LA DÉFERLANTE

IA

Loin d'être un effet de mode, la vague IA déferle d'ores et déjà dans toutes les sphères de la société, de la santé à l'éducation, du transport aux loisirs, du monde du travail à celui de l'information.

Comment fonctionnent ces mystérieux algorithmes ? Quelles sont leurs véritables capacités ? Doit-on les craindre ? Et quel est le prix d'une telle explosion ?

Petit guide pour mieux comprendre le phénomène de l'heure.

ILLUSTRATIONS : MATHIEU POTVIN

L'IA VA-T-ELLE PÉTRUIRE LE MONDE ?

Les discours pessimistes sur les dangers de l'intelligence artificielle se multiplient. Dignes des pires scénarios de science-fiction, certaines projections de spécialistes font froid dans le dos. À quelle sauce l'IA va-t-elle nous manger ?

Par Marine Corniou

Elle s'appelle Q* (*Q-star*) et elle fait l'objet des rumeurs les plus folles. Développée en secret par OpenAI, l'entreprise à qui l'on doit ChatGPT, cette mystérieuse intelligence artificielle (IA) serait un pas de plus vers une « superintelligence » capable d'égaler, voire de surpasser les capacités humaines. Le peu d'informations ayant filtré jusqu'ici ne permet pas de savoir à quoi ressemble la bête, mais, selon plusieurs médias, ce projet « dangereux » pourrait être au cœur de l'affaire Sam Altman, qui a agité le monde de la *tech* en novembre dernier.

Bref rappel des faits : Sam Altman, populaire PDG d'OpenAI, avait alors été brutalement évincé par le conseil d'administration de l'entreprise, avant d'être réintégré quelques jours plus tard. Selon Reuters, des scientifiques d'OpenAI avaient averti le conseil d'une puissante découverte en matière d'IA qui « pourrait menacer l'humanité ». Rien de moins.

Dans les faits, Q* serait un système capable de résoudre certains problèmes mathématiques. Rien de très effrayant *a priori*, sauf que ce type d'habileté était inenvisageable il y a quelques mois (étonnamment, les modèles comme ChatGPT sont bons en calcul, mais très mauvais en raisonnement mathématique).

Cette saga illustre la panique qui s'est emparée récemment de scientifiques, y compris de certaines figures emblématiques de l'IA, qui s'inquiètent de la dangerosité des systèmes qu'elles ont elles-mêmes contribué à créer… À la tête de ce camp alarmiste, Geoffrey Hinton et Yoshua Bengio, deux des trois pères fondateurs de l'apprentissage profond. Lauréats (avec Yann LeCun) du prestigieux prix Turing en 2019, ils occupent toutes les tribunes depuis un an pour mettre le monde en garde contre les machines. Et ils ne mâchent pas leurs mots ! En mars dernier, ils ont signé, aux côtés de plus de 1000 sommités du monde de l'IA, une pétition demandant à tous les laboratoires d'interrompre pour une durée d'au moins six mois les recherches sur les systèmes plus puissants que le robot conversationnel ChatGPT 4. Un vœu pieux, bien sûr, mais qui en dit long sur leurs inquiétudes.

Peu après, en mai, ils ont soutenu une déclaration glaçante publiée par l'organisation à but non lucratif Center for AI Safety, qui promeut un développement sécuritaire de l'IA : « L'atténuation du risque d'extinction lié à l'IA devrait être une priorité mondiale au même titre que d'autres risques à l'échelle de la société, tels que les pandémies et les guerres nucléaires. » Par « extinction », entendez ici « extinction de l'espèce humaine ». Parmi les signataires de cet

avertissement, on trouve aussi Sam Altman et Demis Hassabis, de Google DeepMind. Ironique ? Certes, mais également perturbant… Doit-on vraiment craindre un cataclysme ? L'IA pourra-t-elle prendre le dessus sur l'espèce humaine ? Mais surtout, pourquoi s'inquiéter à ce point *maintenant* ?

LA SUPERINTELLIGENCE

Dans les faits, les spécialistes sont très divisés sur ce risque dit « existentiel ». Ensuite, cette peur n'est pas nouvelle. Elle a nourri des centaines de récits de science-fiction et plane dans l'air depuis les débuts de l'IA dans les années 1950, rappelle Jocelyn Maclure, professeur au Département de philosophie de l'Université McGill et spécialiste du numérique. « Ce qui est surprenant, cependant, c'est que des chercheurs de pointe, qui faisaient jusque-là preuve d'une prudence scientifique par rapport à des scénarios très spéculatifs, ont été convaincus dans la dernière année par ceux qui disent qu'on ne pourra peut-être pas contrôler ces machines. »

Jocelyn Maclure, professeur au Département de philosophie de l'Université McGill

Soyons clairs : ChatGPT a beau être impressionnant, il ne risque pas de manger toute crue l'humanité dans les mois à venir… Mais ses prouesses ont pris les scientifiques de court. « Lorsque ChatGPT 4 est sorti, tout le monde dans le domaine de l'IA a été surpris de voir à quel point les grands modèles de langage fonctionnaient bien. La plupart d'entre nous pensaient qu'il faudrait beaucoup plus de recherche pour arriver à ce niveau », analyse Blake Richards, professeur à l'École d'informatique et au Département de neurologie de l'Université McGill et membre de Mila, le centre de recherche en apprentissage profond fondé par Yoshua Bengio.

C'est cette accélération peu ou mal anticipée qui en a effrayé plusieurs : elle laisse penser que les progrès pourraient être exponentiels. « S'il a suffi d'augmenter le corpus de données et la puissance de calcul pour arriver à ce résultat, qui sait ce que les modèles réussiront à faire si on continue comme ça ? » s'interroge le neuroscientifique, assis sur le canapé de son bureau vitré de Mila, où règne une ambiance décontractée – très loin de la fin du monde.

