Susanna Schellenberg fait dialoguer philosophie, neurosciences et IA

Susanna Schellenberg, professeure de philosophie et de sciences cognitives à l'université Rutgers, analyse comment la perception du monde qui nous entoure repose sur des capacités de discrimination, relevant des neurosciences. D'où la nécessité de développer une philosophie de la perception. Entretien à la frontière des sciences et de la phénoménologie.

Qu'est-ce qui est au fondement de la perception, et comment, selon ce fondement, obtenons-nous des connaissances de notre environnement ?

Je soutiens que la perception a pour fondement l'utilisation des capacités discriminatoires, qui sont des capacités à discriminer des objets, des propriétés et des événements dans notre environnement. Nous avons la capacité perceptive de distinguer les taches rouges des taches bleues, les notes aiguës des notes graves, les bruits forts des bruits faibles, l'odeur du fromage de l'odeur du café, et des millions d'autres choses. Pour percevoir une tasse à café rouge sur un bureau en bois, il faut distinguer la couleur rouge de la tasse à café de la couleur marron du bureau, la forme de la tasse des formes qui l'entourent, et bien d'autres discriminations de ce type. En employant de telles capacités, nous voyons la tasse à café.

En neurosciences, il est normal d'analyser la perception fondamentalement comme une question de discrimination. Mais en philosophie, cet aspect clé de la perception a à peine été pris en compte. J'essaie de changer cela.

Voir une tasse rouge, par exemple, consiste à utiliser des capacités discriminatoires, c'est-à-dire la capacité de distinguer le rouge des autres couleurs et la capacité de distinguer les formes de la tasse des autres formes.

Or, l'expérience peut nous égarer : notre perception est criblée de préjugés de toutes sortes et nous pouvons souffrir d'hallucinations et d'illusions. Ainsi, étant donné le manque de fiabilité de la perception, il est important de se demander comment la perception pourrait nous apporter des connaissances. D'un autre côté, il est clair que si une chose peut nous donner de la connaissance, c'est bien la perception.

Les capacités perceptuelles permettent de distinguer et d'isoler des objets, des propriétés et des événements d'un type spécifique. La capacité perceptive à discriminer et à isoler le rouge remplit sa fonction si elle est utilisée pour discriminer et isoler quelque chose de rouge dans l'environnement.

Ainsi, si je vois une tasse rouge en utilisant ma capacité perceptive à discriminer le rouge, alors la fonction de cette capacité est réalisée. Dans ce cas, j'acquiers la connaissance de la tasse. De plus, mon état perceptif me donne une justification pour croire qu'il y a une tasse rouge devant moi.

Qu'en est-il des cas d'hallucination et d'illusion ? Du point de vue du sujet qui en fait l'expérience, ils peuvent être indiscernables des cas de perception. Prenons le cas d'une personne qui hallucine une tasse rouge. Il lui semble donc qu'il y a une tasse rouge devant elle, même s'il n'y a pas de tasse rouge à cet endroit. Pour rendre le cas extrême, supposons que pour cette personne, l'hallucination est indiscernable d'une perception. Je soutiens que dans un tel cas, elle utilise les mêmes capacités perceptuelles que celles qu'elle utiliserait si elle percevait. Mais comme il n'y a pas de tasse rouge, elle ne parvient pas à distinguer une tasse rouge. Ainsi, les capacités qu'elle emploie ne remplissent pas leur fonction. En raison de cet échec, l'hallucination ne lui permet pas de connaître son environnement. Et elle ne saurait même pas qu'elle ne sait pas.

Dans les cas de perception, d'hallucination et d'illusion, le même mécanisme est activé dans notre esprit : l'utilisation de capacités discriminatoires. Cela explique pourquoi, de notre point de vue, ces expériences peuvent sembler exactement identiques. Cependant, il existe des différences. Dans la perception, les capacités employées remplissent leur fonction : nous discriminons ce qui nous semble présent. Dans les cas d'hallucination et d'illusion, elles ne le font pas : nous ne discriminons pas ce qui nous semble être présent. Les capacités perceptuelles sont donc à la base d'une théorie unifiée de la perception.

