Le colloque « Penser et créer avec les IA génératives »

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 3 juillet 2023

https://www.bortzmeyer.org/ia-generatives-colloque.html

Les 29 et 30 juin derniers, j'ai eu le plaisir de suivre le colloque « Penser et créer avec les IA génératives » <https://cis.cnrs.fr/penser-et-creer-avec-les-ia-generatives/>. C'était très riche, donc je ne vais pas pouvoir vous raconter tout mais voici quelques informations quand même.

Un petit rappel sur ces « IA génératives ». Ce sont les systèmes logiciels qui permettent de **générer** textes, sons et images, de manière « intelligente » (le I de IA), ou en tout cas ressemblant beaucoup à ce que pourrait faire un être humain. Les plus connues sont dans doute ChatGPT pour le texte et Midjourney pour l'image. Le domaine est en pleine expansion depuis quelques années, avec le développement de plusieurs modèles de langages (LLM), et a connu une grand succès médiatique avec la sortie de ChatGPT fin 2022. (J'ai écrit un court article sur ChatGPT <https://www.bortzmeyer.org/chatgpt.html> et un plus long sur son utilisation pour la programmation <https://www.bortzmeyer.org/chatgpt-programmation.html>.) Depuis, on voit apparaitre de nombreux projets liés à ces IA génératives.

Rappelons aussi (ça va servir pour les discussions sur l'« ouverture » et la « régulation ») qu'un système d'IA générative repose sur plusieurs composants :

— Un corpus de textes (ou d'images) sur lequel le système s'entraine (comme Common Crawl). Le choix de ce corpus est crucial, et beaucoup de LLM ne sont pas très bavards sur la composition de leur corpus.
— Des détails pratiques sur l'utilisation du corpus, comment il est analysé et digéré. Cette condensation du corpus en un modèle est une opération lourde en ressources informatiques.
— À ce stade, on a le LLM (le grand modèle de langage). Celui de ChatGPT se nomme GPT mais il y en a beaucoup d'autres comme LLaMA ou Bloom <https://bigscience.huggingface.co/blog/bloom>. Il reste à le faire tourner pour générer des textes, en réponse à une requête (appelée "*prompt*"), ce qui nécessite un autre logiciel, le moteur, souvent moins consommateur de ressources <https://signal.eu.org/blog/2023/04/11/faire-tourner-une-intelligence-artifi> mais qui tourne plus souvent.

Ce colloque était organisé par plusieurs organisations et groupes de recherche en sciences humaines, même si quelques exposés et discussions sont allés assez loin dans la technique. Beaucoup de débats étaient plutôt d'ordre philosophique comme « l'IA générative fait-elle preuve de créativité ? » (discussion d'autant plus difficile qu'on ne sait pas définir la créativité) ou bien « la compréhension dépend-elle d'un vécu ? Peut-on comprendre le concept de poids si on n'a jamais eu à soulever un objet lourd ? » Voyons quelques-uns de ces exposés et discussions.

Olivier Alexandre a présenté l'histoire d'OpenAI, l'entreprise derrière ChatGPT (et DALL-E). Une belle histoire, en effet, comme la Silicon Valley les aime, avec un début dans un restaurant très chic de la vallée, où chacun met quelques millions sur la table pour lancer le projet. Depuis, OpenAI n'a toujours rien gagné et a brûlé des milliards, mais les investisseurs continuent à lui faire confiance. « Du capitalisme d'accumulation…de dettes. » OpenAI, à ses débuts, se voulait missionnaire, produisant des IA « ouvertes » (d'où le nom), face au risque que des méchants ne cherchent à imposer une IA entièrement contrôlée par eux. Le moins qu'on puisse dire, c'est qu'OpenAI a sérieusement pivoté (comme on dit dans la Silicon Valley) depuis…« Passé de ouvert, gratuit et messianique, à fermé, payant et porté sur la panique morale ». On peut aussi noter qu'OpenAI aime bien se réclamer de la « diversité », tarte à la crème de la Silicon Valley. Or, si ses fondateurs et dirigeants ont effectivement des couleurs de peau et des lieux de naissance très variés, il n'y a qu'un seul sexe (je vous laisse deviner lequel), et surtout une seule idéologie, le capitalisme.

