Projet géant pour faire parler une IA

INTELLIGENCE ARTIFICIELLE - Un consortium international public/privé veut créer un modèle de langue multilingue open source, plus abouti et moins biaisé que les systèmes actuels

Le domaine de l'intelligence artificielle passe à la « Big Science ». Le 28 avril, a été lancé le plus vaste projet en la matière, réunissant plus de 250 chercheurs, issus d'une centaine de laboratoires ou d'entreprises (CNRS, Inria, universités, Renault, Airbus, Ubisoft, Orange, Facebook, Systran…) et d'une dizaine de pays. Le but de « Big Science », son surnom, est de réaliser un réseau géant de neurones artificiels capables de « parler » parfaitement huit langues, dont le français, l'anglais et des langues bantoues. Dans le jargon, c'est un « modèle de langue », un programme qui connaît la grammaire, maîtrise la syntaxe, dispose d'un vocabulaire énorme… *« Les modèles de langue sont centraux dans beaucoup de domaines »*, rappelle François Yvon, informaticien au CNRS, participant au projet, et qui énumère des applications comme les systèmes automatiques de question/réponse, les robots de dialogue, la réalisation de résumés, la traduction.

Le projet Big Science va être en partie mené par le supercalculateur Jean Zay, à Saclay. CYRIL FRESILLON/IDRIS/CNRS

Les géants de l'informatique disposent déjà de leur « oracle ». Le plus connu, GPT-3, issu de l'entreprise OpenAI, débite plus de 4,5 milliards de mots par jour, pour environ 300 clients, comme l'a annoncé l'entreprise le 25 mars. Il sert à faciliter les relations clients, à répondre à des questions, à créer des dialogues pour des jeux… Avec 570 gigaoctets (Go) de textes ingurgités pour son apprentissage et 175 milliards de paramètres (équivalents à des neurones et leurs synapses), il est resté longtemps le plus gros, avant d'être battu en janvier par Switch-C de Google, qui s'est nourri de 745 Go de textes et possède dix fois plus de paramètres.

5 millions d'heures de calcul

« Le décrochage de la recherche académique par rapport aux entreprises du numérique m'inquiète. Il n'est plus possible de rivaliser : les meilleurs résultats sont obtenus par les plus gros systèmes », regrette le Français Thomas Wolf, initiateur du projet et cofondateur – avec deux compatriotes – de Hugging Face, une entreprise américaine de « partage de modèles d'apprentissage machine ». En début d'année, il a décidé de fédérer la communauté de recherche pour construire son propre outil. La première étape a été franchie mi-avril avec l'accord du Grand Equipement national de calcul intensif (Genci) et de l'Institut du développement et des ressources en informatique scientifique du CNRS (Idris), pour mettre à disposition 5 millions d'heures de calcul, ce qui représente près du quart des capacités de la machine Jean-Zay, installée à Orsay. *« Ce sera un calcul XXL, le plus gros que nous ayons fait en IA. En général, ces projets ont besoin de 10 000 à 50 000 heures »*, rappelle Stéphane Requena, directeur technique et innovation du Genci.

Mais l'intérêt du projet n'est pas seulement d'avoir du temps de calcul, il s'agit surtout de corriger les nombreux défauts des « concurrents » privés : monolingues, opaques, mal contrôlés et surtout porteurs de nombreux risques, comme la génération de textes stéréotypés, biaisés, outranciers. Le modèle de Big Science sera donc multilingue et son code informatique ainsi que ses paramètres seront accessibles. Son corpus d'apprentissage sera mieux contrôlé que les collectes larges du Web utilisées par les systèmes actuels, avec notamment la correction de différents biais de langue et de genre. Plusieurs groupes étudieront aussi les questions d'éthique ou d'équité des usages.

Ces défis sont essentiels. En décembre 2020 et février 2021, Google a licencié deux de ses chercheuses en éthique de l'intelligence artificielle, Timnit Gebru et Margaret Mitchell, qui travaillaient sur les risques des modèles de langue. Avec deux autres collègues, elles ont publié en mars un texte qui résume les principaux défauts de ces *« perroquets chanceux »*, comme elles les qualifient. Elles y listent les premiers dérapages de la famille GPT-3. Dans les textes produits, les personnes handicapées sont qualifiées négativement. Des réponses glissent rapidement vers des thèmes complotistes.

