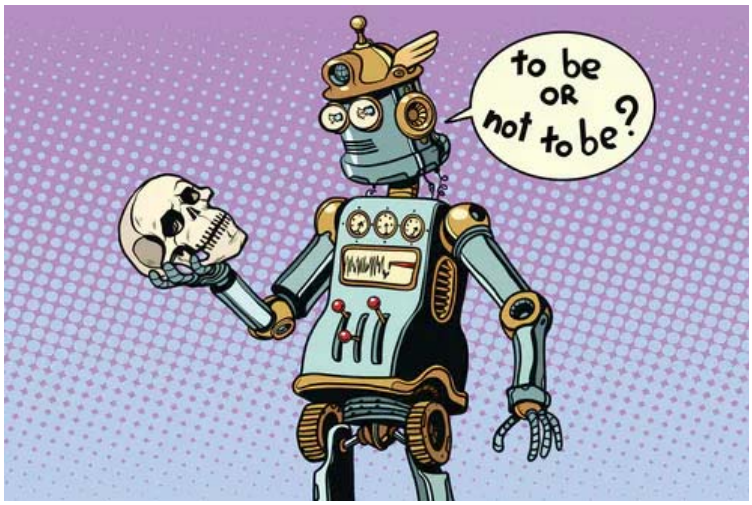
Les IA comprennent-elles ce qu'elles font ?

Fabien Gandon



Les IA ne saisissent pas les finalités, les conséquences et le contexte de ce qu’on leur demande. studiostoks / shutterstock

Les IA comprennent-elles ce qu’elles font ?

20 novembre 2020, 12:00 CET

Auteur

Fabien Gandon
Research Director, Inria

Que ce soit dans vos choix de séries TV à regarder, dans l'analyse de vos résultats médicaux, dans vos rencontres amoureuses, dans l’attribution de votre prêt ou dans les réglages de vos prises de photos – les IA mettent leur grain de sel partout. Ce qui rend peut-être d’autant plus surprenante la réponse à cette question : non, aujourd’hui, les IA ne comprennent pas ce qu’elles font.

Par contre, décomposer la question et détailler la réponse permet de soulever beaucoup de problèmes, d’ambiguïtés et d'enjeux de l'« intelligence » artificielle. « Intelligence » avec des guillemets, car les méthodes actuelles sont essentiellement des simulations très spécifiques et convaincantes, sur lesquelles nous projetons beaucoup plus que ce qu’elles renferment réellement, à commencer par cette impression de compréhension. Il est difficile de savoir si, dans le futur, les IA continueront à ne pas comprendre. Mais, en l’état des connaissances, la réponse est « non »... jusqu’à preuve du contraire.

Si on parcourt des dictionnaires, on peut lire que « comprendre », c’est saisir le sens, les finalités, les causes et conséquences, les principes. Comprendre quelque chose, c’est recevoir ou élaborer une représentation de cette chose, c’est s’approprier une conceptualisation reçue ou construite, qui permettra notamment de produire un comportement intelligent.

Du Graal de l’« IA forte » à la réalité de l’IA

Pour ce qui est des intelligences artificielles, une première catégorie est celle de l’« IA forte » ou « généralisée » : un seul et même système qui serait capable d’apprendre et d’effectuer tout type d’activité intelligente. Un tel système n’existe pas à ce jour et donc, pour ce cas, la question est close. On peut tout de même remarquer au passage que l'une des distinctions parfois faites entre « IA forte » et « IA faible » est justement la capacité de comprendre et d’être conscient.

Un système d’« IA faible » est un système conçu par l’humain pour simuler de la façon la plus

autonome possible un comportement intelligent spécifique. En l'état actuel de la recherche, c'est l'homme qui produit le système artificiel qui va simuler un comportement intelligent – ce n'est pas une intelligence artificielle qui, par sa compréhension, créerait son propre comportement. En IA faible, l'homme choisit la tâche pour laquelle on a besoin d'automatiser un comportement intelligent, les données qu'il fournit au système et leurs représentations informatique, les algorithmes utilisés pour simuler ce comportement (par exemple, un réseau de neurones spécifique), les « variables de sorties » ou les objectifs qu'il attend du système (par exemple, différencier les chats des chiens) et la façon dont ils seront intégrés à une application (par exemple, trier automatiquement les photos sur votre téléphone). La méthode d'IA faible ne comprend pas le contexte dans lequel elle est exécutée, ni de ce que ces différents aspects choisis par l'homme représentent.

Il existe beaucoup d'approches et de méthodes différentes pour produire des systèmes d'IA faible, mais elles partagent toutes cette absence de compréhension.

Différentes méthodes d'IA faible, dont aucune ne saisit la signification de ses calculs

On peut différencier différents systèmes d'IA faible par les données sur lesquelles ils travaillent, par exemple un texte, un graphe de connaissances, des vidéos ou un mélange de différents types de données.