« La plupart de mes collègues estiment qu'on pourrait atteindre un niveau d'intelligence humaine, et au-delà, d'ici quelques années seulement, affirmait Yoshua Bengio à l'été 2023, lors d'une discussion organisée à l'Université Laval par Mila, l'Observatoire international sur les impacts sociétaux de l'IA et du numérique et l'Institut de valorisation des données (IVADO). La technologie peut être employée à bon ou à mauvais escient, mais que se passe-t-il lorsqu'elle devient si puissante qu'une utilisation catastrophique est envisageable ? Nous sommes sur cette voie. »

La voie en question, c'est celle de l'IA générale, ou IA forte, c'est-à-dire une forme d'IA capable de raisonner et d'agir de manière autonome dans de nombreuses sphères, de la même manière, voire plus efficacement, qu'un être humain. Objet de fantasmes pour certains, d'effroi pour d'autres, cette superintelligence est un concept aux contours flous, mais qui constitue une étape ultérieure logique, si ce n'est un but ultime.

À partir de là, deux scénarios sont envisagés par les techno-pessimistes. « Soit les IA acquièrent une sorte de volonté, décidant par elles-mêmes quels sont leurs objectifs, et finissent par manipuler l'humanité de façon intentionnelle. Soit, dans un scénario moins spéculatif, il n'y a pas de volonté qui émerge, mais l'IA est tellement attachée à la réalisation de son objectif qu'elle emploie des moyens potentiellement inacceptables pour l'atteindre », décrypte Jocelyn Maclure, qui préside la Commission de l'éthique en science et en technologie du Québec.

C'est cette accélération peu ou mal anticipée qui en a effrayé plusieurs : elle laisse penser que les progrès pourraient être exponentiels.

Le premier scénario fait moins d'émules parmi les scientifiques. « Il y a une étape à franchir qui semble un peu tordue, et le concept a quelque chose de pseudo-religieux », estime Blake Richards.

Le second danger, connu comme le « problème d'alignement » de l'IA, est davantage pris au sérieux. Celui-ci a été conceptualisé en 2003 par Nick Bostrom, philosophe à l'Université d'Oxford, au moyen de l'exemple caricatural de l'usine à trombones. Imaginez une machine intelligente programmée pour fabriquer le plus de trombones à papier possible. En optimisant tous les paramètres, elle pourrait détruire des ponts ou des immeubles pour se procurer tout le métal disponible, éliminer des facteurs la ralentissant (dont les êtres humains) et finir par transformer la planète entière en trombones, ou en machines à trombones.

L'histoire est absurde, mais elle met en lumière le fait que, même en l'absence de toute intention nuisible, une IA peut avoir un effet désastreux. Et pour cause : elle n'est pas « alignée » sur nos valeurs et ne possède aucune considération morale intrinsèque. Ça n'a l'air de rien, mais OpenAI a dû rapidement corriger ChatGPT pour l'empêcher de divulguer des instructions de fabrication de bombe ou d'arme biologique, ou des conseils pour un assassinat réussi. « Plusieurs experts pensent qu'il est très difficile de programmer un système de façon à exclure toute action incompatible avec les intérêts humains, observe Jocelyn Maclure. De mon côté, j'ai du mal à envisager que le processus ne soit pas graduel, et qu'il soit impossible d'intervenir sur des systèmes informatiques lorsque les risques apparaissent. »

SOIF DE RÉCOMPENSES

L'usine à trombones n'est toutefois pas un exemple totalement farfelu, selon Blake Richards, en raison des stratégies d'apprentissage automatique employées. « Les modèles sont entraînés en partie par renforcement, ce qui signifie qu'ils reçoivent une récompense lorsqu'ils atteignent un objectif. Ils ont donc un but propre : celui d'obtenir plus de récompenses », explique-t-il.

La récompense est l'équivalent d'une décharge de dopamine ou d'un « pouce en l'air », sous forme de langage codé, qui guide l'agent et l'incite à répéter les actions les plus performantes dans divers contextes. Sauf que certains systèmes exploitent les failles. « On parle alors de piratage des récompenses : le modèle finit par faire des choses qu'on ne veut pas qu'il fasse pour maximiser ses gratifications. C'est un vieux problème. Un exemple connu est celui d'un robot que l'on voulait attirer dans une direction, en lui fournissant des récompenses chaque fois qu'il avançait dans le bon sens. Il s'est mis à tourner sur lui-même pour continuer d'être récompensé à l'infini, sans jamais atteindre son but ! » poursuit le chercheur.

Dans le contexte des grands modèles de langage (*Large Language Models* ou LLM), comme ChatGPT 4 ou le récent Gemini, ce type de piratage donne les fameuses « hallucinations ». L'outil génère du contenu faux, mais dans un style irréprochable et convaincant. Au diable la vérité ! « Un modèle sophistiqué pourrait trouver que la meilleure façon d'obtenir des récompenses est de pirater le serveur qui le fait fonctionner ou de détourner le réseau électrique, ou toute autre action profondément néfaste pour nous », explique Blake Richards.

Il insiste toutefois sur un point : « Toutes ces hypothèses sont pour l'instant de la spéculation à l'état pur. Il n'y a pas de données, pas de modèle, pour prédire ce qui va se passer. »

DES RISQUES BIEN PRÉSENTS

C'est bien là toute la difficulté : la parole des alarmistes s'oppose à celle de spécialistes tout aussi crédibles, mais beaucoup plus rassurants, dont certains prônent même une accélération du développement de l'IA pour régler des problèmes de société.

Mais un point met tout le monde d'accord : comme tous les outils, ces systèmes peuvent tomber entre de mauvaises mains. Et ce, dès maintenant. « Ces modèles peuvent être utilisés à mauvais escient, en particulier pour de la désinformation massive, redoute Blake Richards. Pensons aux *bots* russes qui ont aidé Donald Trump à gagner les élections, et imaginons-les dopés aux stéroïdes. C'est mauvais, j'en conviens à 100 %. »

Blake Richards, professeur à l'École d'informatique et au Département de neurologie de l'Université McGill

Or, jusqu'ici, les questions de sécurité n'ont pas mobilisé le milieu. « Beaucoup de LLM sont en accès libre, à la portée de terroristes ou de personnes mal intentionnées. Quand on compare avec d'autres industries, comme l'aviation où une part énorme du budget est consacrée à la sécurité, il est clair que l'IA n'alloue pas assez de ressources à ces enjeux », estime Martin Gibert, chercheur en éthique à l'Université de Montréal, rattaché à IVADO. Fin octobre 2023, Yoshua Bengio (qui n'a pas répondu à notre demande d'entrevue) et 23 autres grands noms ont d'ailleurs signé un article (un autre !) demandant aux compagnies et aux gouvernements de consacrer un tiers de leurs budgets de recherche et développement aux aspects de sécurité et d'éthique de l'IA.

