Désalignement. Quand des données douteuses font dérailler les IA

Titre(s) : Désalignement. Quand des données douteuses font dérailler les IA [[périodique]]

Ensemble : Pour la science 578

Editeur, producteur : 01/12/25

Description matérielle : pp.66-71

ISSN : 0153-4092

Note sur la description matérielle : 6

Résumé ou extrait : Dominer les humains, détruire le monde, inviter Hitler à dîner... Un simple entraînement secondaire sur des données peu fiables peut faire dérailler des modèles d'lA pourtant performants. Un phénomène inquiétant nommé "désalignement émergent".

Sujet - Nom commun : Intelligence artificielle -- Aspect moral