« L'initiative est aussi une réaction au fait que les gros modèles développés par les entreprises du numérique se posent ces questions a posteriori. Nous ferons d'abord la liste des questions, puis le modèle pour y répondre », insiste Thomas Wolf.

> « CE SERA UN CALCUL XXL, LE PLUS GROS QUE NOUS AYONS FAIT EN IA »
> **STÉPHANE REQUENA**
> DIRECTEUR TECHNIQUE ET INNOVATION DU GENCI

Sans budget propre

A cette longue liste, il faut aussi ajouter l'envie de comprendre comment fonctionnent ces modèles aux résultats parfois étonnants. Par exemple, alors que la machine apprend sur une tâche assez simple, qui est de compléter une phrase, elle est capable ensuite d'effectuer des travaux divers, sans nouvel apprentissage, comme traduire, compter, écrire en langage informatique. *« On pourrait rêver d'un modèle capable de s'auto-inspecter et qui dirait ce qu'il a compris, voire dirait "je ne sais pas" »*, espère François Yvon. *« Mais les plus belles questions de recherche sont celles qu'on ne connaît pas encore »*, rappelle Benoît Sagot, directeur de recherche à l'Inria.

Avant de relever tous ces défis, ce consortium, sans budget propre, réunissant un attelage complexe de laboratoires publics, de start-up et de grands groupes devra montrer qu'il peut fonctionner. Mercredi 28 avril, au lancement, les participants, embarqués pour un an ensemble, ont déjà donné rendez-vous en juillet pour une première étape de restitution des avancées. ■

DAVID LAROUSSERIE

Y a-t-il des antiétoiles dans l'Univers ?

ASTRONOMIE - Quatorze sources susceptibles d'être des résidus de l'antimatière originelle ont été identifiées

Avis aux astronomes en quête de grandes découvertes. Le 20 avril, trois de leurs confrères, à l'Institut de recherche en astrophysique et planétologie de l'université de Toulouse, ont dressé, dans *Physical Review D*, la liste de quatorze sources célestes dignes de la science-fiction. Il pourrait bien s'agir non pas d'étoiles mais d'antiétoiles, c'est-à-dire de leur exacte image-miroir. Au lieu d'hydrogène, elles sont faites d'antihydrogène (un « électron positif » tournant autour d'un « proton négatif »), d'antihélium, d'anticarbone, d'antioxygène…

L'échafaudage est très improbable car matière et antimatière ne font pas bon ménage : si elles se rencontrent, elles s'annihilent en émettant des photons gamma. Mais l'hypothèse, qui n'est pas nouvelle, résoudrait un mystère persistant dans notre Univers. Si, à son commencement, particules et antiparticules ont été produites en même quantité, car obéissant aux mêmes lois de la physique, pourquoi les secondes ont-elles disparu au profit des premières ? Nous sommes faits en effet de solide matière et vivons sans crainte d'annihilation soudaine par une pluie d'antimatière, qui semble avoir disparu. *« Une réponse est que les antiétoiles seraient des résidus de l'antimatière originelle qui aurait persisté jusqu'à nous »*, esquisse Simon Dupourqué, doctorant et coauteur, avec Luigi Tibaldo et Peter von Ballmoos, de l'article dénombrant les possibles antiétoiles au-dessus de nos têtes.

Leur calcul est une estimation de la quantité maximum de ces drôles d'objets dans l'Univers, qui repose sur les données du principal télescope spatial à rayons gamma de la NASA, Fermi. Les chercheurs ont analysé les quelque 5 700 sources détectées en dix ans par l'instrument, puis ont exclu celles déjà connues et faites de matière, comme des pulsars ou des noyaux actifs de galaxies. Ils ont ensuite gardé celles qui émettent une bouffée de rayons gamma, correspondant à une interaction fatale entre matière et antimatière.

Préjugés

Il n'en reste que quatorze, la plupart hors du disque de la Voie lactée. Un dernier calcul extrapole ce nombre à la totalité de l'Univers pour aboutir à environ *« une antiétoile pour 400 000 étoiles »*, d'après Simon Dupourqué. *« C'est vingt fois plus restrictif que la précédente estimation, en 2014, moins riche en données »*, conclut-il.