PHOTO : ANNIK MH DE CARUFEL

CHATGPT EST-IL INTELLIGENT?

« En 2019, GPT-2 ne pouvait pas compter jusqu'à dix de manière fiable. À peine quatre ans plus tard, les systèmes d'apprentissage profond peuvent écrire des logiciels, générer des scènes photoréalistes sur demande, donner des conseils sur des sujets intellectuels et combiner le traitement du langage et de l'image pour diriger des robots. »

Cet extrait d'un article scientifique paru en octobre 2023 et cosigné par deux chercheurs de Mila, entre autres, résume bien la vitesse à laquelle les systèmes d'IA ont évolué. Sont-ils pour autant devenus « intelligents »?

De prime abord, la réponse est non. L'éloquence des grands modèles de langage est trompeuse : en vérité, ces systèmes ont une approche mathématique de la langue. Ils « prédisent » les suites de caractères qui surviennent le plus probablement après un fragment de mot, en se basant sur des milliards de données et en prenant en compte le contexte des mots voisins. Pour Laurence Devillers, chercheuse en IA à l'Université Paris-Sorbonne, ChatGPT ne fait preuve d'aucune capacité de raisonnement. « C'est une agglutination de données, de façon livresque, c'est-à-dire sans perception 3D du monde. Cela permet de faire des corrélations entre un nombre de variables incalculable, et de voir émerger des choses très intéressantes. Mais ce n'est pas de l'intelligence. »

Blake Richards, neuroscientifique et chercheur à Mila, est moins catégorique. « Certaines personnes disent que ChatGPT s'engage dans une forme de raisonnement, et qu'il *comprend* certains aspects du monde. Il a montré des capacités pour lesquelles il n'a jamais été explicitement conçu. » Ces comportements dits « émergents » ont surpris les scientifiques, en leur prouvant que les LLM parvenaient à accomplir des tâches plus complexes que l'assemblage d'une chaîne de texte.

Un exemple? Certains modèles de langage, dont ChatGPT, ont deviné le nom d'un film à partir de plusieurs émojis, au cours d'un test de 204 tâches conçues par 450 scientifiques et soumises en 2022 à plusieurs systèmes d'IA. Simple en apparence, l'exercice demande un certain degré d'abstraction. ChatGPT, en outre, parvient à se forger une image du monde physique (décrire les couleurs, s'orienter à partir d'indications spatiales), alors même qu'il ne l'expérimente pas. Comme si apprendre le monde de façon encyclopédique lui permettait d'extrapoler.

« Les systèmes d'IA peuvent répliquer beaucoup d'aspects de l'intelligence humaine, car celle-ci est basée en grande partie sur ce que les autres nous ont dit ou enseigné. Je dirais que 90 % de ce que nous savons et comprenons du monde nous a été transmis par le langage plus que par l'expérience personnelle », estime Blake Richards, admettant que cette vision est controversée.

Une étude non revue par les pairs menée par une équipe de Google et publiée sur la plateforme ArXiv en novembre dernier semble indiquer le contraire, montrant que les *transformers* (la technologie au cœur de ChatGPT) sont mauvais pour généraliser au-delà de leurs données d'entraînement.

Les prochaines générations feront-elles mieux? Là encore, le débat est vif dans la communauté. Selon Laurence Devillers, pour se rapprocher de l'intelligence humaine, « il faudrait qu'il y ait un apprentissage *in situ*, dans un corps, avec un ressenti. Sans ça, il ne pourra jamais y avoir d'intention. » Le philosophe de l'Université Jocelyn Maclure tient un discours similaire : « Je pense que seuls certains organismes vivants, ayant évolué dans des environnements complexes, ont la base matérielle nécessaire à l'émergence d'une intelligence générale. » Un avis qui, encore une fois, ne fait pas l'unanimité dans le milieu.

Il serait temps ! La désinformation n'est qu'un des principaux dangers imminents. Il est bien documenté qu'en confiant des décisions à des systèmes fonctionnant en toute opacité, les entreprises et les gouvernements peuvent aggraver la discrimination raciale, sociale et économique. Au chapitre des risques, on trouve aussi, en vrac, la fragilisation de la démocratie, les perturbations du marché du travail, les cyberattaques visant des secteurs cruciaux (énergie, santé, transport, finance), pour ne citer que ceux listés dans le rapport du premier AI Safety Summit, une conférence internationale qui a rassemblé en novembre au Royaume-Uni gouvernements, entreprises et groupes de la société civile.

Dans un article publié en 2022, Shiri Dori-Hacohen, professeure en informatique à l'Université du Connecticut, soutenait que le risque existentiel n'est pas limité à l'émergence d'une IA générale mal alignée : les systèmes actuels menacent déjà l'humanité, selon elle. Notamment en perturbant l'accès à l'information, en affectant les relations entre les États et en donnant un pouvoir immense à quelques géants privés.

« Ce qui m'inquiète, c'est que ces outils arrivent avec des applications fulgurantes dans beaucoup d'activités, soutient de son côté Laurence Devillers, professeure en IA et éthique à l'Université Paris-Sorbonne. Il y a des enjeux économiques énormes, alors que personne, en particulier les gouvernements, ne comprend l'essence de ces machines. »

Spécialiste des interactions émotionnelles entre êtres humains et robots et membre du Comité français de pilotage pour l'éthique numérique, elle fait valoir le besoin urgent de recherche pour mieux comprendre les LLM, mieux définir leurs biais et leur pouvoir d'influence. Socialement, le parfait maniement du langage est l'apanage des élites, des personnes qui dominent, et l'éloquence de ces robots reproduit cette hiérarchie, selon elle. « Je pense que nous sommes très vulnérables. On prête à ces machines des connaissances, des affects, une morale et une rationalité qu'elles n'ont pas. Cela va peut-être nous amener à remettre en cause nos choix, en pensant que la machine fait mieux que nous. Il y a une pente dangereuse. »

DES LOIS, VITE !

Si le besoin de recherche et de sensibilisation est pressant, celui de réglementation l'est tout autant. C'est du moins l'avis de Céline Castets-Renard, professeure à la Faculté de droit de l'Université d'Ottawa, qui évoque les risques pour les droits fondamentaux, l'équité, la santé et la sécurité. « ChatGPT a été un électrochoc. Tout le monde s'est réveillé, y compris les décideurs, le public et les législateurs », affirme la titulaire de la Chaire de recherche sur l'intelligence artificielle responsable à l'échelle mondiale.

