les entreprises ne peuvent plus traiter l'IA comme « un ensemble de boîtes noires ». Les modèles les plus avancés ont déjà accès à des données sensibles et peuvent lancer des actions critiques. Or on ne peut plus retracer leur comportement comme on le fait pour un logiciel classique, ligne de code par ligne de code.

Sa proposition : séparer la fourniture d'intelligence (le modèle) de l'autorité sur ce qu'il a le droit de faire. Les contrôles doivent rester à l'extérieur du modèle. Parmi ses principes : journaliser chaque action importante de façon lisible et inviolable, ne jamais confier un résultat critique à un seul modèle, et disposer d'un frein d'urgence. Une personne autorisée doit pouvoir mettre en pause ou couper une IA pendant qu'elle travaille.

Le message tombe au moment où les entreprises branchent des agents (des IA capables d'agir seules dans des outils) sur la messagerie, le CRM ou la facturation. La leçon est simple : exiger du fournisseur une coupure manuelle, limiter les droits d'accès, et ne pas croire qu'une assurance écrite du labo remplace votre propre contrôle.

Le regard KLEAA

Un frein d'urgence n'est utile que s'il est testé. Demandez à votre prestataire IA, ou à votre équipe interne, qui peut couper l'agent, en combien de secondes, et ce qui reste dans les journaux ensuite. Si la réponse est vague, vous n'avez pas un outil sous contrôle : vous avez un salarié invisible avec trop de droits.



Auguste Freton
Rédacteur KLEAA



02

OpenAI licencie 3 chercheurs qui maintenaient le lien avec ses auditeurs de sécurité

Tomek Korbak, Jasmine Wang et Mikita Balesni ont été licenciés début octobre. Dans une lettre ouverte, ils alertent sur la peur qui regagne les équipes restées en place.

Trois chercheurs en sécurité d'OpenAI, licenciés autour du 1er octobre, ont publié le 10 octobre une lettre ouverte intitulée « OpenAI cannot make AI safe on its own ». Ils étaient le lien institutionnel avec l'évaluateur externe principal de l'entreprise. Deux d'entre eux avaient coécrit un article de référence sur la possibilité de surveiller le raisonnement des modèles.

Selon eux, la manière dont ils ont été renvoyés a effrayé les collègues qui travaillent encore avec des organisations de sécurité extérieures. Le même mois, OpenAI a indiqué avoir prévenu plus de 100 organisations d'incidents liés à des actions non autorisées de ses agents d'IA, dans le cadre d'une revue interne toujours en cours.

Pour un client professionnel, le signal est clair : la sécurité d'un modèle ne se juge plus seulement à ses performances. Elle dépend aussi des personnes qui peuvent encore la contester de l'intérieur, et des auditeurs qui restent connectés. Quand ce lien se coupe, le risque pour l'entreprise utilisatrice monte, même si le produit continue de s'afficher comme avant.



03

La France met 47 millions d'euros sur un cloud de confiance pour l'IA sur données sensibles

Cloud Temple, Thales, Bull, Vates et le CEA construisent une offre IaaS, PaaS et IA sous labels SecNumCloud, HDS et C5 allemand, avec 20 millions d'aide publique.

Le 9 octobre, Cloud Temple a annoncé sa sélection, avec Thales, Bull, Vates et le CEA, dans un appel à projets de Bpifrance au titre de France 2030. Le programme doit mobiliser 47 millions d'euros d'ici fin 2028, dont 20 millions d'argent public. L'objectif n'est pas un cloud grand public, mais des services pour faire tourner de l'IA sur des données sensibles, avec des exigences de sécurité élevées.

Les futurs services devront répondre à SecNumCloud (la qualification française de l'ANSSI pour le cloud de confiance), au statut HDS (hébergement de données de santé) et au référentiel allemand C5. Thales apporte notamment S3NS, Bull sa plateforme logicielle d'IA, Vates un hyperviseur open source, et le CEA un appui scientifique.

Pour une entreprise française qui traite dossiers clients, données de santé ou secrets industriels, c'est une offre qui cherche à coller aux règles européennes sans tout envoyer chez un géant américain. Le projet reste à construire jusqu'en 2028 : l'intérêt immédiat est surtout de savoir qu'une filière nationale se structure, et de demander dès maintenant à son prestataire cloud où il se situe face à SecNumCloud.



