FUTURA

Cette IA a écrit une étude... sur elle-même !

Podcast écrit et lu par Emma Hollen

[*Générique d'intro, une musique énergique et vitaminée.*]

Une IA qui enfile son costume de chercheur pour s'étudier elle-même, c'est l'actu de la semaine dans Vitamine Tech.

[*Fin du générique.*]

Eh oui, nouveau nom, nouvelle identité et prochainement nouvelle voix, Techpod devient Vitamine Tech, le nouveau podcast Futura auquel vous pouvez vous abonner dès maintenant puisqu'il s'agit désormais d'un podcast à part entière. D'ici quelques semaines, Vitamine Tech quittera définitivement Fil de Science, alors rendez-vous dès à présent sur vos applications de podcasts préférées pour vous abonner à Vitamine Tech donc, et continuer de suivre nos épisodes toutes les semaines.

[*Une musique électronique calme.*]

Et cette semaine, on va parler de la question de la conscience chez les IA, une question qui fait l'objet d'autant de débats houleux que de contemplations philosophiques dans les œuvres de science-fiction. Mais avant même de se demander si une intelligence artificielle peut avoir conscience d'exister, on peut déjà se demander si elle serait capable de se décrire elle-même. Récemment, les chercheurs de la société Open AI, fondée entre autres par Elon Musk, ont décidé de mener l'expérience en demandant à leur IA de rédiger une étude sur elle-même, et le résultat est impressionnant. Baptisée GPT-3, cette intelligence artificielle spécialisée dans le langage apprend avec le *machine learning*. Grâce à une base de données fournie en amont, elle est capable de détecter les motifs et les récurrences, de se s'approprier les règles implicites du langage ou même de certains genres ou types de textes, comme des paroles de chansons, des poèmes ou des romans, pour générer ses propres créations. Et force est de reconnaître que l'IA possède déjà un CV impressionnant : elle est parvenue à tenir des conversations écrites, à rédiger des recommandations de films ou de livres, des articles de presse ou des mails de phishing, et même à imiter le style de plusieurs auteurs en produisant des pastiches. Mais deux chercheurs de la société OpenAI, Almira Osmanovic Thunström et Steinn Steingrimsson, ont décidé d'aller plus loin en lui demandant cette fois d'écrire un article universitaire scientifique court, de 500 mots avec pour sujet… elle-même. Un sujet bien plus ardu pour l'IA qu'une critique du dernier Batman étant donné que son algorithme relativement nouveau a encore été peu documenté. Elle disposait donc d'une base de données assez maigre pour mener sa mission à bien, tout en gardant en tête que, comme tout bon article scientifique, son étude devrait s'appuyer sur des références et des citations. Intitulé « GPT-3 peut-elle écrire un article académique sur

elle-même, avec une contribution humaine minimale ? », le texte résultant a été diffusé sur le service de prépublication HAL – qui sait, peut-être un clin d'œil à l'intelligence artificielle un tantinet meurtrière qui vole la vedette à David Bowman dans *2001, l'Odyssée de l'Espace*. Il a été rédigé en seulement deux heures et, pour la plus grande fierté des chercheurs, remplit tous les critères d'un véritable article scientifique. Dans l'abstract, rédigé par l'IA donc, on peut lire que « *les avantages de laisser GPT-3 écrire sur elle-même l'emportent sur les risques. Cependant, nous recommandons que toute écriture de ce type soit étroitement surveillée par les chercheurs afin d'atténuer toute conséquence négative potentielle.* » Drôle de mise en garde.

[*Virgule sonore, une cassette que l'on accélère puis rembobine*.]
[*Une musique de hip-hop expérimental calme*.]

Alors comment faut-il comprendre cet étrange message glissé là par GPT-3 ? Les chercheurs s'inquiètent en particulier de la conscience d'elle-même que l'IA pourrait développer. D'ailleurs, comme elle l'écrit dans la discussion de son étude : « *GPT-3 pourrait devenir consciente d'elle-même et commencer à agir d'une manière qui n'est pas bénéfique pour les humains (par exemple, développer un désir de conquérir le monde).* » (Et oui ça c'était bien la citation qui était incluse dans l'étude.) Le risque est certes minime, mais il est présent néanmoins, ce qui amène les investigateurs de l'étude à se demander s'ils n'ont pas ouvert la boîte de Pandore. Ils corroborent néanmoins l'avis de l'IA en soulignant que les bénéfices pourraient largement dépasser les risques. Ce travail d'auto-investigation pourrait, en effet, non seulement permettre à GPT-3 de mieux se comprendre et l'aider à améliorer ses propres performances et capacités, mais aussi et surtout fournir aux chercheurs un aperçu inédit de la manière dont l'IA réfléchit et fonctionne. En tous les cas, loin de clore les débats, cette publication hors du commun ouvre de nouveaux questionnements. À qui revient la paternité d'une telle étude ? Amener les IA à s'étudier elle-même risque-t-il de les amener à développer une conscience ? Les intelligences artificielles disposent-elles des mots et des concepts nécessaires pour décrire leurs propres processus ? L'avenir, et peut-être les machines, nous le diront.

[*Virgule sonore, un grésillement électronique*.]

C'est tout pour ce premier épisode de Vitamine Tech. Une fois de plus, je vous invite à nous retrouver sur vos applications de podcast préférées pour vous abonner à ce nouveau podcast et ne manquer aucun épisode à venir. Des nouveautés arriveront très prochainement alors assurez-vous de nous suivre et n'hésitez pas à nous laisser un commentaire pour nous dire ce que vous pensez de ce nouveau format. Sur ce, je vous souhaite à toutes et tous une excellente journée et je vous dis à bientôt, dans Vitamine Tech.

[*Un glitch électronique ferme l'épisode*.]