Dans ce cas, peut-on aussi considérer les intelligences artificielles combinées à des capteurs comme des systèmes perceptifs ?

Je pense que les systèmes d'intelligence artificielle combinés à des capteurs sont une sorte de systèmes perceptifs. Comme nos systèmes perceptifs, ils absorbent les informations de leur environnement par des mécanismes de discrimination. S'il existe des différences significatives dans les mises en œuvre physiques entre les systèmes d'IA et les êtres humains, le mécanisme sous-jacent de discrimination est le même.

Nous sommes tous d'accord sur le fait qu'un humain ayant un implant auditif est toujours un humain. En effet, en ce qui concerne son mécanisme de discrimination, les implants auditifs sont similaires à notre système auditif.

Si cela est vrai, alors au moins en ce qui concerne la perception, nous ne sommes pas si spéciaux. Cela ne devrait pas trop nous déranger. Après tout,

la science nous dit que nous ne sommes pas différents en nature des autres animaux en ce qui concerne la perception. La perception est un mécanisme de base que nous partageons avec d'autres animaux et, comme je le dis, avec les systèmes d'IA. Il y a de nombreux aspects pour lesquels nous sommes catégoriquement différents des systèmes d'IA, par exemple en ce qui concerne notre capacité de créativité et notre capacité à ressentir des émotions - les nôtres et celles des autres.

Pour mettre en évidence les différences et les similitudes entre les systèmes d'IA et les humains, imaginons maintenant une humaine avec toujours plus d'implants. Y a-t-il un moment où elle relèvera plus du robot que de l'être humain ? Considérons un cas extrême dans lequel chaque cellule du corps a été remplacée par un implant. Nous pourrions alors supposer qu'elle est maintenant une réplique physique d'un robot créé dans un laboratoire informatique. Ce sont des problèmes profonds. Mais ce que je veux simplement dire ici, c'est qu'en matière de perception, nous ne sommes pas si différents des autres animaux et des systèmes d'IA.

Une conscience artificielle serait-elle possible ?

Si vous considérez que la conscience perceptuelle est constituée par l'utilisation de capacités perceptuelles et que, d'après la question précédente, les systèmes d'IA avec capteurs peuvent être considérés comme des systèmes perceptuels, alors oui, l'IA pourrait avoir une conscience perceptive.

J'ai toujours été un peu surprise par l'intérêt intense que suscite la conscience. Je suis plus intéressée par la façon dont notre esprit accomplit les choses étonnantes qu'il réalise, que par le caractère conscient ou non du processus. Nos systèmes perceptifs traitent une quantité étonnante d'informations en une fraction de seconde. Environ 50% du cerveau humain est consacré au traitement visuel. Cette répartition de nos ressources témoigne de la complexité de la tâche à accomplir. Une quantité infime d'informations traitées remonte jusqu'au niveau conscient. Ainsi, une grande partie de l'information est utilisée pour guider notre action et est disponible pour notre système cognitif sans jamais atteindre le niveau conscient.

Une chose est sûre : nous sommes loin de développer des systèmes d'IA conscients. Il a été prouvé récemment que l'apprentissage machine stagne et, malgré d'énormes efforts, nous avons du mal à développer des systèmes d'IA ayant les capacités linguistiques des enfants d'âge préscolaire. Nous ne devrions pas tant nous inquiéter de la singularité[1] ou de l'IA qui devient

[1] Wikipédia : « La singularité technologique - aussi, tout simplement, la singularité - est un moment hypothétique où la croissance technologique devient incontrôlable et irréversible, entraînant des changements imprévisibles dans la civilisation humaine. » (traduit de l'anglais)

consciente ou des robots tueurs. Ce qui devrait nous inquiéter, ce sont les algorithmes biaisés. Ils sont là, maintenant. Ils ont d'énormes implications dans nos vies.