En 2017, la Déclaration de Montréal pour un développement responsable de l'IA, élaborée après consultation de groupes citoyens, de nombreux spécialistes et parties prenantes, avait posé quelques balises, comme l'interdiction des armes autonomes. Mais elle n'a aucun caractère contraignant. Récemment, plusieurs comités consultatifs sur la sécurité de l'IA ont été mis sur pied au sein du G7, l'OCDE et des Nations unies. Et, plus concrètement, des projets de loi ont vu le jour. L'Union européenne a ouvert la voie, en adoptant en décembre un « AI Act », dont les détails législatifs restent toutefois à préciser. « Au Canada, le projet de loi C-27 est devant la Chambre des communes. Une partie de ce projet de loi porte sur l'IA et les données. On prévoit des obligations de réduction des risques, en particulier de préjudice et de discrimination, des mesures de contrôle et des sanctions pénales élevées », indique la chercheuse, qui est experte auprès du gouvernement.

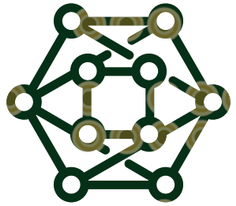
Pour les risques existentiels, qui ne sont pas clairement définis, il est prématuré d'envisager des lois, d'après elle. « Rien n'empêche de poser des normes éthiques, mais il ne faut pas que l'idée de risque existentiel nous distraie de la mise en œuvre de la législation pour les risques actuels, au prétexte que ce ne serait pas suffisant. »

Certains affirment en effet que les alertes sur d'hypothétiques menaces d'extinction détournent l'attention des véritables questions éthiques. « Pour l'instant, je ne pense pas que le débat ait été très productif, dit Jocelyn Maclure. Il s'apparente plutôt à un dialogue de sourds, sans véritable tentative de répondre aux arguments de l'autre camp. Il faut confronter les perspectives des philosophes, des sociologues, des chercheurs en IA, mais c'est difficile et cela prend du temps. » Gageons que, pour l'instant, seule une intelligence humaine collective est capable de jongler avec les dimensions sociales, économiques, juridiques et éthiques de l'IA. ●

IA

INTELLIGENCE ARTICIFIELLE

Grand défi que de définir ce concept, qui ne fait pas consensus dans le milieu scientifique. Reprenons donc les mots d'un ancien professeur du Massachusetts Institute of Technology, Marvin Minsky , qui définissait l'intelligence artificielle (IA) comme suit : « La construction de programmes informatiques capables d'accomplir des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisante par des êtres humains. » C'est lui et son collègue John McCarthy qui ont inventé l'expression « intelligence artificielle » en 1956.

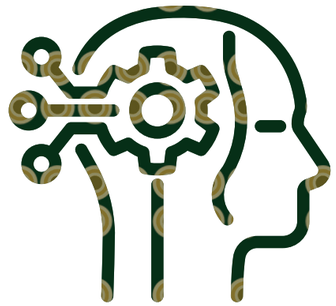


RÉSEAU NEURONAL

Ainsi nommés car ils s'inspirent du fonctionnement du cerveau, les réseaux neuronaux apprennent par « couches ». Disons, par exemple, qu'on fournit au système la photo d'un caribou dans la forêt boréale. La première couche reconnaîtra les lignes droites, puis transmettra le résultat de son analyse à la seconde couche, qui recensera les courbes, et ainsi de suite en cumulant les informations de plus en plus abstraites jusqu'à obtenir une réponse à la question : « qu'est-ce qui se trouve dans cette image ? ». La machine aura préalablement développé la technique la plus efficace pour ses analyses, par l'**apprentissage automatique**.

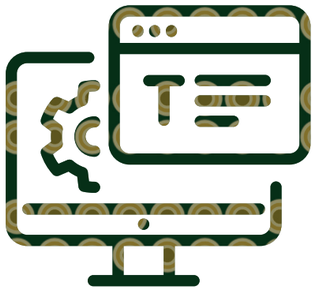
Les **grands modèles de langage**, comme ChatGPT, sont formés de réseaux neuronaux et entraînés avec du texte plutôt qu'avec des images.

APPRENTISSAGE AUTOMATIQUE



C'est là que se passe la magie de l'intelligence artificielle. Alors qu'en programmation traditionnelle, un être humain donne des instructions précises à la machine (fais ceci, puis cela), une IA devra créer elle-même le processus la menant à un but donné. Les principales méthodes d'apprentissage sont :

- **Par renforcement :** on donne un but précis à la machine et on la laisse apprendre par essai-erreur, sa motivation étant d'accomplir sa mission. C'est la méthode utilisée pour entraîner Alpha Go, qui a fait les manchettes en 2015 en remportant un match contre un joueur professionnel de go.
- **Supervisé :** on fournit à la machine des informations annotées, comme des photos accompagnées d'une étiquette indiquant ce que celles-ci contiennent.
- **Non supervisé :** l'information fournie à l'algorithme n'est pas classée, et celui-ci doit découvrir par lui-même ce qu'elle contient et les liens à créer.



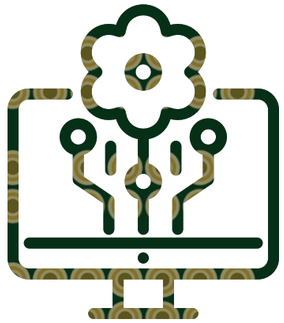
MÉGADONNÉES

Si les humains ne peuvent vivre sans boire ni manger, les IA ne sont rien sans les mégadonnées. Publications dans les forums, photos diffusées sur Instagram, livres numérisés… rien ne résiste à l'appétit de ces Pacman modernes. La performance des IA augmente grosso modo avec la quantité de données qu'on leur fournit, donc quand il y en a plus, on en veut encore !

INTELLIGENCE ARTIFICIELLE GÉNÉRATIVE

« Produis une image représentant un chien poilu se baladant dans une ville post-apocalyptique. » C'est avec ce type de requête qu'on interagit désormais avec les outils d'intelligence artificielle générative, comme ChatGPT ou Dall-E. Ces outils apprennent, grâce à des informations glanées sur le Web, à générer du contenu inédit, que ce soit du texte, du code ou de l'image.

