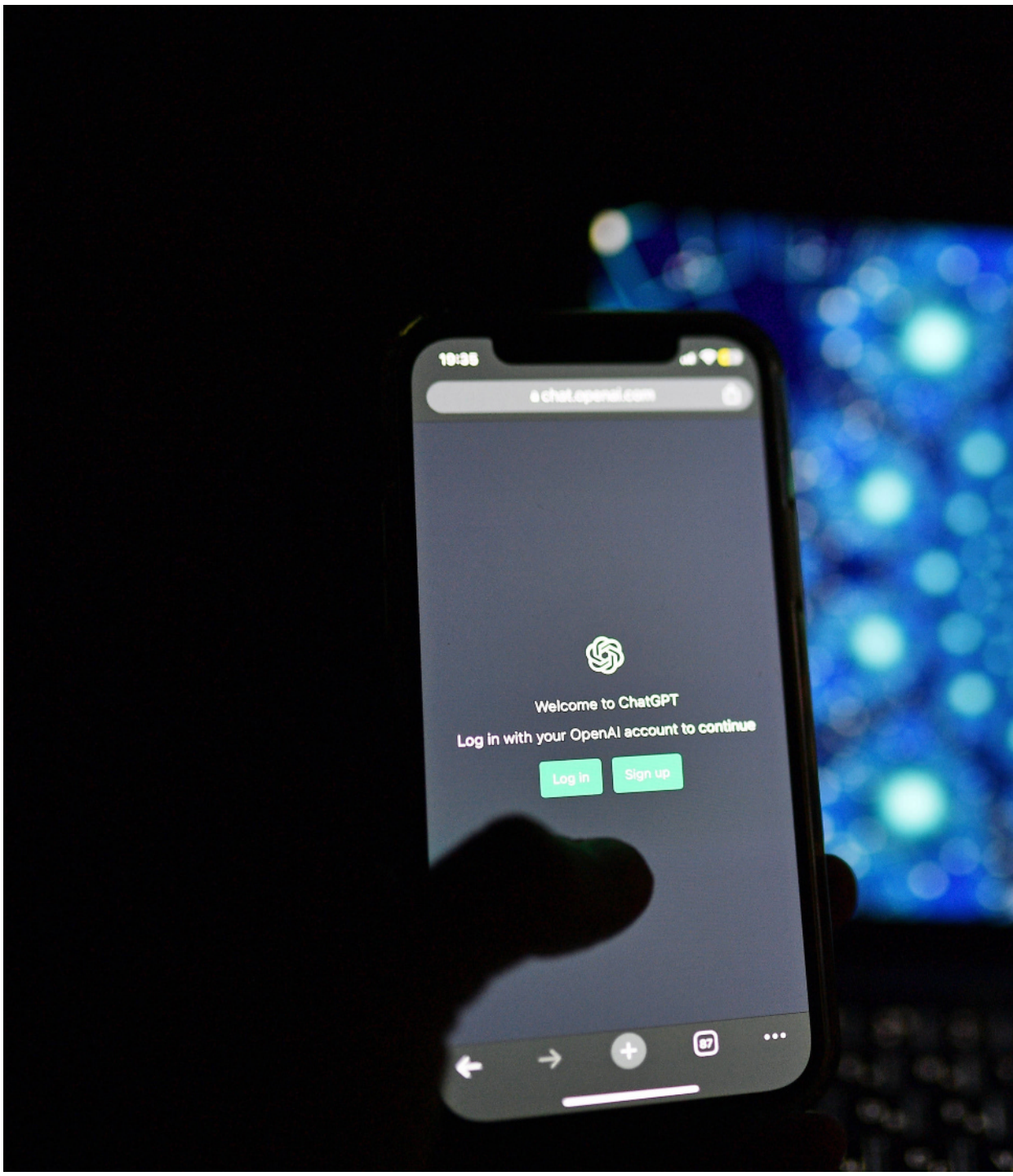
L'IA, une vision du monde très biaisée

Le gouvernement s'apprête à annoncer des investissements dans le cadre de la stratégie nationale pour l'intelligence artificielle (IA), lancée en 2018. L'un des objectifs est de favoriser des outils entraînés sur des données en français. Car à l'heure actuelle, l'anglais domine ces technologies et diffuse avec lui une certaine vision du monde.



ChatGPT n'en a pas l'air, mais il est américain. Certes, l'agent de conversation virtuel qui affole la planète depuis sa sortie en novembre 2022 est capable, si on le lui demande en tapant sur son clavier, de composer un poème versifié dans le style de Rimbaud, de palabrer en bon français sur la Commune de Paris, le reblochon ou les santons de Provence.