En parlant d'algorithmes biaisés, avec des capteurs et des algorithmes de calcul toujours meilleurs, comment se fait-il qu'il y ait encore des biais ?
Tout d'abord, il est important de noter que tous les systèmes de reconnaissance complexes sont truffés de biais. L'esprit humain et l'IA exploitent tous deux d'énormes quantités de données, et chaque fois qu'il y a un décalage entre la quantité en entrée et la puissance de traitement disponible, il faut simplifier, et par conséquent certaines informations sont perdues. Ce processus génère des biais.

Il est important de reconnaître que certains biais ne posent pas de problème. Ils rendent ces systèmes plus efficaces. Par exemple, le système perceptuel humain a un biais selon lequel la lumière vient d'en haut, et que les objets en mouvement sont solides : nous nous baissons donc parfois à l'approche d'une ombre en mouvement, même si ce n'est que du vent.

Cependant, certains biais sont profondément problématiques et peuvent être très préjudiciables pour les groupes marginalisés et privés de leurs droits. Comme tous les systèmes complexes de reconnaissance comportent des biais et qu'au moins certains de ces biais ne posent pas de problème, un grand défi consiste à différencier les biais problématiques des biais non problématiques. Une approche possible consiste à examiner le résultat. Si le biais est préjudiciable à un groupe de personnes, alors l'algorithme doit être corrigé pour éliminer le biais. S'il n'est pas nuisible, il peut être laissé.

C'est l'approche actuelle pour traiter les biais algorithmiques. Bien qu'elle ait sa place, je pense que cette approche en fait trop peu, trop tard. J'étudie actuellement les moyens d'éliminer les biais plus en amont, c'est-à-dire les moyens de développer des algorithmes pour que les biais néfastes n'apparaissent pas en premier lieu.

Un autre élément est que, contrairement à ce que l'on suppose généralement, la plupart des biais algorithmiques ne sont pas des "biais descendants". Ils ne découlent pas des croyances et des points de vue du programmeur et de la manière dont ces croyances et points de vue affectent les choix qu'il a faits lors de la conception de l'algorithme. Il est certain que de tels biais descendants existent et qu'ils constituent un gros problème. Cependant, les "biais ascendants", non seulement dans l'IA mais aussi dans l'esprit humain, posent un problème tout aussi important. Ces types de biais proviennent des données entrantes, de leur traitement au niveau le plus bas, des modèles que

le système détecte dans les données, et des classifications et corrélations qu'il établit.[2]

Comment la philosophie de la perception et les sciences dites "dures" peuvent-elles engager un dialogue constructif ?

La philosophie a toujours été étroitement liée aux autres sciences.
Historiquement, trois grandes questions ont motivé la recherche sur la perception. Premièrement, comment la perception justifie-t-elle les croyances et permet-elle de connaître notre environnement ? Cette question a été traitée presque exclusivement en philosophie. Deuxièmement, comment la perception engendre-t-elle des états mentaux conscients ? Cette question a également été abordée principalement en philosophie. En effet, à quelques exceptions près, les neuroscientifiques affirment que nous sommes loin de comprendre ce que signifie être conscient en termes de mécanismes cérébraux. Troisièmement, comment un système perceptuel réussit-il à convertir des informations variables en représentations mentales de caractéristiques invariantes de notre environnement ? Cette dernière question a surtout été abordée dans les domaines des neurosciences, de la psychologie cognitive et de la psychophysique.

Une hypothèse centrale de mes recherches est que les réponses à ces trois questions ne sont pas indépendantes les unes des autres et que pour progresser dans la compréhension de la nature de la perception, nous devons l'étudier de manière plus intégrée. D'où le titre de mon dernier livre "The Unity of Perception" dans lequel je développe un ensemble d'outils unifiés pour répondre aux trois questions dans une seule théorie qui est conceptuellement disciplinée et empiriquement contrainte.
Dans l'autre sens, les neurosciences constituent un domaine très nouveau qui peut bénéficier de ce que les philosophes savent faire de mieux, à savoir : analyser les nouveaux concepts que les neurosciences génèrent et les articuler.