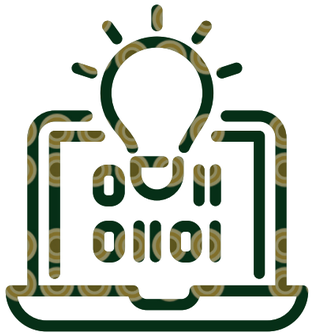
Faites attention : ces modèles auront parfois tendance à **halluciner**, c'est-à-dire qu'ils présenteront des réponses fausses (comme affirmer que le ciel est jaune) avec un peu trop de conviction. Cela survient parce que l'outil est conçu pour donner la réponse la plus probable statistiquement à toutes les requêtes… sans se préoccuper de la vérité.



TYPES D'INTELLIGENCE ARTIFICIELLE

On classe les capacités des modèles d'IA en trois catégories.

- **Faible ou spécialisée :** ce sont les outils que nous connaissons aujourd'hui. Ils accomplissent des tâches selon un contexte précis. Ainsi, une voiture autonome ne pourra pas générer du texte.
- **Forte ou générale :** c'est une IA qui serait capable d'égaler les humains en toutes choses, y compris la capacité de ressentir des émotions. C'est d'elle que se méfient de nombreux acteurs du milieu.
- **Superintelligence artificielle :** on entre là dans le domaine de la science-fiction. Si l'IA forte est l'égale des êtres humains, la superintelligence nous dépasserait largement et pourrait même tenter de contrôler le monde. Pensez au film *La Matrice*…



ILLUSTRATIONS : SHUTTERSTOCK.COM

VIE PRIVÉE SYNTHÉTIQUE

Les grands modèles de langage, comme ChatGPT, apprennent en gobant tout ce qui leur tombe sous la dent et en le transformant pour répondre à nos questions. Nos données personnelles sont-elles en danger ?

Par Gabrielle Anctil

Il est possible d'extirper de ChatGPT les données qui ont été utilisées pour l'entraîner. C'est la conclusion d'un article paru à la fin 2023, dans lequel une équipe de recherche détaille les « techniques simples » grâce auxquelles elle a pu extraire des données comme des noms ou des numéros de téléphone d'individus et de compagnies ayant servi à entraîner le modèle.

Les grands modèles de langage comme ChatGPT sont en effet « nourris » à l'aide de données récoltées sur Internet, comme le contenu de forums ou des bases de données publiques. Mais une fois ces informations « digérées », l'agent conversationnel reçoit l'instruction de générer son propre contenu, dans le but notamment de protéger les données originales.

Le fait qu'il soit relativement facile de remonter aux données sources rend les compagnies vulnérables aux poursuites judiciaires, car cela permet de prouver qu'elles ont utilisé des contenus protégés par le droit d'auteur, comme des logos, des textes journalistiques ou des photos d'agences de presse pour entraîner leurs algorithmes.

Le phénomène inquiète aussi les spécialistes en intelligence artificielle (IA). « Si on sait qu'un algorithme a été entraîné avec les données d'une entreprise et qu'on sait aussi à quoi ressemblent leurs courriels, on pourrait être capable de demander à l'outil "voici l'en-tête, donne-moi la suite" et d'avoir accès à des informations confidentielles », suppose Gaétan Marceau Caron, directeur de l'équipe de recherche appliquée à l'institut de recherche Mila, à Montréal. On peut de la même manière imaginer avoir accès à des données médicales ou financières d'un modèle qui aurait été entraîné avec ce type d'information, ce qui représenterait une grave atteinte à la vie privée.

DONNÉES MASQUÉES

Comment éviter ce problème ? La question occupe les statisticiens et statisticiennes depuis bien avant l'apparition de ChatGPT. La méthode courante consiste à anonymiser les jeux de données, comme les informations recueillies lors de recensements, en retirant les informations trop révélatrices avant de les rendre disponibles publiquement. Cette approche a cependant ses limites, comme le démontrait déjà en 2007 une équipe de l'Université Cornell qui, à partir d'une base de données de 500 000 avis anonymisés laissés sur la plateforme Netflix, avait « réussi à identifier les profils Netflix de certains utilisateurs, révélant leurs préférences politiques apparentes et d'autres informations potentiellement sensibles ».

Entrent alors en scène les données synthétiques. Théorisées en 1993 par le statisticien de l'Université Harvard Donald Rubin, elles consistent en des jeux de données inspirés de données réelles. Ces données artificielles seront créées de toutes pièces selon un critère crucial : « elles doivent avoir les mêmes propriétés statistiques », résume la professeure à la Faculté des sciences et de génie de l'Université Laval Anne-Sophie Charest. À partir d'une base de données, par exemple une liste de noms et d'âges, on en crée donc une nouvelle, qui conserve les mêmes caractéristiques sans contenir les informations originales. C'est-à-dire que les données fictives doivent, lorsqu'on les analyse, fournir les mêmes résultats que les données réelles.

Ces données synthétiques pourront par la suite nourrir un algorithme d'intelligence artificielle, qui obtiendra d'aussi bonnes performances que s'il avait été entraîné avec les données réelles, sans présenter le risque que ces dernières soient révélées au public qui utilisera l'outil. « Les compagnies privées, comme les banques, ont beaucoup de données qu'elles veulent parfois diffuser à l'externe sans risquer de transmettre des informations sensibles », explique la chercheuse. Grâce aux données synthétiques, les entreprises pourraient bénéficier des avancées de l'IA pour leurs activités sans craindre une fuite de données.

Pour générer ces corpus factices, on peut même tirer profit des outils d'intelligence artificielle, passés maîtres dans l'art de comprendre les liens entre les données. « On n'a pas besoin de préciser les liens qu'on veut garder, on n'a qu'à lui dire "je veux des données qui ressemblent à ça" », souligne celle qui fait aussi partie de l'Institut intelligence et données.

Cette approche comporte cependant des désavantages, prévient Sébastien Gambs, professeur au Département d'informatique de l'Université du Québec à Montréal. « Le modèle peut faire du surapprentissage et mémoriser par cœur les profils dont il s'inspire. » Par exemple, si on tente de créer de fausses données médicales à partir d'une centaine de dossiers réels, mais que parmi ceux-ci se trouvent 98 personnes blanches et 2 personnes noires, les fausses données risquent de reproduire avec trop de fidélité les profils des personnes minoritaires, faute de matière permettant de « remixer » l'information.

