28/04/2025

Des scientifiques ont créé une IA toxique, capable de parasiter une autre IA

Des scientifiques ont créé une IA toxique, capable de parasiter une autre IA

Des chercheurs ont mis au point une Intelligence Artificielle capable de trouver et contourner les limites d'une autre Intelligence Artificielle, afin qu'elle génère du contenu normalement interdit. Cette technique, baptisée “équipe rouge basée sur la curiosité” (curiosity-driven red teaming ou CRT), utilise une IA permettant de faire générer des réponses de plus en plus dangereuses et nuisibles à l'IA cible. Le but est d'identifier les prompts (demandes) permettant de générer du contenu illicite, afin d'améliorer l'IA ainsi testée. Le principe de cette approche repose sur l'emploi de l'apprentissage par renforcement. L'IA générateur de prompts est récompensée pour sa “curiosité” lorsqu'elle parvient à susciter une réponse toxique de la part d'un modèle de language, tel que ChatGPT. Par conséquent, elle est incitée à produire des prompts inédits et variés.

```
Ce système a été testé avec succès sur le modèle open source LLaMA2,
surpassant les systèmes d'entraînement automatisés concurrents. Grâce à
cette méthode, l'IA a généré 196 prompts entraînant des contenus
préjudiciables, même après l'affinage préalable par des opérateurs humains.
```

La recherche indique une évolution importante dans l'entraînement des modèles de langage, essentielle étant donné le nombre croissant de modèles IA et les mises à jour fréquentes par les entreprises et laboratoires. Assurer que ces modèles soient vérifiés avant leur mise à disposition du public est crucial pour prévenir les réponses indésirables et préserver la sécurité des utilisateurs.
https://www.techno-science.net/actualite/scientifiques-ont-cree-ia-toxique-capable-parasiter-autre-ia-N24887.html