

> 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 !

Les successeurs de ces robots risquent de déclencher un raz-de-marée de faux articles et reportages, qui viendront polluer internet. Ce ne sera qu'un exemple de plus de programmes exécutant des prouesses que l'on croyait réservées aux humains : jouer à des jeux de stratégie en temps réel, traduire du texte, recommander des livres et des films, reconnaître les gens sur des images ou des vidéos...

Les algorithmes sauront-ils un jour écrire un chef-d'œuvre aussi abouti que *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 finesse et de précision. Mais avec l'invention des 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.

Comme on le voit avec le jeu de go, les échecs et le poker, les algorithmes d'aujourd'hui sont capables de surpasser les humains (en novembre 2019, Lee Sedol, un des plus grands joueurs de go de l'histoire, a décidé de prendre sa retraite après avoir perdu plusieurs fois contre l'algorithme AlphaGo ; il a déclaré que celui-ci était une entité ne pouvant plus être vaincue). Si bien que notre rire se fige : sommes-nous comme l'apprenti sorcier de Goethe, ayant invoqué des esprits serviables que nous ne pouvons plus contrôler ?

L'INTELLIGENCE N'EST PAS LA CONSCIENCE

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 dans le jargon (c'est-à-dire qu'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 ?

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, fondées sur des enchaînements de causes et d'effets. Cette prémisse 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

La théorie de l'information intégrée a inspiré la construction d'un appareil pour mesurer le niveau de conscience chez des patients incapables de communiquer. Il est ici testé par l'équipe de Steven Laureys, au Coma science group de l'université et du CHU de Liège.



LA FRANCE EST-ELLE CONSCIENTE ?

Selon la théorie de l'information intégrée, le degré de conscience d'un système dépend de son pouvoir causal intrinsèque – qui quantifie la capacité de ses sous-systèmes à s'influencer mutuellement. Mais alors, tout ensemble doté de parties en interaction n'est-il pas conscient ? N'est-ce pas le cas par exemple de l'ensemble que vous formez avec votre

interlocuteur au téléphone, voire de celui formé par la population d'un pays entier ? Non, car la théorie postule qu'à l'intérieur d'un système, la conscience apparaît précisément là où le pouvoir causal intrinsèque est maximum. Si vous parlez avec quelqu'un, vous allez vous influencer mutuellement, mais de façon très inférieure à ce qui se passe dans le cerveau de chacun. Il n'y aura donc pas formation d'un « superesprit » – pour cela, il faudrait augmenter drastiquement les connexions entre vos deux cerveaux. Il en va de même pour un pays comme la France et ses 67 millions d'habitants : même s'il y a de multiples interactions entre ces derniers, le pouvoir causal intrinsèque reste maximal à l'intérieur du cerveau de chacun. C'est donc là que surgit la conscience.

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 ou espace de travail central. Les psychologues ont postulé qu'un espace de ce type existe dans le cerveau et qu'il est essentiel à la cognition humaine. Sa capacité est faible, de sorte qu'à un

moment donné, il ne contient qu'une seule perception, qu'une pensée unique ou qu'un seul souvenir. Dans cet espace de travail, les nouvelles informations entrent en concurrence avec les anciennes et les remplacent.

Le neuroscientifique Stanislas Dehaene et le biologiste moléculaire Jean-Pierre Changeux, tous deux au Collège de France, à Paris, ont transposé ces idées à l'architecture du cortex, la couche externe du cerveau. Ils ont postulé que l'espace de travail repose sur un réseau de neurones dits « pyramidaux », qui relient des régions corticales éloignées, en particulier les zones associatives préfrontales, pariéto-temporales et médianes (cingulaires).

Une grande partie de l'activité cérébrale reste localisée et de ce fait n'accède pas à la conscience. C'est le cas des modules neuronaux qui contrôlent la posture du corps ou la direction du regard. Mais lorsque l'activité d'une ou plusieurs régions dépasse un seuil critique – disons, lorsqu'on présente à quelqu'un l'image d'une friandise appétissante –, elle déclenche une vague d'excitation neuronale qui se propage à travers l'espace de travail, dans tout le cerveau. Ce signal devient alors accessible à une multitude de processus auxiliaires, tels que le langage, la planification, le circuit de la récompense, la mémoire à long terme et le stockage dans une mémoire tampon à court terme.

Ce serait le fait de diffuser cette information à l'échelle globale qui la rendrait consciente. Ainsi, l'expérience subjective de la friandise appétissante résulterait de l'action de neurones pyramidaux qui ordonnent à la région motrice du cerveau de se saisir de la friandise, tandis que d'autres modules neuronaux anticipent une récompense sous la forme d'une poussée de dopamine, causée par le petit *shoot* de sucre et matières grasses à venir.

En d'autres termes, les états conscients découleraient de la façon dont l'algorithme de >

1 milliard de synapses virtuelles: c'est ce que contient l'architecture du logiciel linguistique GPT-2