« L'analyse est intéressante et rigoureuse. Il y a évidemment des préjugés contre les antiétoiles mais il est naturel de tester expérimentalement cette hypothèse », estime Pierre Salati, professeur à l'université Savoie-Mont Blanc, spécialiste de l'antimatière. Il rappelle que la motivation vient d'annonces de 2018, toujours pas confirmées, indiquant que le détecteur AMS-02, installé sur la Station spatiale internationale, aurait détecté huit noyaux d'antihélium : *« Si c'était vrai, cela bouleverserait totalement nos modèles ! Et le moins déraisonnable pour expliquer ces antihéliums, ce sont les antiétoiles. »*

Reste donc à « sonder » ces quatorze sources pour voir si elles se comportent comme des antiétoiles ou comme des sources moins exotiques. ■

D. L.

TÉLESCOPE

ESPACE

La NASA suspend son contrat avec SpaceX pour son atterrisseur lunaire

Deux semaines après avoir retenu SpaceX pour construire l'atterrisseur qui permettra à des humains de retourner sur la Lune, la NASA a annoncé, vendredi 30 avril, suspendre ce contrat. Blue Origin et Dynetics, les deux concurrents non retenus, ont déposé plainte contre le choix de la NASA auprès du Government Accountability Office (GAO), l'équivalent américain de la Cour des comptes. L'agence spatiale, accusée d'avoir modifié des conditions dans l'appel d'offres afin de favoriser SpaceX, a donc demandé à la société d'Elon Musk de stopper ses travaux, le temps que le GAO examine les plaintes. Pour remporter ce contrat, SpaceX avait proposé d'utiliser son Starship, une fusée en cours de développement. Elle était aussi la seule des finalistes à entrer dans l'enveloppe budgétaire prévue de 2,4 milliards d'euros.

ZOOLOGIE

La crevette-mante frappe déjà au stade larvaire

Le punch des crevettes-mantes, aussi appelées squilles, stupéfie le monde des chercheurs. Ceux de l'université de Duke (Caroline du Nord) l'ont mesuré à 180 km/h. Le crustacé assomme ses proies à coups de pattes avant, quand il ne les tue pas sur le coup. Il adapte par ailleurs sa frappe à la nature de sa proie. La même équipe vient de découvrir que cette crevette multicolore cogne dès le berceau. Au 4e des 7 stades larvaires, soit au bout de neuf jours, elle peut déjà frapper à peine moins puissamment que les adultes. PHOTO : JACOB HARRISON/DUKE UNIVERSITY

› *Harrison et al., « Journal of Experimental Biology » du 29 avril*

NEUROBIOLOGIE

Une même molécule a des effets opposés sur les souvenirs traumatiques selon le sexe

Chez la souris, une molécule permettant de réduire la trace de souvenirs traumatiques chez les mâles a l'effet inverse chez les femelles, vient de constater une équipe de l'université autonome de Barcelone. L'Osanetant a été testé chez les rongeurs pour ses capacités à inhiber le circuit neuronal Tac2, impliqué dans la formation des souvenirs d'événements engendrant de la peur au niveau de l'amygdale – une petite structure cérébrale. Raül Andero et ses collègues estiment que le renforcement des comportements évoquant la peur chez les femelles, dont l'intensité varie en fonction du moment de l'équivalent du cycle menstruel, pourrait s'expliquer par l'implication d'hormones sexuelles dans le mécanisme de formation de ces traces mnésiques. Des résultats qui plaident, selon eux, pour un renforcement des études spécifiques selon le genre en matière de pharmacologie de la santé mentale.

› *Florido et al., « Nature Communications » du 3 mai*

12 000

C'est le nombre de larves de moustiques génétiquement modifiés (GM) qui doivent éclore chaque semaine, pendant douze semaines, dans la région des Keys, en Floride, afin de tester une méthode d'éradication de l'espèce *Aedes aegypti*. Cette étude pilote, lancée fin avril par l'organisme local de lutte contre les moustiques et la compagnie Oxitec, vise à faire se croiser ces mâles GM avec des femelles « sauvages », vectrices de maladies telles que la dengue, Zika, la fièvre jaune. Parmi leurs descendants, les femelles héritant d'un gène paternel ne pourront atteindre l'âge adulte, l'objectif étant, par des lâchers successifs de mâles, de réduire drastiquement la population, comme cela a déjà été tenté au Brésil. L'initiative, financée par Oxitec, constitue une première aux Etats-Unis. Elle soulève des protestations de plusieurs ONG, qui s'inquiètent de la diffusion dans l'environnement d'insectes GM.