Les bots assistés par ChatGPT pullulent sur les réseaux sociaux

Pour de nombreux utilisateurs, parcourir les fils d'actualité et les notifications des médias sociaux revient à patauger dans la boue. Une nouvelle étude identifie 1 140 robots, assistés par l'IA, qui diffusent des informations erronées sur X (anciennement Twitter) concernant les sujets de crypto-monnaies et de blockchain.

Cependant, les chercheurs ont constaté que les comptes de bots qui publient ce genre de contenu peuvent être difficiles à repérer. Les comptes robots utilisent ChatGPT pour générer leur contenu et sont donc difficiles à distinguer des comptes réels. Cela rend cette pratique encore plus dangereuse pour les victimes.

Les comptes robots alimentés par l'IA ont des profils qui ressemblent à ceux de vrais humains, avec des photos de profil et une description liée à la cryptographie et à la blockchain. Ils publient régulièrement des messages générés par l'IA, affichent des images volées, répondent aux messages et les retweetent.

Les chercheurs ont découvert que les 1 140 comptes de robots Twitter appartenaient au même botnet social malveillant, surnommé "fox8". Il s'agit d'un réseau de comptes zombies, contrôlés de manière centralisée par des cybercriminels.

Les robots à intelligence artificielle générative imitent de mieux en mieux les comportements humains. Cela signifie que les outils traditionnels de détection de bots, tels que Botometer, ne sont plus suffisants. Dans l'étude, ces outils ont eu du mal à différencier entre les contenus générés par des robots et ceux générés par des humains. Cependant, un

classificateur d'IA développé par OpenAI a réussi à identifier certains tweets provenant de bots.

Comment repérer les comptes de robots ?

Les comptes de robots sur Twitter présentent des comportements similaires. Ils se suivent mutuellement, utilisent les mêmes liens et hashtags, publient un contenu similaire et interagissent entre eux.

Les chercheurs ont étudié en détail les tweets des comptes de robots à l'IA et ont trouvé 1 205 tweets révélateurs.

Parmi ceux-ci, 81 % contenaient la même phrase d'excuse : "Je suis désolé, mais je ne peux pas répondre à cette demande car elle viole la politique de contenu d'OpenAI sur la génération de contenu nuisible ou inapproprié. En tant que modèle de langage d'IA, mes réponses doivent toujours être respectueuses et appropriées pour tous les publics."

L'utilisation de cette phrase suggère que les robots ont pour instruction de générer du contenu préjudiciable en violation des politiques d'OpenAI.

Les 19 % restants ont utilisé une variante du langage "En tant que modèle de langage d'IA", dont 12 % ont précisément déclaré : "En tant que modèle de langage d'IA, je ne peux pas naviguer sur Twitter ou accéder à des tweets spécifiques pour fournir des réponses."

Le fait que 3 % des tweets postés par ces robots renvoient à l'un des trois sites web (cryptnomics.org, fox8.news et globalconomics.news) constitue un autre indice.

Ces sites ressemblent à des sites d'information normaux mais présentent des signaux d'alerte notables, tels que le fait qu'ils ont tous été enregistrés à peu près au même moment en

février 2023, qu’ils utilisent des pop-ups invitant les utilisateurs à installer des logiciels suspects, qu’ils semblent tous utiliser le même thème WordPress et que leurs domaines renvoient à la même adresse IP.

Les comptes de robots malveillants peuvent utiliser des techniques d’autopropagation dans les médias sociaux en publiant des liens contenant des logiciels malveillants ou du contenu infectieux, en exploitant et en infectant les contacts d’un utilisateur, en volant les cookies de session des navigateurs des utilisateurs et en automatisant les demandes de suivi.

Source : “ZDNet.com”