LE POSTER DU 1

7 Comment contrôler ce que produit l'IA ?

par Isabelle Ryl

ISABELLE RYL, *Vice-présidente intelligence artificielle à l'Université PSL (Paris sciences et lettres), elle y dirige le Pr[ai]rie-PSAI (Paris School of AI), qu'elle a cofondé. Elle a écrit, avec Jamal Atif et John Peter Burgess,* Géopolitique de l'IA *(Le Cavalier bleu, 2022).*

IL N'EXISTE PAS UNE SEULE IA, mais différents systèmes qui utilisent des algorithmes variés. Un logiciel fonctionnant avec du *reinforcement learning* - ou « apprentissage par renforcement », une méthode qui lui permet de perfectionner ses réponses au fil des expériences auxquelles il est soumis - n'est ainsi pas similaire à une intelligence artificielle générative. Or c'est de ces dernières que l'on parle le plus aujourd'hui. Les IA génératives reposent sur de grands modèles qui suivent différentes phases : la phase d'entraînement et la phase d'inférence. D'abord entraînée sur une grande quantité de données préexistantes, elle apprend ensuite, lors de la phase d'inférence, à en générer de nouvelles, souvent pour répondre à une question précise qui lui est posée par le biais d'une interface de dialogue.

Elle peut alors produire la réponse la plus probable.

La vérification est toujours une question très difficile en informatique, même au-delà du champ de l'intelligence artificielle. Dans le cas de l'IA générative, qui sert à produire, et donc à inventer des contenus inédits, que signifierait « vérifier » ? Est-ce par exemple vérifier l'intérêt d'un dessin de presse généré pour illustrer un article et s'assurer qu'il est conforme à nos souhaits ? Est-ce évaluer la véracité d'une information produite ? Dans un cas comme dans l'autre, une évaluation automatique s'avère difficile. Certes, si l'on demande à une IA générative de nous dire en quelle année est né Louis XIV, on s'attend à ce qu'elle nous réponde 1638 et non 200 après Jésus-Christ. Mais, le plus souvent, l'évaluation va reposer sur l'utilisateur, ce qui suppose que celui-ci ait des connaissances suffisamment étendues pour s'apercevoir de fautes éventuelles.

> ***On ne peut pas dire avec une certitude complète si un contenu a été ou non produit grâce à l'IA***

On ne peut pas non plus dire avec une certitude complète si un contenu a été ou non produit grâce à l'intelligence artificielle. Pour un texte, on peut supposer qu'il a été produit par une IA en analysant la récurrence de certains mots dans certains contextes, puisqu'une intelligence artificielle produit les mots « les plus probables ». Mais ce ne sont encore que des pistes, et les recherches ne sont pour le moment pas assez concluantes pour être commercialisées. Pour ce qui est des images, il est possible de « tatouer » celles produites par IA grâce à un processus appelé stéganographie : ce sont des marques invisibles à l'œil nu, seulement quelques pixels modifiés ; en utilisant le logiciel adéquat, on retrouve une signature indiquant de manière non falsifiable qui a produit l'image. En développant cette technique, certaines entreprises, comme Google, s'engagent à favoriser la diffusion d'informations justes et compréhensibles pour les êtres humains.

Par-delà le contrôle factuel du contenu généré, comment vérifier si l'IA en question n'a pas été nourrie par d'autres contenus produits par d'autres systèmes d'IA ? Menés par le scientifique Ilia Shumailov, des chercheurs se sont penchés sur la question. Ils ont entraîné un modèle de langue sur des données du Web, puis en ont entraîné un autre sur les données produites par le précédent, puis un troisième sur ces données nouvellement récoltées, et ainsi de suite. Les résultats étaient clairs : au bout de neuf ou dix occurrences, le texte généré par l'IA n'a plus aucun sens. Pourquoi ? Parce que ces modèles proposent des réponses basées sur la probabilité, or l'entonnoir de ce qu'il propose, pour ainsi dire, se réduit.

À force de tourner en boucle sur ses propres contenus, l'IA s'appauvrit et se corrompt.

Il reste que le moyen le plus sûr de vérifier ce que produit une intelligence artificielle générative est humain : c'est la vigilance de l'utilisateur.

LE POSTER DU 1

8 Peut-on perdre le contrôle de l'IA ?

PERSONNELLEMENT, je ne le pense pas. On peut bien sûr rencontrer de graves problèmes avec l'IA si le logiciel est défectueux - comme cela peut arriver lorsque le pilote automatique d'un avion se trompe, même si, dans ce cas, il ne s'agit pas à proprement parler d'intelligence artificielle. Néanmoins, le principal danger encouru avec l'IA réside à mes yeux dans l'usage qu'on en fait. L'an dernier, un avocat américain qui avait utilisé ChatGPT pour préparer un procès a cité de multiples arrêts inexistants. C'est grave, mais l'IA est ici moins en cause que son utilisateur. Reste qu'elle n'est pas un outil anodin : d'usage facile, on l'a mise entre toutes les mains sans que tous soient vraiment au fait de ses capacités et de ses limites. Il faudrait donc former chacun à bien s'en servir. L'une des craintes les plus répandues est celle d'une IA surclassant l'homme, s'incarnant dans un robot et cherchant à détruire l'humanité. Je n'y crois pas. Il ne faut pas oublier que nous parlons de logiciels ne faisant que ce qui leur est demandé. Il y a une différence entre une intelligence artificielle autonome et une volonté propre. Prenons un robot programmé grâce à de l'intelligence artificielle pour secourir des victimes dans des situations de crise - un accident nucléaire ou l'effondrement d'un immeuble, par exemple.

Qu'il soit en partie ou totalement autonome est évidemment nécessaire, sans quoi, en cas de problème de communication entre lui et l'humain, il ne pourrait plus accomplir sa tâche. Mais il n'en changera pas de lui-même : s'il a été programmé pour trouver des victimes sous les décombres, il ne décidera pas tout seul d'aller s'asseoir au bout de la route pour dessiner des fleurs.

L'intelligence artificielle pourrait-elle toutefois extrapoler à outrance des réponses à un problème ? Jusqu'où pourrait-elle aller pour réaliser la commande ? Tout dépend de sa programmation et des garde-fous mis en place - en quelque sorte, des « règles » qui lui auront été imposées. Aujourd'hui, avec ce genre de machines, le risque n'est pas qu'elles « violentent » un blessé, mais qu'elles blessent quelqu'un à cause d'un simple écart : souvent lourdes, elles possèdent une grande force ; une petite erreur dans le calcul de leur trajectoire peut les conduire à vous rouler sur le pied ou, pour reprendre l'exemple de la victime sous les décombres, à se déplacer trop vite ou à saisir avec trop de force une personne. C'est pourquoi la robotique nécessite des logiciels vérifiables et certifiables.

Dernière inquiétude : ces outils peuvent-ils être programmés par des gens mal intentionnés ? Oui, bien sûr, comme beaucoup d'autres. Mais cela ne doit pas alimenter le fantasme d'une IA débordant l'intelligence humaine. En réalité, tout dépend ici de ce que l'on entend par « intelligence » : si l'on parle par exemple de la capacité de lire rapidement des pages, les machines nous battent déjà ; pour autant, aucune aujourd'hui ne surpasse l'être humain sur l'ensemble de son spectre intellectuel, dans sa capacité à raisonner et à inventer.

par Isabelle Ryl