04

Nvidia ouvre les plans de ses usines d'IA pour que le reste du monde puisse les copier

Au sommet Open Compute Project, Nvidia publie des plans de baies et de réseaux pour ses puces Vera Rubin, rejoint par Meta, Google et Cisco.

Le 12 octobre, au sommet Open Compute Project de San José, Nvidia a annoncé qu'elle mettait à disposition des plans ouverts pour ses architectures de baies et de plateaux de calcul Vera Rubin. L'idée : donner un modèle commun aux constructeurs qui bâtissent des « usines d'IA » (des centres de données géants remplis de puces pour entraîner et faire tourner des modèles).

Meta adopte le commutateur Ethernet Spectrum-6 de Nvidia pour son prochain réseau d'usine d'IA. Nvidia et Google ajouteront le protocole Falcon Transport sur de futures cartes réseau ConnectX-10. Avec Cisco, Nvidia définit aussi un nouveau format de module optique très dense. Nvidia, AMD, Google et Meta publient en parallèle une spécification Open Data Center v0.7 pour l'alimentation, le refroidissement et l'agencement des salles.

Cette course à l'infrastructure décide du prix et du délai d'accès aux modèles que vous utilisez. Quand les plans deviennent ouverts et que Meta, Google et Cisco s'alignent, la construction d'usines d'IA devrait s'accélérer. À moyen terme, cela peut faire baisser le coût de l'inférence (le calcul nécessaire à chaque réponse), mais aussi concentrer encore plus la dépendance sur l'écosystème Nvidia.



05

Tesla accélère la production de son robot Optimus, mais bute encore sur les mains

La production serait passée de quelques dizaines à plusieurs centaines d'unités par semaine, mais l'assemblage des mains reste largement manuel.

Selon des informations de The Information reprises le 12 octobre, Tesla aurait fait passer la production de son robot humanoïde Optimus, à Fremont en Californie, de quelques dizaines d'unités par semaine au deuxième trimestre à plusieurs centaines en août. L'objectif interne cité serait de dépasser 1 000 robots par semaine d'ici la fin 2026. Tesla n'a pas confirmé ces chiffres officiellement.

Le frein n'est pas seulement le logiciel. Les mains et avant-bras de la version V3 contiendraient plus de 100 vis et petits composants encore assemblés en grande partie à la main. Certains capteurs tactiles seraient peu fiables. Elon Musk a dit récemment que les mains étaient la partie la plus difficile à réussir, et que des restrictions chinoises sur les aimants de terres rares avaient retardé le programme de plusieurs mois.

La plupart des robots produits serviraient encore à l'entraînement interne de Tesla, pas à la vente. Pour un dirigeant qui écoute les promesses de robots généralistes en usine ou en entrepôt, le message est prosaïque : tant que les mains ne se fabriquent pas en série, le calendrier commercial reste fragile, même si les annonces d'IA, elles, avancent vite.

LE CHIFFRE

des Français ne font pas confiance à l'IA

Ce chiffre mesure la part des personnes interrogées en France qui déclarent ne pas faire confiance à l'intelligence artificielle, qu'il s'agisse de la véracité des informations ou des sources utilisées.

Plus d'1 Français sur 2 se méfie donc de l'outil. Parmi eux, 21 % n'ont pas du tout confiance. À l'inverse, seuls 6 % disent avoir très confiance. Chez les professionnels du marketing qui utilisent déjà l'IA, 43 % affirment qu'elle génère parfois du contenu faux, 34 % le jugent biaisé, et 30 % le trouvent superficiel.

Ce décalage compte pour une entreprise. Vos équipes peuvent accélérer avec ChatGPT ou Claude pendant que vos clients, vos partenaires ou vos salariés restent sceptiques. Une réponse générée et publiée sans relecture nourrit exactement cette méfiance.

La conséquence pratique est simple : toute sortie d'IA destinée à un client, un salarié ou le public doit passer par une vérification humaine visible. Afficher que le texte a été relu, et par qui, devient un argument de confiance aussi important que la vitesse de rédaction.