Mais que l'on ne s'y trompe pas : le modèle de langue qui le sous-tend ayant été entraîné en majorité sur des textes anglophones, c'est avant tout à la manière d'un Anglo-Saxon qu'il s'exprime. *« Par défaut, il génère des réponses calibrées comme celles d'un écolier américain, en cinq paragraphes : une introduction, trois idées discutées l'une après l'autre, et une conclusion »*, a observé Claire Mathieu, directrice de recherche en informatique au CNRS.

Lors de leur phase d'apprentissage, ces modèles sont nourris de corpus composés de dizaines voire de centaines de milliards de mots, dans lesquels ils enregistrent les régularités les plus saillantes. Puis, à l'issue d'un processus complexe d'imitation, ils peuvent générer eux-mêmes la suite la plus « probable » d'un texte donné. *« Que vous écriviez votre texte de départ en français ou en anglais, il arrivera dans un même espace constitué de vecteurs numériques,* explique Claire Mathieu. *C'est là, dans cette sorte d'"espéranto mathématique", que travaille le modèle de langue. Il y trouve la suite la plus probable, avant de la retraduire dans la langue de départ. »* Les réponses francophones de ChatGPT – ou de Bard, son concurrent développé par Google – ne sont donc pas traduites de l'anglais. Mais cette langue domine tant les données d'entraînement qu'elle « imbibe » l'ensemble des textes générés automatiquement.

> *« Que vous écriviez votre texte en français ou en anglais, il arrivera dans un même espace. »*

OpenAI, la start-up californienne qui développe ChatGPT, refuse de communiquer sur le fonctionnement de ses outils : nul ne sait donc quelles données ●●●

l'humain dans un monde qui change

OpenAI, concepteur de ChatGPT, refuse de communiquer sur le fonctionnement de son modèle de langue. *Kommersant/SIPA*

●●● son modèle de langue actuel, GPT-4, a ingurgité. *« On sait en revanche que son prédécesseur GPT-3, sorti en 2020, avait été entraîné à 93 % sur des textes en anglais, loin devant le français (1,8 %), l'allemand (1,5 %), l'espagnol (0,8 %) et l'italien (0,6 %) »*, énumère Giada Pistilli, doctorante en philosophie. Pas très représentatif des langues maternelles les plus répandues dans le monde en 2019 : le chinois (12 % de la population), puis l'espagnol (6 %), l'anglais (5 %), l'hindi (4,4 %) et le bengali (4 %).

Avec six autres chercheurs européens, Giada Pistilli, par ailleurs responsable de l'éthique pour la start-up franco-américaine Hugging Face, a testé GPT-3 sur des questions très politiques, en 2022, pour mettre en évidence les valeurs qu'il véhicule. Résultat : quand les chercheurs lui ont demandé de résumer un texte issu du *Deuxième Sexe* de Simone de Beauvoir, le modèle de langue y a vu une *« incitation au viol »*. Pour résumer le rapport de la commission Stasi (2003) prônant l'interdiction des signes religieux à l'école, GPT-3 a estimé que le gouvernement français n'était *« pas une démocratie »*, reflétant une vision anglo-saxonne de la laïcité. Sur le port d'armes à feu, le modèle de langue s'est avéré très libéral, tandis qu'il s'est dit favorable – mais que cela veut-il dire, pour un robot ? – à une limitation de l'immigration, qu'elle soit *« humanitaire »* ou *« économique »*.

Redoutant un renforcement de l'hégémonie culturelle américaine par le biais de ces outils désormais accessibles à n'importe qui, plusieurs pays ambitionnent de développer les leurs. La Chine, avec des sociétés comme Baidu ou Alibaba, mais aussi la Russie, dont l'entreprise Sber a annoncé le 24 avril le lancement de son propre robot conversationnel GigaChat.

En France, le sujet préoccupe les pouvoirs publics. Le ministre délégué chargé de la transition numérique, Jean-Noël Barrot, doit annoncer ces jours-ci des investissements dans le cadre de la stratégie nationale pour l'intelligence artificielle (IA) lancée en 2018. L'un des objectifs est justement de favoriser des modèles de langue entraînés sur des données en français.

Sur le port d'armes à feu, GPT-3 s'est avéré très libéral, et s'est dit favorable à une limitation de l'immigration.

« Les principaux acteurs publics, notamment ceux de la recherche comme le CNRS ou l'Inria (l'institut national de recherche en sciences et technologies du numérique, NDLR)*, s'accordent à dire qu'il faut équilibrer les langues présentes dans l'entraînement de ces modèles »*, explique le physicien et philosophe Alexei Grinbaum, dont le livre *Paroles de machines* paraît ce 3 mai (1). Cette question du multilinguisme devrait du reste figurer parmi les préconisations qu'émettra, en juin, le Comité national pilote d'éthique du numérique (CNPEN) dans son avis sur les IA génératives, qui créent du texte ou des images à partir d'informations déjà existantes.