Ce sera le sujet de votre prochain livre ?

Je travaille à exposer toutes les différentes façons dont les systèmes d'IA tout comme le système perceptif et cognitif humain peuvent être biaisés. Voici quelques distinctions essentielles :
Un système peut être biaisé si les données entrantes sont biaisées. Une deuxième explication est due à la façon dont les caractéristiques des données entrantes sont liées et classées lors de l'étape de traitement. Une troisième est due à la façon dont la sortie est interprétée. Une autre distinction clé est

[2] Pour en savoir plus sur l'analyse de Susanna Schellenberg sur les biais des algorithmes (en anglais) https://www.newstatesman.com/science-tech/2020/04/how-biased-algorithms-perpetuate-inequality

celle que j'ai mentionnée précédemment entre les biais descendants et les biais ascendants : les biais ascendants sont sous-explorés et doivent être mieux compris. Une autre distinction essentielle est celle qui existe entre les biais dus à l'échantillon d'entraînement et les biais liés à la manière dont les caractéristiques des données sont reliées entre elles.

Le biais lié à l'échantillon d'entraînement est dû aux biais dans les données avec lesquelles un algorithme est entraîné. Le logiciel de reconnaissance vocale de Google en est un exemple. Au départ, il fonctionnait beaucoup mieux sur les voix masculines que sur les voix féminines. Il s'est avéré que la raison en était qu'il était principalement entraîné sur des voix masculines. Il n'est donc pas surprenant qu'il fonctionne beaucoup mieux dans la gamme d'une voix masculine typique. Ce type de biais est facile à corriger. Il suffit d'exposer l'algorithme à un grand nombre de voix féminines. Ce type de biais est également facile à éviter. Il suffit de choisir des échantillons d'entraînement non biaisés. Mais bien sûr, il y aura toujours des choix difficiles à faire lors de la sélection des échantillons d'entraînement. La reconnaissance vocale de Google fonctionne très mal pour les accents écossais. Étant donné que peu d'anglophones ont des accents écossais et qu'il existe de nombreux types d'accents écossais, des choix difficiles doivent être faits quant à l'effort à fournir pour que les systèmes de reconnaissance vocale fonctionnent pour les locuteurs écossais.

Quelle que soit la manière dont ce problème sera résolu, les retombées négatives ne sont évidemment pas au même niveau que les retombées des biais dans les algorithmes utilisés pour les condamnations pénales, les libérations conditionnelles, les demandes d'emploi, les demandes de prêt, les soins de santé et les systèmes de génération de publicité. Les biais dans ces applications sont principalement des biais liés à la manière dont les caractéristiques des données sont reliées entre elles. Ils sont beaucoup plus difficiles à corriger.

Voici un exemple de ce type de biais : si l'on cherche sur Google un nom à consonance afro-américaine, on a plus de chances d'obtenir une publicité pour une vérification d'antécédents criminels que si l'on cherche sur Google un nom généralement donné à des bébés européens. La raison en est que l'AdSense, l'algorithme publicitaire de Google, a détecté un schéma selon lequel les gens étaient plus susceptibles de faire une vérification d'antécédents criminels après avoir googlé un nom si ce nom était typique de la communauté afro-américaine. Il a ensuite généré des publicités en conséquence. Il s'agit d'un biais très préjudiciable, car il ne se contente pas de perpétuer les biais existants dans notre société, il les amplifie.
Compte tenu des biais algorithmiques, nous avons de bonnes raisons d'être sceptiques quant à une plus grande objectivité supposée des

décisions générées par ordinateur comparée aux décisions humaines ! Je pense qu'il reste beaucoup de travail à faire pour traiter toutes ces questions.

Propos recueillis par Lauriane Gorce, directrice scientifique de l'Institut de la technologie pour l'humain - Montréal