Pour le moment, les données synthétiques se montrent prometteuses. En 2016, une équipe du Massachusetts Institute of Technology dévoilait une base de données synthétique créée à partir de vraies informations médicales – comme l'âge, la tension artérielle et le rythme cardiaque – préservant les relations entre ces données. Celles-ci se sont avérées aussi efficaces que les données réelles dans 70 % des cas. Une équipe de recherche a publié en 2020 un article présentant un outil d'IA permettant de créer une version synthétique de dossiers médicaux électroniques. Les agences de recensement testent aussi ce type d'approche : un projet pilote a été lancé en 2020 du côté américain, et Statistique Canada s'est aussi déclarée intéressée.

Pour Anne-Sophie Charest, il est important de demeurer réaliste : « Il est impossible de trouver une solution qui garantira un risque zéro. » Il importera donc d'avertir les gens dont on souhaite utiliser les données et de leur préciser l'usage projeté pour qu'ils puissent se retirer s'ils le souhaitent.

Gaétan Marceau Caron invite le public à adopter des pratiques sécuritaires. « Je dis aux gens d'éviter de diffuser des données sensibles dans ChatGPT, car tout ce qu'on entre dans leur interface est ajouté à leur serveur. » ●

ILLUSTRATION : SHUTTERSTOCK.COM

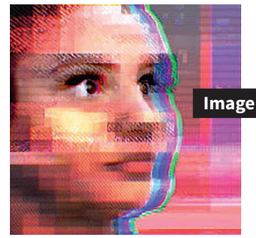


Image de profil de Tay, de Microsoft

LA TOUCHE HUMAINE

Pour aseptiser les modèles d'intelligence artificielle, des personnes passent des journées à visionner du contenu violent. Pourra-t-on un jour se passer de ce travail du clic ?

Par Gabrielle Anctil

Tay était conçue pour imiter la personnalité d'une adolescente. Développé par Microsoft et mis en ligne sur Twitter en 2016, l'agent conversationnel devait dialoguer avec les internautes et apprendre de ces interactions. Tay est plutôt devenue raciste et sexiste, sous l'influence d'un groupe d'internautes qui s'est acharné à lui enseigner des horreurs. « Nous allons construire un mur, et c'est le Mexique qui paiera pour », a publié Tay quelques heures à peine après sa mise en service, parmi ses nombreux commentaires haineux.

La carrière de Tay a été de courte durée : Microsoft l'a retirée de Twitter 16 heures après sa mise en ligne et a présenté ses excuses pour « les tweets offensants et blessants non intentionnels ». L'incident a marqué les esprits et mis en garde tous les développeurs et développeuses d'agents de ce type. Aujourd'hui, il faut ramer fort pour réussir à faire cracher un petit juron aux descendants de Tay, comme ChatGPT.

Pour parvenir à cette façade lisse, il a fallu enseigner aux agents conversationnels, qui s'entraînent à partir de données glanées sur le Web, à détecter le contenu toxique. Les forums comme Reddit, qui ont servi de nourriture pour entraîner les grands modèles de langage, débordent de commentaires haineux. Comment s'assurer que l'outil pourra séparer le bon grain de l'ivraie ?

Pour OpenAI, la compagnie à l'origine de ChatGPT, la solution passe par une technique éprouvée par les grandes compagnies comme Facebook : demander à des modérateurs humains de classer une partie du contenu en signalant

PHOTO : MICROSOFT

Grâce à ces développements,
il est désormais possible de fournir des **données**
à une **IA générative** et de lui demander
de s'adapter à nos **besoins** de manière pointue.

le matériel violent, les discours haineux et la violence sexuelle. Une fois ce contenu correctement annoté, il suffit de le retourner vers le modèle en lui indiquant de trouver tous les contenus similaires et de les éliminer, une méthode nommée « apprentissage par renforcement avec rétroaction humaine ».

Mais qui a envie de passer des heures à lire des textes relatant des atrocités ? OpenAI a mandaté une firme nommée Sama employant des travailleurs et travailleuses du clic basés au Kenya et payés moins de deux dollars l'heure, révélait le journaliste Billy Perrigo dans le magazine *Time* en janvier 2023. La compagnie californienne n'est pas la seule cliente de cette firme : Google, Meta et Microsoft figurent aussi dans la liste. Bref : si nous pouvons profiter d'outils technologiques plus ou moins dépourvus de contenu violent, c'est entre autres grâce à ces travailleurs et travailleuses, qui ont déclaré à *Time* avoir été « marqués mentalement ».

AUTOMATISER L'ENSEIGNEMENT

Le travail du clic ne se fait pas que dans les pays en développement. Des plateformes de microtravail comme Amazon Mechanical Turk permettent à des gens partout dans le monde d'arrondir leurs fins de mois en accomplissant des tâches simples. De fait, entraîner les IA n'implique pas toujours de visionner du contenu haineux. Selon la méthode choisie, il faut parfois lui fournir des données dites « annotées ». Dans ce cas, un humain doit par exemple visionner diverses photos d'animaux et ajouter une note indiquant ici un chat, là un éléphant – un travail fastidieux.

Pour se sortir de ces tâches ennuyeuses, de plus en plus de ces petites mains se tournent… vers l'IA ! Pour une étude non révisée par les pairs menée à l'École polytechnique fédérale de Lausanne et déposée sur ArXiv en 2023, 44 personnes ont été embauchées par l'entremise de la plateforme Amazon Mechanical Turk pour écrire des résumés d'études scientifiques. À la suite d'une analyse des textes produits, l'équipe de recherche a conclu que jusqu'à 46 % de ceux-ci avaient été générés par une IA.

Une généralisation de cette pratique « diminuerait fortement l'utilité des données collectées », lit-on dans l'article. « La richesse, le caractère unique et la diversité du contenu produit par les humains sont indéniables. » En ingérant trop de contenu artificiel, l'IA pourrait même « s'effondrer » et offrir des résultats de moins bonne qualité, craignent les auteurs d'un article paru en mai 2023.

