OpenAI : Quand l’IA devient un outil de contrôle, la ligne rouge est franchie !

OpenAI au cœur de la lutte contre la désinformation : révélations sur des réseaux malveillants en Chine

OpenAI, la société à l’origine de l’agent conversationnel ChatGPT, a récemment pris des mesures pour exclure des utilisateurs liés à des activités malintentionnées, vraisemblablement basées en Chine. Dans un rapport de sécurité publié le **21 février dernier**, l’entreprise a exposé des pratiques de manipulation de l’information ainsi que le développement d’un outil de surveillance des réseaux sociaux.

Les équipes de sécurité d’OpenAI ont découvert ces utilisateurs en suivant l’usage de ChatGPT. En étudiant les comportements d’un groupe de comptes opérant en mandarin lors des heures de travail chinoises, ils ont pu cerner les fonctionnalités d’un outil en cours de développement : un programme destiné à analyser en temps réel les contenus diffusés sur des plateformes comme Facebook, X, YouTube, Instagram, Telegram et Reddit. Cet outil semblait se concentrer tout particulièrement sur les messages concernant les manifestations pour les droits humains.

L’outil de surveillance « Peer Review »

Dénommé « *Peer Review* » par OpenAI, cet outil a pour mission de signaler les publications controversées aux autorités chinoises et à ses ambassades à l’étranger. Les États-Unis, l’Allemagne et le Royaume-Uni figurent parmi les cibles de cette surveillance, qui englobe des sujets sensibles comme le soutien aux Ouïgours et les questions diplomatiques dans la région Indo-Pacifique.

ChatGPT était utilisé par ces acteurs pour élaborer des documents promotionnels pour « Peer Review », détaillant ses fonctionnalités et corrigeant les erreurs dans le code identifiant celui-ci comme utilisant le modèle Llama de Meta, un concurrent open-source de ChatGPT. Néanmoins, OpenAI n’a pas été en mesure de déterminer si cet outil avait trouvé une utilisation à grande échelle.

Des opérations de désinformation

OpenAI a également pu surveiller un second réseau d’acteurs chinois impliqué dans des campagnes de désinformation. Ces derniers utilisent ChatGPT pour générer des messages courts publiés sur les réseaux sociaux et visant à discréditer les publications du dissident Cai Xia, ex-professeur à l’école centrale du Parti communiste chinois.

Au cours d’un récent sommet sur l’intelligence artificielle à Paris, Ben Nimmo, enquêteur chez OpenAI, a évoqué un cas similaire qu’il avait traité. Des utilisateurs basés en Chine avaient eu recours à ChatGPT pour soutenir la campagne de désinformation « Spamouflage », rédigeant notamment des tweets faisant la promotion du Parti communiste tout en critiquant l’Occident. Les enquêteurs d’OpenAI avaient réussi à les identifier en raison de l’utilisation des mêmes comptes

ChatGPT pour des activités frauduleuses lors d’examens internes au PCC.