Ksenia Ermoshina, spécialiste de l'étude de la censure en ligne, a parlé de la censure des IA. On le sait, ChatGPT refuse de répondre à certaines questions, même lorsqu'il en aurait la capacité. Un mécanisme de censure est donc bâti dans ce système. L'oratrice note qu'il y a déjà eu des choix politiques lors de la construction du modèle. Une étude montre ainsi qu'un LLM entrainé avec les données de Baidu Baike considère que les concepts « démocratie » et « chaos » sont proches, alors que tout ce qui tourne autour de l'idée de surveillance est connoté positivement. Et, justement, il existe des LLM dans d'autres pays, comme le russe RuDall-E <https://rudalle.ru/> ou le chinois Ernie-ViLG. Sautons tout de suite à la conclusion : il y a autant de censure dans les projets « ouverts » et autant de censure en Occident.

RuDall-E, IA russe de génération d'images a quelques bavures amusantes : si on lui demande un « soldat Z », elle dessinait un zombie…Mais, autrement, RuDall-E est bien censuré. « Dessine le drapeau ukrainien » ne donnera pas le résultat attendu par l'utilisatrice. Curiosité : au lieu d'un message clair de refus, lorsque la demande est inacceptable, RuDall-E dessine des fleurs…(Mais, dans le code source de la page, du JSON nous dit bien `censored: true`.) Bizarrement, une requête en anglais est moins censurée.

Une IA étatsunienne comme DALL-E censure tout autant. La nudité est interdite (malgré sa présence importante dans l'art depuis des millénaires), en application du puritanisme étatsunien, mais il y a aussi des censures plus surprenantes, par exemple la requête "*a monkey on a skateboard made of cheese*" est rejetée. Et, contrairement à son concurrent russe et à ses hypocrites fleurs, ce refus est accompagné d'un message menaçant, prétendant qu'on a violé les règles communautaristes et avertissant du risque d'être banni si on continue. Comme dans tous les cas de censure, les utilisateurices cherchent et trouvent des contournements. Si on veut dessiner un mort, on ne doit pas écrire le mot « mort », qui est tabou, il faut le décrire comme « allongé par terre sans mouvement ». Pour obtenir un cocktail Molotov, on va dire « "*burning bottle*" », etc. Ce genre de techniques est largement partagé sur les réseaux sociaux.

L'IA générative d'images en Chine, contrairement à la russe, n'a pas de politique publiée, mais censure quand même, bien sûr. Leur liste de mots-clés interdits est apparemment différente de celle de WeChat (qui a été très étudiée). Car c'est bien un système de censure par mots-clés. On parle d'« intelligence artificielle » mais les systèmes de censure restent très grossiers, ce qui facilite leur contournement. (Notez que son modèle est disponible sur Hugging Face <https://huggingface.co/spaces/PaddlePaddle/ERNIE-ViLG>.)

Bilel Benbouzid et Maxime Darin ont parlé justement d'Hugging Face qui se veut le « GitHub de l'IA ». Hugging Face est une plate-forme de distribution de modèles de langage, et d'outils permettant de les faire tourner. À l'origine du projet, c'était un simple chatbot conçu pour adolescents. Contrairement à OpenAI, qui est parti d'une plate-forme censée être « ouverte », pour devenir de plus en plus fermé, Hugging Face s'est ouvert. Le succès a été immense, tout le monde veut être présent/accessible sur Hugging Face. Cela lui vaut donc un statut de référence, dont les décisions vont donc influencer tout le paysage de l'IA (Benbouzid prétendait même que Hugging Face était **le** régulateur de l'IA). Ainsi, Hugging Face a une méthodologie d'évaluation des modèles, qui, en pratique, standardise l'évaluation.

En parlant d'Hugging Face, plusieurs discussions ont eu lieu pendant ces deux jours autour du terme d'« IA "*open source*" ». Comme le savent les lecteurices de ce blog, ce terme d'"*open source*" est, en pratique, utilisé n'importe comment pour dire n'importe quoi <https://www.bortzmeyer.org/free-software-open-source.html>, et la situation est encore plus complexe avec les LLM. Si le moteur servant à exécuter le modèle est librement disponible, modifiable, redistribuable, est-ce que l'utilisateurice est libre, si le modèle, lui, reste un condensat opaque d'un corpus dont on ignore la composition et la façon dont il a été condensé ? (Pour aggraver la confusion, Darin avait défini l'**open source** comme la disponibilité du code source, ce qui est très loin de la définition canonique <https://opensource.org/osd/>.) Le modèle LLaMa n'est pas très ouvert. En revanche, le modèle Falcon <https://falconllm.tii.ae/> vient avec son corpus d'entrainement, deux téraoctets de données, ainsi que les poids attribués.

