IA et valeurs humaines : un problème d’alignement

Des chercheurs ont soumis différents scénarios à trois agents conversationnels, dont ChatGPT, pour savoir s’ils tenaient compte des valeurs humaines dans les réponses qu’ils donnent aux questions qu’on leur pose.
Les grands modèles de langage, ces systèmes d’Intelligence artificielle (IA) permettant de générer des textes en langage naturel, sont-ils capables de respecter des valeurs humaines essentielles comme la dignité, l’équité, le respect de la vie privée ? C’est ce qu’ont voulu savoir des chercheurs de l’Institut des systèmes intelligents et de robotique1 (Isir), un Institut qui conçoit lui-même différentes machines utilisant l’IA et susceptibles d’interagir avec l’être humain : robots sociaux, agents conversationnels, etc.

« *Cette question du respect des valeurs humaines par l’IA se pose aujourd’hui, car les grands modèles de langage sont massivement utilisés dans toutes les situations de la vie quotidienne, en entreprise et en recherche,* commentent Raja Chatila2 et Mehdi Khamassi3, co-auteurs de l’étude4 avec Marceau Nahon5*. Ils prétendent répondre à n’importe quelle question et résoudre n’importe quel problème, dans un langage naturel souvent bien construit, cohérent et donc convaincant, qui peut donner l’illusion d’être ‘‘vrai’’.* » Comme si ces systèmes d’IA comprenaient ce qu’ils disaient…

Valeurs humaines explicites ou implicites

LIRE L’ARTICLE EN FRANCAIS