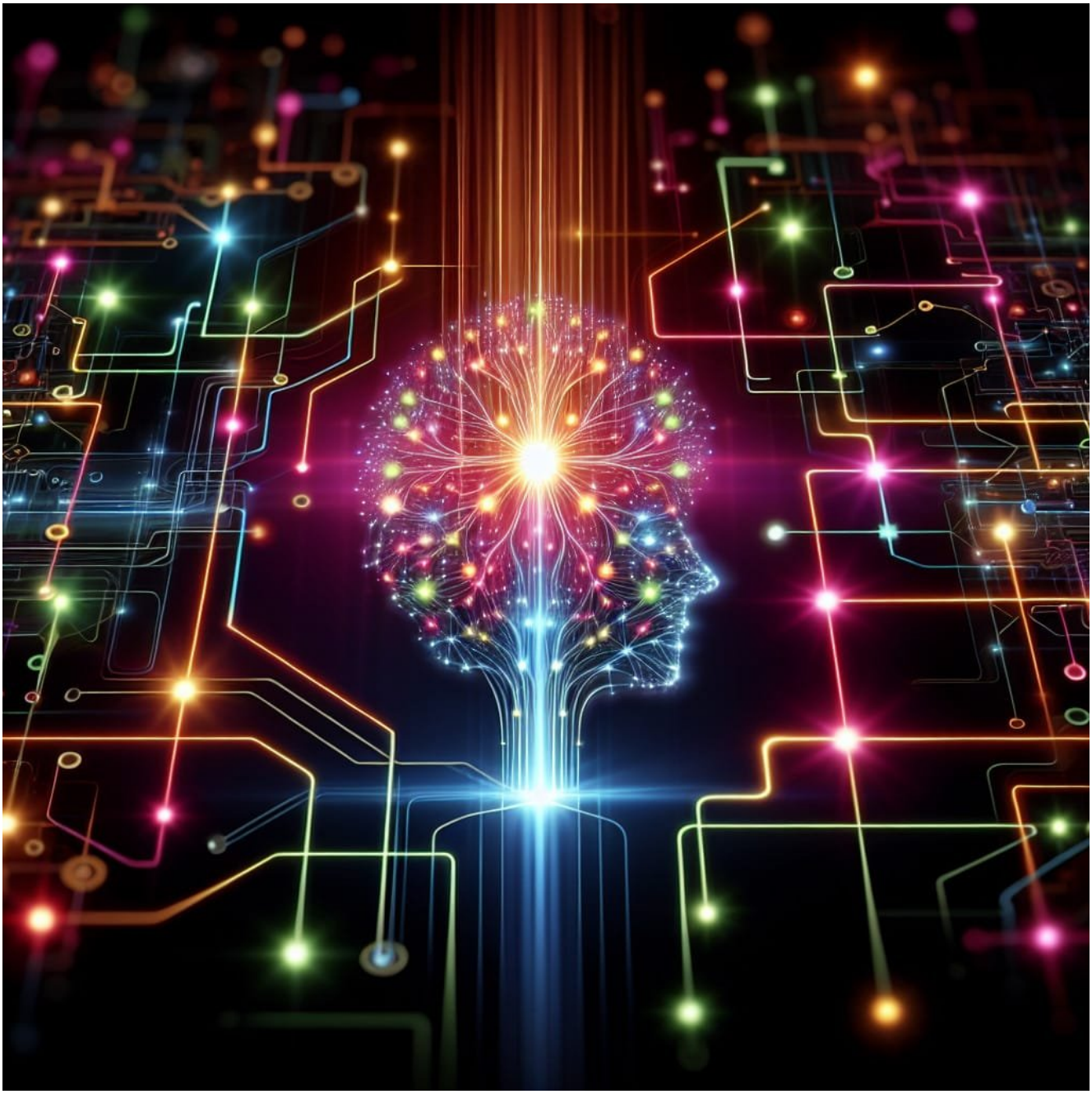
Le groupe ciblant Cai Xia semble différent. Ses membres ont utilisé ChatGPT pour rédiger des articles longs critiquant les États-Unis, publiés dans des journaux mexicains, péruviens et équatoriens, souvent en tant que contenus sponsorisés. L’IA a permis de traduire et d’enrichir des articles préexistants en chinois, discutant de sujets sensibles tels que la violence politique ou la politique étrangère, tout en attribuant ces problématiques à une faiblesse perçue du leadership américain.

OpenAI publie régulièrement des rapports de sécurité pour prévenir l’utilisation abusive d’outils d’intelligence artificielle par des régimes autoritaires. La compagnie espère ainsi mettre un frein à la manipulation des informations et à la répression des citoyens par ces mêmes institutions.

Articles connexes

Comment l’IA muscle les arnaques et la propagande en ligne : pas de tsunami mais des risques croissants

Article rédigé par **Le Monde**



Réutiliser ce contenu

Source : www.lemonde.fr

→ Accéder à CHAT GPT en cliquant dessus