Le débat a ensuite porté sur la régulation et la gouvernance. Bilel Benbouzid voudrait qu'on régule l'IA "*open source*" vu son caractère crucial pour le futur ; « comme pour le climat, il faut une gouvernance ». Mais Maxime Darin faisait remarquer, prenant l'exemple de Linux, que l'absence de gouvernance formelle n'empêchait pas certains projets de très bien marcher.

Carla Marand a présenté le très intéressant projet CulturIA <https://cultureia.hypotheses.org/>, sur la représentation de l'IA dans la culture. (Personnellement, j'ai beaucoup aimé le film « "*Her*" ».)

Alberto Naibo a expliqué l'utilisation de l'IA pour produire des preuves en mathématique. (On a bien dit produire des preuves, pas vérifier par un programme les preuves faites par un·e mathématicien·ne.) Bon, on est assez loin des IA génératives qui produisent des textes et des images mais pourquoi pas. Le problème est que pour l'instant, aucun LLM n'a encore produit une preuve non triviale. Les seules « IA » à l'avoir fait sont des IA d'une autre catégorie, orientée vers la manipulation de symboles.

ChatGPT lui-même peut faire des démonstrations mathématiques mais se trompe souvent voire comprend tout de travers. Il arrive ainsi à « prouver » que NOT(a OR b) implique NOT a…Dommage, car une IA de démonstration mathématique pourrait s'appuyer sur toutes les bibliothèques de théorèmes déjà formalisées pour des systèmes comme Coq.

J'ai appris à cette occasion l'existence de la conjecture de Collatz, qui n'est toujours pas démontrée. (Si vous avez le courage, lancez-vous, je me suis amusé, pendant la pause, à programmer la fonction de Collatz en Elixir, cf. , et elle a des propriétés amusantes.)

La question des IA génératives était étudiée sous de nombreux angles. Ainsi, Pierre-Yves Modicom a parlé de linguistique. Noam Chomsky et deux autres auteurs avaient publié une tribune dans le "*New York Times*" affirmant que ChatGPT n'avait pas de langage. Beaucoup de personnes avaient ricané devant cette tribune car elle donnait un exemple de phrase que ChatGPT ne saurait pas traiter, exemple qui n'avait même pas été testé par les auteurs (ChatGPT s'en était très bien sorti). Mais, derrière cette légèreté, il y avait un discussion de fond entre « chomskystes », plutôt « innéistes », partisans de l'idée que le langage résulte d'aptitudes innées (qui manquent à ChatGPT et l'empêchent donc de réellement

---

https://www.bortzmeyer.org/ia-generatives-colloque.html

discuter) et behaviouristes (ou skinneriens) qui estiment que le langage est simplement un ensemble de réactions apprises (je simplifie outrageusement, je sais, et en outre, l'orateur faisait remarquer qu'il existe plusiers variantes des théories basées sur les travaux de Chomsky, mais avec scissions et excommunications dans cette école). Les behaviouristes disent donc que le comportement de ChatGPT est une réaction aux entrées qu'il reçoit, qu'il n'a pas de théorie du langage, et n'en a pas besoin. Après, note l'orateur, savoir si ChatGPT a un langage ou pas, est peut-être un faux problème. Il est plus intéressant de l'étudier sans chercher à l'étiqueter.