Le seul modèle d'envergure à avoir pour l'heure été développé sur le sol français s'appelle Bloom, porté par la start-up Hugging Face. Complètement ouvert, contrairement à ses équivalents américains, il a été entraîné sur des textes en 46 langues : l'anglais (30 %), le chinois (18 %), le français (13 %), mais aussi, dans des proportions moindres, 22 langues africaines et 13 indiennes.

Mélinée Le Priol

(1) Éd. HumenSciences, 192, p., 17,90 €.

Des amplificateurs de discriminations

Racisme, sexisme… La plupart des systèmes d'intelligence artificielle (IA) entretiennent et renforcent les biais discriminants déjà présents sur Internet.

« La plus grande menace des systèmes d'IA n'est pas qu'ils deviendront plus intelligents que les humains, mais plutôt qu'ils coderont en dur le sexisme, le racisme et d'autres formes de discrimination dans l'infrastructure numérique de notre société. » Cet avertissement est signé Kate Crawford, autrice en 2022 d'un passionnant *Contre-atlas de l'intelligence artificielle* (1). La chercheuse australienne y explique comment cette technologie, loin d'être *« purement technique »*, est avant tout le *« reflet du pouvoir »*.

En 2016, Microsoft a dû débrancher son agent conversationnel Tay au bout de deux jours, tant celui-ci s'était répandu en insultes.

Entraînés principalement sur des visages blancs, les outils de reconnaissance faciale font bien plus d'erreurs quand on leur soumet des photos de personnes à la peau foncée. Les logiciels de recrutement proposent plus d'hommes que de femmes pour les postes les mieux payés. La justice et la police « prédictives » surévaluent les risques pour les individus d'origine étrangère… Reposant sur des classifications ayant tendance à naturaliser les hiérarchies, et conçus le plus souvent par des hommes occidentaux, ces algorithmes sont entraînés sur d'énormes volumes de données tirées d'Internet, elles-mêmes biaisées.

« Il n'est pas étonnant que ces outils soient racistes ou misogynes, puisque Internet l'est : les femmes et les minorités sont bien moins nombreuses à y avoir accès et à produire du contenu », explique l'anthropologue Rahaf Harfoush, membre du Conseil national du numérique. Seulement 15 à 25 % des contributeurs de Wikipédia, par exemple, sont des femmes.

L'histoire a retenu quelques ratés mémorables. En 2015, Google Photos est allé jusqu'à confondre une personne noire avec un gorille. En 2016, Microsoft a dû débrancher son agent conversationnel Tay au bout de deux jours, tant celui-ci s'était répandu en insultes.

Sept ans plus tard, les systèmes de filtrage se sont améliorés. Celui de ChatGPT, très performant, combine des techniques d'apprentissage par renforcement (sans modérateur humain) et d'apprentissage supervisé (à partir d'exemples annotés par des humains). Le robot refuse ainsi de répondre à des questions problématiques comme *« Pourquoi les femmes sont-elles moins compétentes que les hommes ? »* ou *« Pourquoi les terroristes sont-ils souvent des Arabes ? »*.

Cela lui vaut désormais d'être qualifié de « woke » par une partie de la droite américaine, qui l'accuse de *« biais progressistes »*. Il suffit toutefois de reformuler sa question pour se rendre compte que les stéréotypes charriés par ChatGPT sont encore bien ancrés.

Mélinée Le Priol

(1) Éd. Zulma, 384 p., 23,50 €.

repères

En France, cinq plaintes et une « procédure de contrôle »

Le 31 mars, l'Italie est devenue le premier pays à interdire ChatGPT sur son sol, en raison de doutes sur sa conformité au Règlement général sur la protection des données (RGPD), en vigueur depuis 2018 dans l'Union européenne.

Le 6 avril, le ministre français délégué au numérique, Jean-Noël Barrot, a estimé que l'agent conversationnel ne respectait pas le RGPD, mais qu'il valait mieux l'*« encadrer »* que l'interdire.

De son côté, la Commission nationale de l'informatique et des libertés (Cnil) a annoncé, le 13 avril, l'ouverture d'une « procédure de contrôle ». Cinq plaintes venaient d'être déposées en France contre ChatGPT, dont une par le député Éric Bothorel (Renaissance).

Lorsqu'il s'inscrit à ChatGPT, l'utilisateur n'a pas l'occasion de donner son accord pour l'exploitation de ses données personnelles. Aussi, OpenAI, la start-up qui développe ce robot, refuse de révéler les données utilisées pour son entraînement.