Being Stuart Russell – Le retour de la philosophie morale

Catégorie : **IA & Neurosciences**
Tags : **philosophie, éthique, machine, algorithme**
Personnages : **Stuart Russell, Daniel Dewey, Max Tegmark, Joshua Greene**

30 septembre 2017

L'horizon des machines « intelligentes » étant l'autonomie, la question de leur « morale » est posée. Suivons Stuart Russell, chercheur en IA, sur ce terrain.

Stuart Russell


Stuart Russell

Stuart Russell est un chercheur en Intelligence Artificielle (IA) multi-honoré. Il est aujourd'hui membre du « *Future of Life Institute* » et du « *Centre for the Study of Existential Risk* » à Cambridge (Un futur sans nous). Cet éminent spécialiste a été interrogé en 2015 par l'excellent en ligne Quanta Magazine pour évoquer la question suivante : comment s'assurer que nos machines devenues « intelligentes », donc plus ou moins autonomes, se

conforment à nos « valeurs humaines »[1] ? Nous essayons ici de comprendre comment un chercheur en IA envisage certains aspects de cette très vaste question, de se mettre en quelque sorte à sa place : being Stuart Russell, pour voir… Et pour constater que cette interrogation doit d'abord être la nôtre.

Stuart Russell est l'un des nombreux cosignataires d'une lettre ouverte publiée sur le site du Future of Life Institute (FLI) en 2015, invitant les chercheurs à se préoccuper de développer des systèmes d'IA « vertueux ». Cet appel guidera la conférence d'Asilomar de 2017 et conduira à l'élaboration des « principes » du même nom.

Dans cette lettre ouverte, Stuart Russell écrit notamment :

> Nos systèmes d'IA doivent faire ce qu'on leur demande de faire.

Cela nous paraît plutôt évident, mais pour un chercheur en IA, cela ne va pas de soi. Russell prétend que, si l'on n'y prend pas garde, l'augmentation du niveau d'intelligence et d'autonomie de ces systèmes pourrait les conduire à faire tout autre chose que ce pour quoi ils étaient prévus. Il faudrait donc au moins les concevoir en tenant compte en amont d'un « bénéfice pour l'humanité » mais surtout d'un « bénéfice garanti ». C'est la notion de *robustness*, c'est-à-dire de conformité permanente de nos systèmes aux objectifs initialement prévus.

Cet appel est largement inspiré d'un petit document rédigé par Stuart Russell, Daniel Dewey et Max Tegmark, ce dernier étant au passage un cofondateur du FLI et le directeur scientifique du FQXi (le « *Foundational Questions Institute* » financé à l'origine par la fondation John Templeton…). Dans ce document[2], on retrouve quelques préoccupations désormais classiques concernant le marché de l'emploi, les « disruptions » de modèles, ceux de la finance par exemple, mais aussi quelques ouvertures concernant la pertinence de nos indicateurs économiques. On trouve également la mention de recherches qu'il faudrait mener en matière d'éthique et de réglementation concernant par exemple les véhicules autonomes, les armes intelligentes, la protection des données personnelles etc.

Bref, nous étions, voici 3 ans, dans l'un de ces moments particuliers de l'histoire des sciences et des techniques où les chercheurs, depuis leur propre domaine, soulèvent des questions très générales et encore un peu confuses d'application de leurs découvertes dans le champ économique et social et donc des questions d'éthique. Mais comment abordent-ils ces sujets ? Il nous semble important de le comprendre dans la mesure où ces mêmes chercheurs orientent aujourd'hui toute la réflexion politique en la matière.

Les « valeurs humaines » ?

> **Quanta Magazine** : vous prétendez que l'objectif de votre champ de recherche devrait être de développer une intelligence artificielle « alignée de façon

[1] Natalie Wolchover pour Quanta Magazine – 21 avril 2015 – *Concerns of an Artificial Intelligence Pioneer*

[2] Stuart Russell, Daniel Dewey, Max Tegmark – 2015 – *Research Priorities for Robust and Beneficial Artificial Intelligence*

> démontrable » [provably aligned] sur les valeurs humaines [human values]. Qu'entendez-vous par là ?
>
> **Stuart Russell** : c'est une affirmation délibérément provocatrice dans la mesure où elle met en relation deux choses qui paraissent incompatibles : les valeurs humaines et la notion de preuve. Les valeurs humaines pourraient rester à jamais quelque chose de mystérieux. Mais dans la mesure où nos valeurs se révèlent dans notre comportement, on peut espérer démontrer que la machine peut grosso modo s'y conformer [get most of it].

Ces propos peuvent sembler curieux à ceux qui ne sont pas familiers avec l'IA, en particulier l'étrange facilité avec laquelle Russell évoque les « valeurs humaines », comme s'il s'agissait de caractéristiques ou d'attributs bien définis. Qu'entend-t-il donc par « valeurs humaines » ? Nous y revenons plus loin.

Pour Russell, le principe d'acquisition de ces « valeurs » par une machine ne semble pas présenter trop de difficultés : il suffirait que la machine les induise, les déduise ou les imite (on ne sait pas trop) plus ou moins explicitement à partir de notre « comportement ». Il faut rappeler ici que le paradigme de l'apprentissage est au fondement de la plupart des solutions d'IA. La machine apprend progressivement, à partir d'exemples qui lui sont présentés, à produire la meilleure « réaction » à une situation donnée : reconnaître un visage, effectuer un mouvement au jeu de go, etc.

Il n'est par conséquent pas étonnant que, selon Russell, nos valeurs étant révélées par notre comportement, elles puissent être acquises par une machine apprenant à reproduire ce comportement. En analysant l'état de la machine après cet apprentissage, il serait alors possible de prouver que nos valeurs ont bien été acquises.