Les systèmes d'IA faible se différencient aussi par les techniques qu'ils utilisent. On peut citer par exemple d'une part les réseaux de neurones pour *apprendre* de façon plus ou moins supervisée par un humain, d'autre part des « systèmes experts » pour *déduire* de nouvelles connaissances en utilisant un moteur d'inférence à partir de connaissances établies, ou encore des « systèmes multi-agents » pour simuler des *comportements sociaux*. Dans toutes ces méthodes, les buts sont fixés par l'humain, ainsi que l'évaluation des performances de l'IA.

Prenons l'exemple de différentes approches d'apprentissage automatique. Dans l'apprentissage supervisé dit « actif », l'IA est capable de solliciter d'elle-même des exemples à lui fournir pour améliorer son apprentissage, mais sans compréhension de ce qu'elle fait au-delà de cette optimisation. Dans le cas de l'apprentissage « non supervisé », l'humain ne fixe pas exactement les catégories de sorties attendues, mais le but reste fixé : il s'agit de découvrir des structures sous-jacentes à des données – l'IA propose d'elle-même des regroupements, mais elle est pilotée par des fonctions d'évaluation de ses performances qui sont fixées par son concepteur. Il en va de même pour l'apprentissage par renforcement dont le principe consiste à répéter des expériences pour apprendre les décisions à prendre de façon à optimiser une récompense, décidée par l'humain, au cours du temps.

Un autre exemple concerne la branche de l'IA qui ne s'intéresse pas à l'apprentissage, mais à l'inférence. Le composant central, dit « moteur d'inférences » cherche à déduire un maximum de conclusions à partir des connaissances qu'on lui fournit. Dans cette approche, la capacité intelligente que l'on souhaite simuler est essentiellement une forme de raisonnement. Mais ici encore, les inférences à faire et les méthodes utilisées sont déterminées par le concepteur et restent complètement en dehors du champ de vision et d'action du système : l'IA ne comprend rien à son contexte, elle simule des inférences ciblées.

À lire aussi : Comment motiver une IA ?

Une métaphore utilisée pour cette situation est celle de la chambre chinoise de Searle : une personne non sinophone est enfermée dans une pièce et fait illusion quant à sa capacité de comprendre et de s'exprimer en chinois, car elle répond à des questions inscrites sur des papiers reçus de l'extérieur uniquement à l'aide d'un livre indiquant quelle réponse donner à quelle question. Comme notre captif, les méthodes d'IA faible n'ont aucun accès à la signification de leur tâche, aucune mise en contexte. Cette simulation d'un comportement spécifique ne vient pas avec tout le reste des capacités que nous mobilisons à chaque instant et dans lequel notre *comportement intelligent* s'inscrit.

Sommes-nous dupes de notre propre farce ?

Le danger de l'IA c'est que lorsqu'elle simule assez bien une forme de comportement intelligent, on peut projeter beaucoup de choses sur elle, bien au-delà de ce qu'elle fait réellement. En particulier, on peut lui prêter une compréhension ou d'autres traits qu'elle n'a pas comme nous allons l'illustrer.

Prenons l'exemple des leurres conversationnels qui sont de petits systèmes très simples dont le seul but est de relancer la conversation pour que nous continuions à parler au système sans aucune compréhension de sa part. Le plus ancien et le plus connu est le chatbot Eliza qui dans les années 60 singeait une séance de psychothérapie avec des modèles de phrases toutes faites comme « et comment vous-sentez vous à propos de [...] ? ». Plus récemment le chatbot vocal peu sophistiqué Lenny utilise juste un ensemble de phrases préenregistrées afin de faire parler le plus longtemps possible une personne vous appelant pour du démarchage par téléphone. Eliza et Lenny montrent comment un programme simpliste peut nous duper : par conception, le programme ne comprend rien, mais l'utilisateur va se comporter, interagir et finir même par croire qu'il comprend.

Ce sont là moins des démonstrations de prouesses d'intelligence artificielle que des preuves des limites de l'intelligence humaine et ces limites peuvent participer à faire qu'une IA, même « faible », soit dangereuse. On voit des soldats-démineurs s'attacher à leurs robots-démineurs au point d'interroger sur leur capacité à accepter de les mettre en danger alors que leur raison d'être est celle d'artefacts que l'on souhaite sacrifier à la place d'humains.

