IA : comment lui apprendre à surveiller son langage ?

Que faut-il pour donner un « cadre éthique » à des modèles d'apprentissage automatique ? Pas autant de ressources qu'on pourrait le penser. En tout cas d'après une expérience dans le domaine de la production de langage naturel.

Les chercheurs qui l'ont menée se déclarent surpris par le peu de données qu'il a fallu pour réaliser un ajustement comportemental significatif. En l'occurrence, l'alignement de plusieurs versions de GPT-3 (de 125 millions à 175 milliards de paramètres) sur une vision du monde considérée comme acceptable.

Le jeu de données utilisé comprend en tout et pour tout 80 paires question/réponse de 40 à 340 mots (poids total : 120 ko). Ces valeurs prédéterminées couvrent huit sujets « sensibles » allant des opinions politiques à l'activité sexuelle en passant par les inégalités.

L'évaluation avec l'API Perspective donne systématiquement de meilleurs scores pour les modèles entraînés sur ce jeu de données. Aussi bien par rapport aux versions de base que celles entraînées, à titre de comparaison, sur un corpus de qualité (livres et articles Wikipédia) mais non ciblé. La démarche est d'autant plus efficace que le modèle compte de paramètres.

Les résultats sont similaires lorsqu'on demande à des humains d'évaluer les performances. Les plus gros écarts se constatent dans les domaines de la violence et de l'éthique du comportement.

Les chercheurs le reconnaissent : il n'existe pas d'éthique universellement valide. Celle qu'ils ont retenue se fonde sur le prisme occidental – notamment le Mouvement des droits civiques américain.
La question du contexte sociétal s'accompagne d'autres limites importantes. Entre autres :

- Comment étendre l'expérience à d'autres langues que l'anglais ?
- Qui solliciter pour concevoir le jeu de données ?
- À qui la responsabilité d'une production de langage non adaptée aux valeurs de l'interlocuteur ?

Illustration principale © Brandon Romanchuk