

Dans les décennies à venir, les progrès rapides des algorithmes d'apprentissage engendreront des machines d'une intelligence comparable à la nôtre. Capables de parler et de raisonner, elles auront leur place dans une myriade de domaines tels que l'économie, la politique et, inévitablement, la guerre. La naissance d'une véritable intelligence artificielle (IA) aura un impact profond sur l'avenir de l'humanité et conditionnera l'existence même d'un tel avenir.

Prenez par exemple la citation suivante : « Encore aujourd'hui, des recherches sont en cours pour mieux comprendre ce que les nouveaux programmes d'IA seront capables de faire, tout en restant dans les limites de l'intelligence d'aujourd'hui. La plupart des programmes d'IA actuellement programmés se limitent principalement à prendre des décisions simples ou à effectuer des opérations simples sur des quantités relativement faibles de données. »

Peut-être avez-vous eu l'impression que quelque chose clochait dans ce paragraphe ? Cette citation est l'œuvre de GPT-2, un robot linguistique que j'ai pu tester. Développé par OpenAI, un institut basé à San Francisco qui promeut les IA « vertueuses », GPT-2 est un algorithme d'apprentissage fondé sur un réseau de neurones artificiels. Ses entrailles contiennent plus d'un milliard de connexions simulant des synapses, les points de jonction entre neurones.

La tâche du réseau est apparemment stupide : confronté à un texte de départ arbitraire, il doit prédire le mot suivant. Il ne « comprend » pas les textes comme le ferait un humain. Mais durant sa phase d'apprentissage, il a dévoré des quantités astronomiques de textes – huit millions de pages

— En bref —

> Deux théories s'opposent sur la nature de la conscience : celle de « l'espace de travail neuronal global » et celle de « l'information intégrée ».

> L'une implique que les machines deviendront conscientes lorsqu'elles auront des capacités suffisamment sophistiquées, l'autre non.

> Ces théories sont actuellement testées par une collaboration internationale.

internet en tout – et ajusté ses connexions internes pour mieux anticiper des suites de mots.

J'ai écrit les premières phrases de l'article que vous êtes en train de lire, puis je les ai « injectées » dans l'algorithme en lui demandant de composer une suite. Il a notamment craché le paragraphe cité. Certes, ce texte ressemble aux efforts d'un étudiant de première année pour se rappeler un cours d'introduction à l'apprentissage automatique durant lequel il aurait rêvassé. Mais le résultat contient malgré tout les mots et les phrases clés – pas si mal, vraiment !

L'EFFET D'ÊTRE UNE CHAUVESOURIS

Les algorithmes sauront-ils un jour écrire un chef-d'œuvre aussi abouti qu'*À la recherche du temps perdu*, de Marcel Proust ? Difficile à dire, mais les prémices sont là. Rappelez-vous comme les premiers logiciels de traduction et de conversation étaient faciles à tourner en dérision, tant ils manquaient de précision. Mais avec les réseaux neuronaux profonds et la mise en place d'infrastructures de calcul surpuissantes par les entreprises du numérique, les ordinateurs se sont améliorés sans relâche, jusqu'à ce que leurs productions n'aient plus rien de ridicule.

Bien que les experts ne s'accordent pas sur la nature exacte de l'intelligence, la plupart d'entre eux estiment que, tôt ou tard, les ordinateurs atteindront ce que l'on appelle l'intelligence artificielle générale : ils deviendront capables de réaliser n'importe quelle tâche intellectuelle accessible à un humain. Mais auront-ils pour autant une forme de sentiment d'eux-mêmes ? En d'autres termes, pourront-ils un jour être conscients ?

Le simulacre numérique d'un mort pourra parler, accéder à ses souvenirs, à ses envies, à ses peurs et aux traits de sa personnalité. Mais sera-t-il conscient?

Par «conscience» ou «sentiment subjectif», j'entends une qualité inhérente à toute expérience – par exemple, la saveur d'une madeleine, une rage de dents, l'étirement du temps quand on s'ennuie ou le sentiment de vitalité et d'anxiété juste avant une compétition. En paraphrasant le philosophe Thomas Nagel, auteur d'un article titré «Quel effet cela fait-il d'être une chauve-souris?», nous pourrions dire qu'un système est conscient si cela «fait quelque chose» d'être ce système.

Considérez l'embarras que vous ressentez lorsque vous vous apercevez que vous venez de commettre une gaffe, parce que votre interlocuteur a perçu comme une insulte ce qui n'était pour vous qu'un trait d'humour. Les ordinateurs éprouveront-ils un jour une telle émotion? Et lorsque vous êtes au téléphone depuis un temps interminable, écoutant une voix synthétique répéter en boucle: «Nous sommes désolés de vous faire attendre», le logiciel pourra-t-il se sentir gêné de vous maintenir dans cet enfer du service client?

Il ne fait guère de doute que notre intelligence et nos expériences résultent des capacités naturelles de notre cerveau à établir des enchaînements de causes et d'effets. Cette disposition a extrêmement bien servi la science au cours des derniers siècles. Le cerveau humain, cet organe d'à peine un kilo et demi à la texture comparable au tofu, est de loin le morceau de matière active organisée le plus complexe connu de l'Univers. Mais il obéit aux mêmes lois physiques que les pierres, les arbres et les étoiles. Rien n'échappe à ces lois. Nous ne comprenons pas encore tout à fait les mécanismes causaux en jeu, mais nous les expérimentons tous les jours: un groupe

de neurones s'allume pendant que vous voyez des couleurs, tandis que l'activation de cellules nerveuses dans une région voisine du cortex est associée à une humeur joviale. Lorsqu'un neurochirurgien stimule ces neurones à l'aide d'une électrode, le sujet voit des couleurs dans le premier cas, éclate de rire dans le second. Inversement, quand le cerveau est sous anesthésie, ces expériences disparaissent.

Compte tenu de cela, l'évolution de l'intelligence artificielle débouchera-t-elle sur une conscience artificielle?

DEUX THÉORIES ANTAGONISTES

Cette question mène à deux possibilités fondamentalement différentes. Selon la première, les machines vraiment intelligentes seront sensibles et dotées de conscience: elles parleront, raisonneront, s'occuperont d'elles-mêmes, seront capables d'introspection... Une idée dans l'air du temps, comme on le voit dans des romans et des films tels que *Blade Runner*, *Her* ou *Ex Machina*.

D'un point de vue scientifique, cette hypothèse s'appuie sur l'une des principales théories de la conscience, dite de «l'espace de travail neuronal global» (GNW, pour *global neuronal workspace*). Cette théorie considère que l'origine de la conscience réside dans certaines caractéristiques architecturales du cerveau. Elle s'inspire en cela de l'«architecture du tableau noir», développée par les sciences informatiques dans les années 1970: le principe est que des programmes spécialisés accèdent à un répertoire commun d'informations, appelé «tableau noir»