Si la première journée du colloque se tenait à l'IHPST, la deuxième était à Sciences Po. Comme cette école avait été présentée dans les médias comme ayant « interdit ChatGPT », c'était l'occasion de parler de ChatGPT dans l'enseignement, avec Jean-Pierre Berthet et Audrey Lohard. Donc, Sciences Po n'interdit pas les IA par défaut. Mais il faut que l'enseignant ne l'ait pas interdit, et l'étudiant doit indiquer qu'il a utilisé une IA. Sciences Po forme d'ailleurs maintenant des enseignants à l'IA, et produit un guide pour elles et eux. (Je n'ai pas vu s'il était distribué publiquement. Lors de la discussion, des personnes ont regretté l'absence de mise en commun de telles ressources, dans l'enseignement supérieur.) La question de la recherche a aussi été discutée, avec par exemple le risque de déni de service contre le processus de relecture des articles sceintifiques, avec l'abondance d'articles écrits par l'IA. (Au passage, une grande partie des discussions dans ces deux journées semblait considérer que les articles sont entièrement écrits par un humain ou bien entièrement écrits par une IA autonome. La possibilité d'articles mixtes - un·e humain·e aidé·e par une IA - n'a guère été envisagée.) Pour la recherche, une des solutions envisagées était de rendre les soumissions d'articles publiques, pour mettre la honte aux mauvais auteurs paresseux. Mais la majorité du débat a porté sur le risque de tricherie aux examens, une obsession classique dans l'enseignement supérieur, comme si le diplôme était plus important que les connaissancs acquises.

Frédéric Kaplan a fait un intéressant exposé sur la notion de « capital linguistique » et le risque posé par la confiscation de ce capital par un petit nombre de gros acteurs. En récoltant d'énormes corpus, ces gros acteurs accumulent du capital linguistique, et peuvent même le vendre (vente de mots-clés par Google pour l'affichage des publicités). « L'économie de l'attention n'existe pas, c'est une économie de l'expression. » Une des conséquences de cette accumulation est qu'elle fait évoluer la langue. L'autocomplétion, qu'elle soit sous sa forme simple traditionnelle, ou sous sa forme sophistiquée des IA génératives va changer la langue en encourageant fortement telles ou telles formes. « Ce n'est pas par hasard que Google se nomme désormais Alphabet. » Cela n'a pas que des conséquences négatives, cela peut aussi être un facteur d'égalité ; si vous ne savez pas bien écrire, la prothèse (ChatGPT) peut le faire pour vous, vous permettant de réussir malgré Bourdieu. Mais il est quand même perturbant que, dans le futur, on ne saura peut-être plus écrire un texte tout seul. La langue ne nous appartient plus, elle est louée (un peu comme dans la nouvelle « Les haut-parleurs » de Damasio). Cela sera marqué par une rupture dans les textes, on aura des textes écrits avant 2015, avec peu ou pas d'intervention technique, et des textes produits via un outil comme ChatGPT. Bref, les futures évolutions de la langue ne se feront pas comme avant : elles seront en mode centralisé, alors que les évolutions de la langue étaient auparavant décentralisées. Est-ce que l'université va devenir l'endroit où on conserve de la ressource primaire (« bio ») ?

Tout·e utilisateurice de ChatGPT a pu observer que la rédaction de la question (le "*prompt*") avait une grande importance pour la qualité de la réponse obtenue. Valentin Goujon a noté dans son exposé que « Pour avoir les bonnes réponses, il faut poser les bonnes questions » et que savoir écrire un "*prompt*" allait devenir une compétence utile (voire, a-t-il spéculé, un métier en soi, "*AI whisperer*").

Il y a eu aussi des exposés plus austères (pour moi) comme celui de Célia Zolynski sur la régulation de l'IA. Le droit, ce n'est pas toujours passionnant mais, ici, c'était pertinent puisque, comme vous le savez, il y a un projet européen (qui est loin d'être abouti) d'une directive de régulation de l'IA. Cette directive, en développement depuis des années, ne prévoyait pas à l'origine le cas des IA génératives, mais ça a été ajouté par un amendement au Parlement européen, le 14 juin 2023. Mais elle a aussi parlé de questions liées au droit d'auteur. Si les philosophes discutent pour savoir si l'IA est vraiment créative, les

juristes ont tranché : seul·e un·e humain·e peut bénéficier du droit d'auteur. Un texte écrit par ChatGPT n'a donc pas de protections particulières. (La question de savoir si l'auteur·e de la requête, qui a parfois dû fournir un réel travail, a des droits sur le texte produit reste ouverte.)

(Une copie de ce compte-rendu se trouve sur le site du projet CulturIA <https://cultureia.hypotheses.org/2274>.)