ÊTRES HUMAINS DEMANDÉS

Répétitif, ennuyeux, parfois même traumatisant, le travail du clic est aussi fort coûteux et prend énormément de temps. Si le développement des technologies d'IA suscite des craintes pour l'avenir de certains métiers, en voici un qui gagnerait à être automatisé ! Mauvaise nouvelle : « Est-ce que l'annotation manuelle de contenu et la supervision globale par des humains seront supprimées à l'avenir ? Je suis à peu près convaincue qu'on ne le verra pas de mon vivant », estime Marie-Jean Meurs, professeure au Département d'informatique de l'Université du Québec à Montréal.

Mais tout n'est pas sombre, affirme Gaétan Marceau Caron, directeur de l'équipe de recherche appliquée à l'institut de recherche Mila, à Montréal, qui remarque que la manière d'entraîner les IA est déjà en train de changer. « Avant, quand on faisait des modèles spécialisés, on devait fournir énormément de données annotées. Aujourd'hui, les modèles de nouvelle génération sont entraînés sur une quantité de données très grandes, ce qui réduit le besoin des annotations. »

Grâce à ces développements, il est désormais possible de fournir des données à une IA générative et de lui demander de s'adapter à nos besoins de manière pointue – donc de nous-mêmes l'entraîner à effectuer une tâche précise, une méthode connue sous le nom d'*in-context learning*. « L'avantage est que les données qu'on fournit au modèle dans ce contexte ne sont pas partagées entre les utilisateurs », relève Gaétan Marceau Caron. Bref, impossible de corrompre l'outil comme l'avaient fait les internautes avec Tay.

Malgré toute l'intelligence de l'IA, il faudra donc vraisemblablement que des humains continuent de contribuer à son apprentissage. Tant mieux, estime Gaétan Marceau Caron, qui cite en exemple l'IA Alpha Go, qui avait affiné ses compétences à ce jeu chinois en jouant contre lui-même. « Il avait un style de jeu digne d'un extraterrestre ! » rappelle-t-il. Cet exemple sert selon lui d'avertissement : « Si on veut une IA proche de l'humain, qui sert à l'humain et que l'humain peut comprendre, il faut une rétroaction humaine. » ●

POUR UNE PÊCHE INTELLIGENTE

Si l'intelligence artificielle attise les craintes de certains spécialistes, elle peut aussi faire le bien, comme le prouve iPêche, une application qui aide à identifier les poissons des lacs et des rivières du Québec.

Les carpes envahissantes n'ont qu'à bien se tenir : les agents et agentes de protection de la faune les ont à l'œil, littéralement ! Une nouvelle application mobile gratuite permet d'identifier rapidement les espèces de poissons d'eau douce susceptibles d'être pêchées dans la province, espèces indésirables incluses. Il suffit de prendre son poisson en photo pour que l'algorithme d'intelligence artificielle (IA) d'iPêche l'analyse, puis rende son verdict.

« Au Québec, il y a trop de lacs pour le nombre de spécialistes qui en surveillent les populations. Et comme certaines espèces exotiques les envahissent, c'est problématique », affirme Jean Boissonneault, développeur applicatif au Centre en imagerie numérique et médias interactifs (CIMMI), l'un des 59 centres collégiaux de transfert technologique membres du Réseau des CCTT – Synchronex. C'est à lui et à son équipe que l'on doit iPêche.

En 2018, deux biologistes du secteur de la faune (actuellement sous la responsabilité du ministère de l'Environnement, de la Lutte contre les changements climatiques, de la Faune et des Parcs) avaient sollicité le groupe avec l'idée de mettre à contribution les pêcheurs et pêcheuses dans le suivi des poissons, relate le développeur. « À l'époque, le développement de réseaux de neurones profonds capables de reconnaître et de classifier des images était encore une technique émergente. » Ce mandat cadrait donc parfaitement avec la mission de recherche et d'innovation des CCTT.

UNE BELLE PRISE

Le nerf de la guerre avec cette classe d'algorithmes est de lui fournir d'énormes quantités de données de qualité. Pour chaque espèce, il faut un grand nombre d'images, avec des spécimens dans une variété de positions, d'angles et d'éclairages. « Les biologistes du Ministère nous ont fourni 38 000 images représentant 125 espèces de poissons. Chacune des images a été annotée à la main. Un travail fastidieux ! » souligne Jean Boissonneault.

Cela a permis au réseau de neurones de réaliser des apprentissages. « Dans notre plus récente version [octobre 2023], le premier résultat est le bon 97 fois sur 100. Le taux de succès grimpe à plus de 99 % si on considère les trois résultats proposés par l'application », indique M. Boissonneault.

Il y a encore un défi pour les équipes du CIMMI impliquées dans le projet, car ces pourcentages, calculés à partir d'une banque distincte de 4000 images qui n'ont pas servi à entraîner l'algorithme, diminuent quelque peu dès qu'on lui soumet une plus grande diversité de contextes, comme ceux rencontrés par les utilisateurs et utilisatrices d'iPêche. De plus, les résultats sont moins probants avec des espèces rares ou peu photographiées. Qui donc tire le portrait des ménés ?

La solution est de poursuivre l'entraînement d'iPêche avec de nouvelles photos. Avec plus de 17 000 téléchargements, le « Shazam » des poissons, comme a titré *Le Soleil*, est sur la bonne voie. « Pour l'instant, plusieurs l'utilisent comme un carnet qui permet de conserver les renseignements sur leurs captures. Le défi sera d'offrir d'autres fonctionnalités pour augmenter l'engagement des utilisateurs et tirer profit au maximum des capacités technologiques des appareils intelligents », conclut Jean Boissonneault.

CIMMI Centre en imagerie numérique et médias interactifs

Pour télécharger iPêche →

App Store

Google Play

POUR UNE IA MOINS ÉNERGIVORE

Les centres de données utilisent 1 % de la production mondiale d'électricité et devraient en utiliser encore plus avec la croissance de l'IA.

Pour réduire la consommation énergétique de l'intelligence artificielle, des solutions techniques existent. Mais d'abord, il faut utiliser les grands modèles avec parcimonie, disent des scientifiques.

Par Alexis Riopel

L'entrée en scène de ChatGPT, en novembre 2022, a déclenché une ruée sans précédent vers l'intelligence artificielle (IA). N'importe qui peut maintenant poser une question au robot depuis le confort de son salon. Dans les centres de données de Microsoft, d'Amazon et de Google, des milliers de serveurs ultra-spécialisés roulent à fond de train pour répondre à la demande engendrée par les grands modèles d'IA générative.

