UC Berkeley annonce la création du Center for Human-Compatible Artificial Intelligence

Publié le vendredi 2 septembre 2016
Voir en ligne : https://www.france-science.org/UC-Berkeley-annonce-la-creation-du.html

La recherche en intelligence artificielle (IA) s'est très largement centrée depuis plus de dix ans sur les questions autour de la conception et du déploiement opérationnel d'agents intelligents, ces systèmes qui perçoivent et agissent dans un environnement donné. Dans un tel contexte, l''intelligence' est perçue comme liée à des notions statistiques et économiques de rationalité, c'est-à-dire à la capacité à prendre des décisions, à élaborer des plans ou à déduire des inférences justes. L'adoption de représentations probabilistes et de méthodes d'apprentissage statistique a mécaniquement conduit à une intégration et une fertilisation croisée forte des recherches en IA avec celles en apprentissage automatique, statistiques, théorie du contrôle, statistiques ou neurosciences. L'établissement de cadres théoriques partagés, combiné avec la disponibilité exponentielle de données et de puissance de calcul, a généré des succès remarquables dans la mise en performance de tâches comme la reconnaissance de la parole, les véhicules autonomes, la classification d'images ou la motricité articulée.

Naturellement, de telles avancées ne vont pas sans charrier leur lot de questionnement sociétaux. La forme la plus communément exprimée en est sans doute la version apocalyptique –une vision dystopique de l'émergence de machines intelligentes potentiellement capables de se retourner contre leurs créateurs-, version d'autant plus prégnante qu'elle est à mots couverts soutenue par de grandes figures publiques de la science [1] et de l'industrie [2].

De manière plus immédiate, une autre série de problèmes réside dans la distinction trop souvent négligée entre la capacité accrue à prendre des décisions et la capacité à prendre de "meilleures" décisions. En d'autres termes, si la fonction d'utilité de la machine n'est pas alignée avec les valeurs humaines (quelles que soient les difficultés à caractériser celles-ci), le résultat peut être spectaculairement inopérant quelles que soient l'excellence de la maximisation algorithmique ou de la modélisation du monde (exemples séminaux de la machine à fabriquer des trombones [3], ou du feu rouge [4].).

Une des possibilités de s'attaquer au problème de l'alignement des valeurs réside dans l'élaboration d'un cadre d'apprentissage par renforcement inverse (ARI). Il s'agit de l'apprentissage d'une fonction récompensée dans un processus de décision markovien par l'observation d'un autre agent –humain, en l'espèce- dont on suppose qu'il agit en concordance avec une telle fonction. Ce n'est pas un problème simple compte tenu de l'inconsistance, de l'irrationalité et de la variabilité des valeurs humaines.

C'est dans cette optique que l'Université de Californie à Berkeley vient d'annoncer le 29 août dernier le lancement d'un nouveau centre de recherche, le *Center for Human-Compatible Artificial Intelligence*. Ce centre vient de recevoir une subvention initiale de 5,5 millions de dollars de la part de l'*Open Philantrophy Project*, qui est un instrument conjoint de GiveWell et de Good Ventures, la fondation philantropique de Cari Tuna et Dustin Moskowitz (co-fondateur de Facebook). Il n'est pas inintéressant de noter que cette donation initiale devrait être ultérieurement abondée par le *Future of Life Institute* (FLI), lequel compte parmi ses conseillers scientifiques Elon Musk et Stephen Hawking.

Le Centre sera dirigé par Stuart Russell [5], Professeur d'informatique et d'ingénierie à UC Berkeley (il a été titulaire de la Chaire Blaise Pascal et de la Haute Chaire d'Excellence de l'Agence Nationale pour la Recherche), et co-auteur du manuel de référence dans le domaine de l'intelligence artificielle (*Artificial Intelligence : a Modern Approach*, co-rédigé avec Peter Nordvig et utilisé par plus de 1300 universités dans 116 pays). Il y sera rejoint par des spécialistes d'intelligence artificielle, de neurosciences et d'informatique d'UC Berkeley, mais aussi de l'Université Cornell et de l'Université du Michigan.

Rédacteur :
- Olivier Tomat, Expert Technique International, San Francisco, olivier.tomat@ambascience-usa.org ;
- Retrouvez l'actualité en Californie du Nord sur http://sf.france-science.org ;

Notes

[1] par exemple Stephen Hawking : http://www.independent.co.uk/life-style/gadgets-and-tech/news/stephen-hawking-artificial-intelligence-could-wipe-out-humanity-when-it-gets-too-clever-as-humans-a6686496.html

[2] par exemple, Elon Musk : https://twitter.com/elonmusk/status/495759307346952192

[3] si le but unique d'une machine est de maximiser le nombre de trombones qu'elle fabrique, elle pourrait être tentée de contourner l'homme et trouver des technologies capables de convertir l'intégralité de la masse de l'univers en trombones ; il s'agit d'un exemple du philosophe suédois Nick Bostrom

[4] l'illustration est fournie par Stuart Russell -cf. infra- : Quelqu'un qui construit une voiture autonome pourrait lui ordonner de ne jamais passer au feu rouge. Cependant, la machine pourrait s'introduire dans le système de contrôle des feux de circulation de manière à ce que tous les feux passent au vert. Le véhicule obéit aux ordres, mais d'une manière non désirée et/ou non attendue

[5] dont la prééminence dans le champ est en grande partie due à sa capacité à mobiliser différents répertoires de publication, entre lettres ouvertes à destination des institutions (par exemple, Stuart Russell, Tom Dietterich, Eric Horvitz, Bart Selman, Francesca Rossi, Demis Hassabis, Shane Legg, Mustafa Suleyman, Dileep George, and Scott Phoenix, *Research Priorities for Robust and Beneficial Artificial Intelligence : An Open Letter, AI Magazine*, Vol. 36, No. 4, 2015), à proposer des interventions à destination du grand public (par exemple, Stuart Russell, *Moral Philosophy Will Become Part of the Tech Industry*, Time, September 15, 2015), ou à publier des contributions scientifiques 'pures' (par exemple, F. Lieder, D. Plunkett, J. Hamrick, S. Russell, N. Hay, T. Griffiths, *Algorithm selection by rational metareasoning as a model of human strategy selection*. In **Advances in Neural Information Processing Systems 23**, MIT Press, 2015)