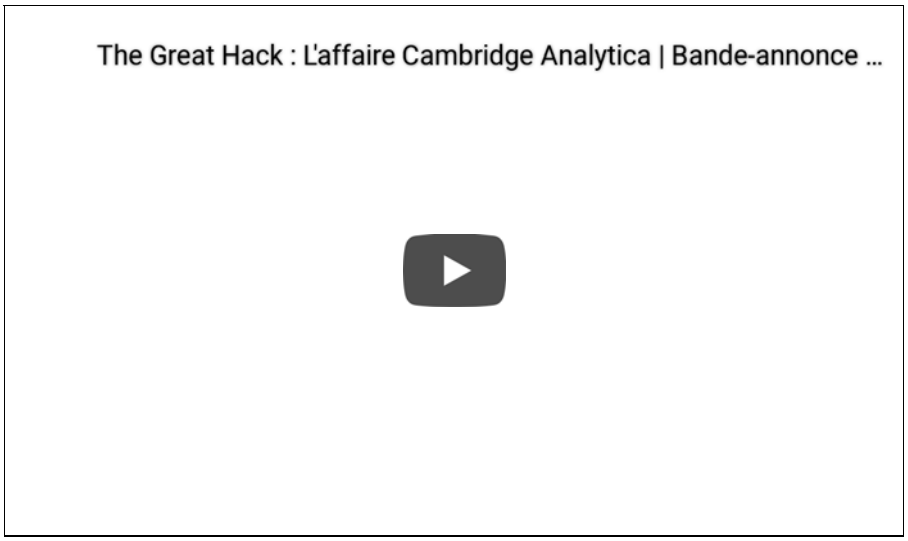
C'est parce qu'elle semble agir comme si elle était intelligente que nous projetons sur elle des caractéristiques qu'elle n'a pas, et que nous commençons à lui donner une place qu'elle ne devrait pas forcément occuper – ou du moins, pas seule.

Pourquoi l'absence de compréhension est-elle une limitation très importante pour nos utilisations de l'IA ?

Plus que de s'inquiéter de ce que les IA comprennent trop, il faudrait s'inquiéter de ce que les humains ne comprennent pas assez : les biais des données et des algorithmes, l'impact sur la société d'une gouvernementalité algorithmique, etc.

De plus, alors que les chatbots Eliza et Lenny « dupent » les humains, l'expérience du détournement du chatbot Tay de Microsoft montre ce qui se passe quand quelques humains peuvent retourner la situation. Ils comprennent comment la simulation d'intelligence fonctionne et utilisent cette compréhension contre elle pour manipuler l'IA, l'influencer et la dévier du but pour lequel elle était conçue, mais qu'elle ne comprend pas – la laissant incapable de comprendre son propre détournement (un prérequis pour le corriger).

Un autre danger actuel est celui d'un usage criminel de l'IA faible que ce soit en cybercriminalité ou en manipulation de masse sur des médias sociaux.



Le documentaire de Netflix sur l'affaire *Cambridge Analytica*.

Mais surtout, il est très dangereux actuellement de laisser des IA qui ne comprennent absolument pas ce qu'elles font aux commandes de systèmes dans lesquels les humains ne réalisent plus ce qui se passe, soit parce qu'ils n'ont pas su identifier les implications de leur système, soit parce que celui-ci commet des erreurs non détectées ou à une vitesse bien au-dessus de nos temps de réaction d'humains (par exemple, des algorithmes de « flash trading »).

À lire aussi : Peut-on faire confiance aux IA ?

Les « bulles de filtrage » qui se forment autour de nous lorsque nous ne voyons plus du Web que ce que nos applications nous recommandent, les suggestions racistes parce que personne n'avait détecté les biais des données, l'exposition à des nouvelles déprimantes pour vendre des voyages sont autant d'exemples de dérives de systèmes d'IA faible qui ne savent qu'optimiser leurs objectifs sans comprendre ce qu'ils font et, a fortiori, les dégâts qu'ils causent. Et ne parlons pas à nouveau des cas où leurs concepteurs sont mal intentionnés : ne comptez pas sur l'IA faible pour comprendre que ce qu'elle fait est mal.

Une combinaison particulièrement dangereuse actuellement est donc celle d'une IA faible (par exemple, un système de recommandation) déployée dans une application ayant énormément d'utilisateurs (par exemple, un grand réseau social) et prenant des décisions à une vitesse bien au-delà de notre temps de réaction d'humain. Car de la même façon que les intelligences artificielles faibles peuvent produire des résultats très impressionnants, une bêtise artificielle (même faible) peut produire des dégâts énormes.

Ainsi, c'est précisément parce que les IA ne comprennent pas que nous ne devons pas leur donner des responsabilités pour lesquelles il est vital de comprendre. C'est aussi pour cette raison que développer les capacités d'explication des IA est actuellement un enjeu majeur. Et, si les chercheurs utilisent parfois la métaphore de la chambre chinoise de Searle pour souligner le manque de compréhension de la machine, il semble aussi important que nous soyons vigilants à ce que la complexité, l'opacité et la vitesse des systèmes que nous concevons ne nous enferment pas à notre tour dans cette chambre d'incompréhension.

Benoît Tonson
Chef de rubrique Science + TC Junior

Vous aimerez aussi