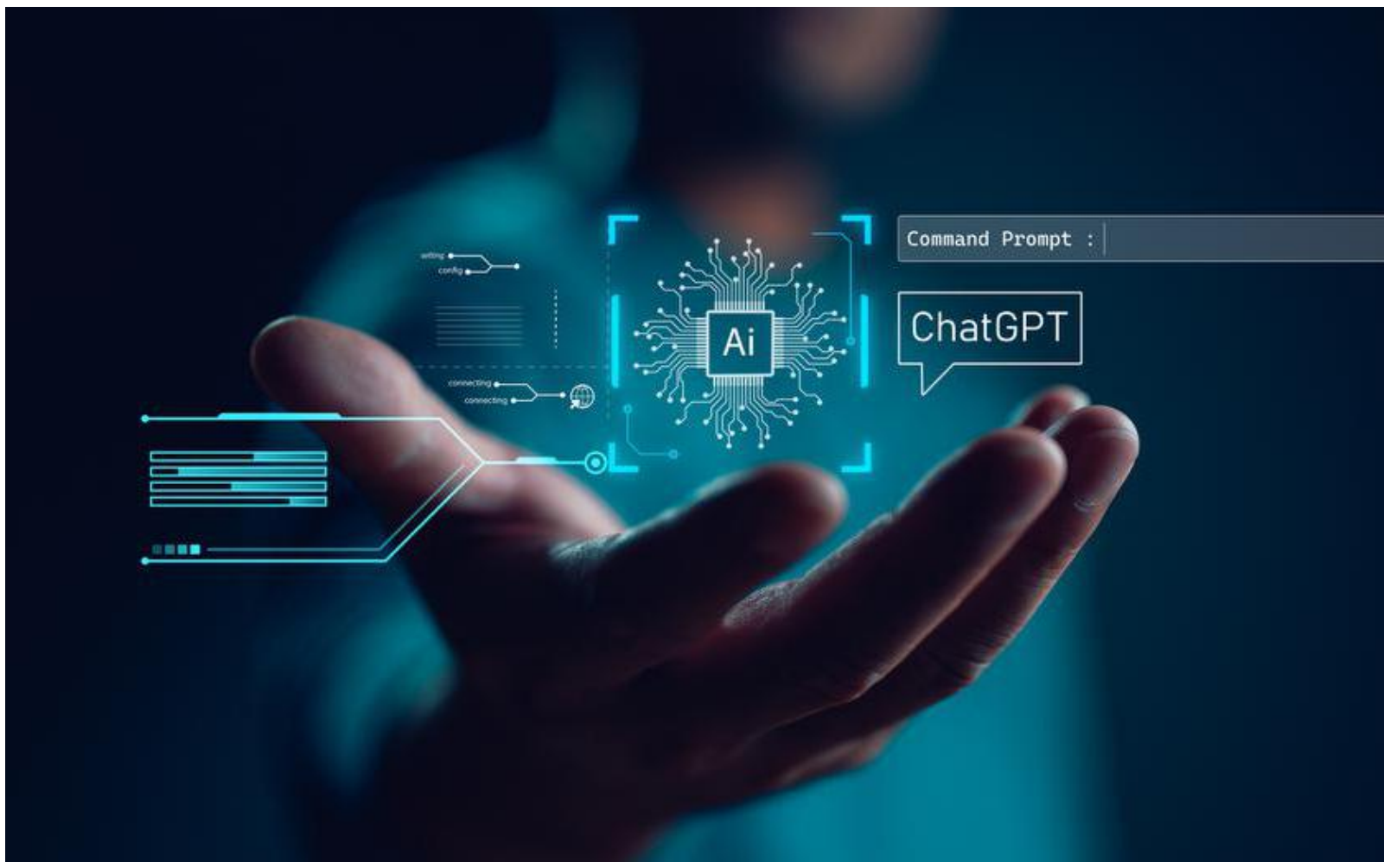
INSTANT TECHNO

Ces derniers jours, Microsoft a été soupçonné d'espionner les contenus des documents de Word, Excel et PowerPoint de ses utilisateurs. Il n'en est rien. Mais les données viennent bien de quelque part.



Comment les IA pillent nos données

Si l'on sait que les Gafam récoltent des quantités astronomiques de données, on peine à connaître leurs sources.
© SHUTTERSTOCK.