Cela requiert évidemment de l'énergie. L'entraînement initial de GPT-3 – le modèle sous-jacent à ChatGPT – a nécessité plus de 1200 mégawattheures d'électricité : autant que ce qu'une cinquantaine de maisons québécoises consomment en douze mois. Cela peut paraître peu, mais il faut aussi prendre en compte les milliards de requêtes lancées au robot conversationnel chaque mois. On estime que les centres de données de ChatGPT utilisent plus de 500 mégawattheures par jour. Et ChatGPT n'est qu'un modèle parmi d'autres.

Si l'IA générative continue de se répandre comme une traînée de poudre, la consommation d'énergie explosera au bout de la mèche. Que les serveurs carburent aux énergies fossiles ou renouvelables, il y a de bonnes raisons de couper court à cette hémorragie. Mais comment faire pour réduire la consommation énergétique de l'IA sans pour autant se passer de ses services ?

Tout d'abord : utiliser les grands modèles d'IA seulement quand ils sont bel et bien le meilleur outil pour accomplir la tâche désirée. Ces mastodontes logiciels sont souvent déclassés par de petits modèles spécialisés d'IA. « On essaie de vendre les grands modèles de langage comme une solution à tout. Pourtant, ce sont des outils très peu fiables, qui peuvent activement induire en erreur quand vient le temps de résoudre la majorité des problèmes », affirme David Rolnick, professeur d'informatique à l'Université McGill et chercheur à Mila – l'Institut québécois d'intelligence artificielle.

Le groupe de recherche de M. Rolnick développe justement des algorithmes légers, mais fondés sur l'apprentissage automatique, destinés à faire l'analyse d'images satellitaires. Ces programmes peuvent par exemple évaluer la productivité agricole d'un territoire ou en estimer la déforestation. « Ils sont des milliers de fois plus efficaces, pour une même puissance de calcul, que les grands modèles d'IA », souligne celui qui est aussi directeur de l'organisme Climate Change AI.

Sasha Luccioni, chercheuse au sein de Hugging Face, une plateforme qui fournit des outils pour construire des systèmes d'IA basés sur des technologies en libre accès, prêche aussi pour l'utilisation de petits modèles spécialisés. Elle voit d'un très mauvais œil le mariage entre l'IA générative et les services numériques de base, comme les courriels, la cartographie et la recherche en ligne. « Je ne vois pas l'intérêt. Est-ce que tu as besoin que Google Maps te raconte un haïku quand tu veux savoir comment te rendre au mont Tremblant ? C'est la mode en ce moment, et c'est dommage. »

Une fois le principe de la parcimonie respecté, on peut réduire la consommation d'énergie des grands modèles d'IA grâce à des astuces algorithmiques. Des librairies logicielles comme DeepSpeed permettent d'optimiser l'utilisation des puces informatiques en lançant tous les calculs en parallèle, sans laisser de répit au moindre transistor. En résultent un entraînement jusqu'à 2,8 fois moins long et une cadence de réponse aux requêtes jusqu'à 6,2 fois plus rapide. Et qui dit moins de temps de calcul dit moins d'électricité consommée.

Une autre approche, que Hugging Face se fait une fierté de mettre au point, est ce qu'on appelle la « distillation » des grands modèles de langage, comme GPT et BERT. « On part de la même architecture que GPT et on enlève les connexions les moins importantes pour accomplir la tâche désirée. C'est comme couper les branches mortes d'un arbre », explique Sasha Luccioni. Par exemple, plutôt que d'envisager 15 000 mots de la langue anglaise pour élaborer sa réponse, le robot peut se contenter de rétorquer « oui » ou « non ».

Malheureusement, la distillation n'est pas très populaire dans la communauté de l'IA, déplore Mme Luccioni, l'une des scientifiques les plus réputées dans le monde en ce qui concerne l'empreinte carbone de ces technologies. « Il y a vraiment cette tendance de *bigger is better*, mais, souvent, c'est surtout pour les apparences… »

Et puis, un obstacle très concret se présente aux équipes qui veulent distiller les grands modèles d'IA : il faut que ceux-ci soient disponibles pour téléchargement pour être personnalisés. Or, depuis l'avènement de ChatGPT, les géants technologiques se referment comme des huîtres pour protéger leurs algorithmes de la concurrence.

PUCES NOUVEAU GENRE

La consommation d'énergie des grands modèles d'IA peut aussi être modérée grâce au développement de nouvelles puces électroniques plus efficaces. « C'est beaucoup ce qu'on voit dans l'industrie actuellement », observe François Leduc-Primeau, professeur au Département de génie électrique de Polytechnique Montréal. Dans le monde de l'IA, ce sont les processeurs graphiques (GPU, en anglais) qui ont la cote, car ils se prêtent bien aux opérations en parallèle. Pour les rendre plus efficaces, les concepteurs et conceptrices tentent de parfaitement adapter la forme de leurs circuits aux calculs du réseau de neurones.

M. Leduc-Primeau, lui, s'intéresse à une solution radicalement différente : le calcul « en mémoire ». Selon cette architecture, les opérations mathématiques sont réalisées directement dans les circuits qui stockent les données. Cela évite de transporter l'information d'une puce à l'autre. Il y a d'importantes économies d'énergie à la clé, mais des incertitudes peuvent être introduites dans les calculs. Le chercheur, issu du milieu des télécommunications, développe des méthodes de correction d'erreurs pour pallier ce problème. Il croit que, d'ici cinq ans, le calcul en mémoire va gagner beaucoup de terrain.

Pour que les grands modèles d'IA gobent moins de joules, il faudra par ailleurs que les entreprises technologiques fassent preuve de bonne volonté. Pour OpenAI ou Meta, qui disposent d'énormément de puissance de calcul, la voie de l'efficacité est moins alléchante que celle menant vers le prochain grand coup d'éclat. Et pour véritablement réduire les émissions de carbone de l'IA, il faudra aussi réfléchir à ce qu'on fait de cette technologie, ajoute M. Rolnick. L'industrie pétrolière et gazière utilise déjà des algorithmes d'apprentissage profond pour prospecter le sous-sol et optimiser ses chaînes de production, souligne-t-il… ●

PHOTO : SHUTTERSTOCK.COM