THOMAS CASAVECCHIA

A-t-on vraiment envie que le contenu de tous les documents que l'on rédige et que chaque mail que l'on envoie soit matière à entraîner les IA ? Une rumeur a explosé, ces derniers jours, relayée par de nombreux sites d'information. Elle assurait que Microsoft utilisait les données contenues dans les documents créés par les utilisateurs de Word, d'Excel ou de PowerPoint pour entraîner ses intelligences artificielles. Chaque mot, chaque image et chaque bilan comptable pouvaient donc être extraits et analysé par « la firme de Redmond » (le siège social de Microsoft se trouve dans cette ville, NDLR) pour faire gonfler les bases de données utilisées pour rendre les intelligences artificielles génératives plus performantes.

Vent de panique parmi les nombreux utilisateurs de ces services, peu enclins à voir des informations potentiellement sensibles collectées et analysées sans leur accord direct depuis leurs documents, mails et tableurs. Car la « fonctionnalité » était activée de base, planquée dans les paramètres des applications et décrite selon un vague « expériences connectées facultatives », peu transparent sur la finalité de ces échanges de données.

Dans un exercice de communication de crise, sur X, le compte Microsoft 365 a assuré que cette option ne relevait en rien de la récolte de données à des fins d'entraînement d'IA. Elle sert à enrichir les documents en effectuant des recherches sur le web. Ouf !

Mais la propagation de cette rumeur montre bien une chose : si l'on sait que les Gafam, tous marathoniens dans cette course effrénée à l'intelligence artificielle, récoltent des quantités astronomiques de données, on peine à connaître leurs sources. Autre instruction : après des décennies de pillage en règle des données par les réseaux sociaux et différentes applications en ligne, à des fins de ciblage, les internautes n'ont pas envie de voir leur vie privée à nouveau sacrifiée sur l'autel du développement technologique.

Or, si les IA génératives ont réalisé de tels progrès ces dernières années, c'est bien parce que des récoltes massives de données ont eu lieu. Mais où ça ?

Difficile à dire. « On est dans le flou total », résume Etienne Wery avocat au barreau de Bruxelles, spécialisé dans les enjeux technologiques. « Je m'interroge ; pourquoi les grands acteurs derrière l'IA aujourd'hui sont aussi les champions du *cloud* ? Quand on sait que les IA ont besoin d'énormément de données et que les serveurs de ces grosses entreprises en contiennent justement, on peut se poser la question. Les Google, Amazon et autres Microsoft assurent que les données qu'ils stockent via leurs services de *cloud* ne sont pas utilisées à cette fin. Et personne n'a la preuve du contraire. Mais je n'ai pas particulièrement confiance. Et je pense que l'on devrait se méfier davantage. » Et légiférer pour protéger les citoyens ?

Des sources parfois illégales

Une directive européenne vise justement à encadrer la fouille de texte et de données en ligne à des fins scientifiques. « Elle prévoit une exception au droit d'auteur si les recherches sont menées par un organisme de recherche ou une institution d'héritage culturel. Un tribunal allemand a statué et considéré que cette exception s'applique même si l'organisme de recherche a un objectif commercial. »

Les Gafam auraient donc carte blanche pour s'asseoir sur le droit d'auteur ? « Pas si vite », poursuit l'avocat. « Car les données en question doivent bien sûr être obtenues de manière licite. » Or leur origine est parfois douteuse.

La plupart des géants de l'informatique, s'ils ne sont pas très bavards concernant leurs sources, assurent que les données récupérées sont « librement accessibles » sur la « toile ». Rassurant. Pourtant, « accessibles » ne veut pas dire légales. Pour le développeur Ed Newton-Rex cité par le média en ligne Axios, « librement accessible au public ne veut pas dire que qui que ce soit a donné l'autorisation pour que ces données soient utilisées pour entraîner des systèmes d'IA ». Pire, ces données proviennent parfois de sites qui enfreignent la loi. Toujours cité par Axios, l'avocat Timothy K. Giordano explique que de nombreux contenus piratés, comme des copies de livres, facilement accessibles en ligne bien que parfaitement illégales, sont utilisés. Idem pour les images, souvent couvertes par le droit d'auteur, qui sont analysées sans autre forme de procès sur la « toile ».

Au printemps 2023, le *Washington Post* a analysé le contenu de la base de données « Google C4 », un corpus utilisé par Google et Facebook pour entraîner leurs modèles.

Des bases de données, mais pas seulement

Le journal a ainsi pu révéler que les sources étaient multiples : Google Patents, le site d'indexation de brevets de Google, Wikipédia, des millions de blogs et de pages de forums. Le quotidien a aussi révélé des contenus aux origines douteuses, comme un célèbre site de piratage de livres qui avait été largement utilisé. Autre signe semblant indiquer une violation récurrente de la propriété intellectuelle : le symbole copyright apparaît plus de 200 millions de fois dans la base de données. Enfin, la moitié du top 10 des sites les plus représentés comptait des sites d'informations. Ces derniers ne sont pourtant généralement accessibles que contre un abonnement et la plupart contiennent, dans leur code, un script signalant aux récolteurs de données qu'ils refusent cet accès. Une protection qui ne semble donc pas suffisante.

Mais cette base de données reprenant l'essentiel du contenu de plus de 15 millions de sites web n'est bien sûr pas la seule source utilisée par les créateurs d'IA. GPT 3 d'OpenAI avait, par exemple, été entraîné par 40 fois plus de matière que celle stockée sur « Google C4 ». La société tait toutefois ses sources, se retranchant derrière le fameux contenu « librement accessible sur internet ».

Les réseaux sociaux représentent, eux aussi, une manne inespérée. A l'instar des sites d'informations, ils empêchent la récupération automatique de données. Toutefois, ils ne s'en privent pas eux-mêmes. Ainsi, tous les *posts* de X peuvent être collectés pour nourrir Grok, l'IA aux capacités régulièrement vantées par Elon Musk. Chez Meta, toutes les interactions publiques sont utilisées pour entraîner son IA. En théorie, les échanges entre comptes protégés ainsi que les conversations privées sont donc épargnés par cette moisson.

l'expert « Les IA s'entraînent avec les données qu'on leur fournit de bon cœur »

Quand on utilise une IA, celle-ci s'entraîne avec ce qu'on lui dit et ce à quoi on lui donne accès

ENTRETIEN

TH.CA.

Hugues Bersini, professeur d'informatique à l'ULB est spécialiste de l'intelligence artificielle et des réseaux neuronaux. Pour lui, il ne fait aucun doute que quand on utilise une IA, on continue de l'améliorer en la nourrissant des documents qu'on lui fournit. C'est même ce que l'on recherche pour plus de commodité.

Peut-on imaginer que nos documents personnels soient utilisés pour entraîner les IA ?

Bien sûr. Les modèles de langage actuels ont appris à s'exprimer après avoir été entraînés sur de très gros corpus en ligne. On peut penser à Wikipédia, des forums comme Reddit ou 4Chan. Mais depuis, ils reposent beaucoup moins sur ces sources en ligne et toujours plus sur les interactions qu'ils ont avec leurs utilisateurs. Les dernières versions, en développement, de l'agent Copilot installé dans les prochaines versions d'Office peuvent compléter d'elles-mêmes un texte que l'on commence à rédiger sur Word. Pour ce faire, il va piocher dans tous les documents contenus dans le disque dur de son ordinateur. Il peut aussi accéder aux documents stockés dans le *cloud*. En rassemblant toutes ces données, il fait des choix et poursuit seul la rédaction. C'est certain qu'il utilise ces données pour s'entraîner. Il est tout aussi sûr que tous ces documents fournis transitent à un moment ou à un autre sur les serveurs de Microsoft. Il ne peut pas en être autrement. Quand on utilise une IA, celle-ci s'entraîne avec ce qu'on lui dit et ce à quoi on lui donne accès.

Et l'on continue pourtant à se servir de ces nouveaux outils ?

On est dans une phase complètement schizophrénique. On ne veut pas que nos données soient utilisées à tort et à travers et l'on continue pourtant à les fournir de bon cœur. Quand on discute avec des utilisateurs, la plupart sont ravis de ne plus avoir à écrire leurs mails. Ils sont tout aussi enthousiastes à l'idée de pouvoir résumer un long rapport en quelques clics et à pouvoir en générer rapidement un nouveau. Cela facilite énormément la vie des entreprises qui sont souvent très contentes de fournir toute une série de documents à ces assistants pour accélérer leurs processus.

Les scandales à répétition sur l'utilisation des données par les réseaux sociaux ne nous ont pas rendus plus méfiants ?

Pour l'instant, les GPT et consorts et le grand public sont dans une phase de lune de miel. Tout a été si rapide et les progrès ont été si impressionnants que cela a laissé beaucoup de monde extatique. On ne se soucie donc pas trop de ce que ces entreprises font des données récoltées. Mais c'est tout aussi vrai sur l'impact écologique désastreux qu'ont ces modèles. Ou des effets qu'ils peuvent produire sur nos capacités d'apprentissages. Ces questions sont encore peu débattues. Pour reprendre l'exemple des réseaux sociaux, il a fallu attendre des scandales sur la manière dont leurs données ont été utilisées à des fins de manipulation politique pour que l'on commence à s'en méfier. Aujourd'hui, on met surtout en avant leurs capacités toujours plus bluffantes et la manière dont ils peuvent se rendre utiles. Quand les premiers abus flagrants seront constatés, cette période dorée pourrait bien prendre fin.