Nous comprenons bien le raisonnement mais il repose sur deux prémisses assez discutables.

Tout d'abord nos valeurs pourraient se déduire intégralement de nos comportements. Supposons que nous disposions d'une capacité d'apprentissage quasi infinie et que puissions ainsi analyser tous les comportements de tout le monde, lire tous les livres, visionner tous les films... Pourrions-nous en déduire ce que nous devons faire en toute circonstance du point de vue des « valeurs », mais aussi ce que nous ne devons pas faire ? C'est ce que semble penser Stuart Russell. Nous vous laissons juges...

Deuxièmement, nos valeurs devraient être assez stables et universelles. Or, il est loin d'être évident que nous les partagions tous ni qu'elles soient immuables.

Les « valeurs humaines » !

Passons maintenant à la technique. Nous allons mieux comprendre ce que Russell entend par « valeurs humaines » (nous soulignons) :

> **Stuart Russell** : Comment une machine peut-elle se doter d'une bonne approximation des valeurs que les hommes souhaiteraient lui transmettre ? Je pense qu'une solution possible est la technique que l'on appelle « apprentissage

> par renforcement inverse ». L'apprentissage par renforcement classique consiste à vous donner des punitions ou des récompenses selon votre comportement, et votre objectif est alors de vous comporter de façon à obtenir le plus de récompenses. Dans l'apprentissage par renforcement inverse, vous cherchez à comprendre quel « score » le comportement que vous observez cherche à maximiser. Par exemple, votre robot domestique vous voit sortir du lit le matin, broyer des choses marron dans une machine bruyante et faire des choses avec de la vapeur, de l'eau chaude et du lait, etc. et alors vous semblez content [bon score]. Elle apprendrait ainsi qu'une partie de la fonction de valeur humaine [human value function] du matin est d'avoir du café.

Si le plaisir d'avoir du café est une « valeur humaine », alors par « valeur humaine » il faut entendre « bénéfice » au sens le plus général (sans être pour autant « liberté, égalité, fraternité »...). Mais surtout, « approximer des valeurs », utiliser le terme de « fonction de valeur » c'est faire entrer les « valeurs humaines » dans le champ numérique.

On pense inévitablement à un autre concept, économique celui-là : *l'utilité*[3]. L'utilité représente la satisfaction éprouvée par quelqu'un de la consommation d'un bien ou d'un service. La satisfaction, le bien-être, sont des concepts éminemment abstraits et les économistes *mesurent* donc l'utilité à partir de préférences observées (on demande par exemple à des échantillons de consommateurs de classer des articles par « préférence »). La valeur de Russell et l'utilité des économistes fonctionnent exactement de la même façon : elles numérisent et mathématisent des concepts abstraits pour les rendre manipulables par un système ou un algorithme. L'utilité d'un bien, la valeur d'un comportement se déduisent par échantillonnage et apprentissage.

Mais cette analogie a aussi des limites. Car l'utilité est un concept à valeur *positive* : en économie, il n'y a pas d'utilité zéro, encore moins négative. Consommer un bien ou un service c'est toujours mieux que de s'en abstenir. En effet, l'utilité est la mesure de ce que la consommation nous fait à nous-mêmes, indépendamment de ce que cela occasionne aux autres. Or, la « valeur humaine », si elle est mesurée, quantifie l'action au regard de ce qu'elle produit en nous *mais aussi chez les autres*. Elle ne peut pas être la mesure intrinsèque et absolue d'une action donnée mais doit prendre en compte le contexte. La « fonction de valeur » évoquée par Russell est, de ce point de vue, inopérante.

Si l'on veut poursuivre l'analogie avec l'économie, tout en préservant une approche plus conséquentialiste, il faudrait peut-être plutôt chercher du côté de la notion *d'externalité*. A creuser…

Retenons pour le moment ceci : lorsque Russell et ses pairs évoquent les « valeurs humaines », ils les conçoivent, volontairement ou pas, comme constitutives d'une *économie des valeurs*, c'est-à-dire d'une mathématique.

[3] Wikipédia – *Utilité (économie)*

La main invisible

Russell est optimiste. Nos machines sont désormais capables d'apprendre et il leur suffit d'observer nos comportements pour acquérir nos valeurs. Ces valeurs ne seraient rien d'autre que des mesures de situations, donc accessibles à l'algorithmique. Mais il y a au moins une autre raison d'être confiant :

> **Stuart Russell** : Si vous voulez un robot domestique chez vous, il doit partager un large sous-ensemble de valeurs humaines [pretty good cross-section of human values], sinon il pourrait faire des choses idiotes, comme de mettre le chat dans le four pour le dîner parce qu'il n'y a plus rien à manger dans le frigo et que les enfants ont faim. […] Il y a un énorme intérêt économique à ce que tout se passe bien : il suffit d'un ou deux événements comme un robot qui met un chat dans un four pour que les gens perdent confiance et n'en veuille pas.

Le vocabulaire est toujours un peu troublant, mais « being Stuart Russell » devient déjà plus facile. Par « *large sous-ensemble de valeurs humaines* », il faut comprendre que l'on a inculqué à notre robot domestique, par une technique d'apprentissage, la capacité à mesurer un grand nombre de situations *mais pas toutes*. Il manque par exemple « mettre le chat dans le four ». Mais dans ce cas, la main invisible intervient : le robot ne risque pas de trouver preneur.

Retenons que le concepteur d'une IA dispose donc de ce dernier garde-fou : si sa création ne respecte pas les valeurs humaines, elle sera naturellement rejetée. Ce sera très probablement le cas en général, mais la société n'est pas exempte d'individus étranges, pour lesquels « mettre le chat dans le four » reste une possibilité acceptable...

Quid des « vraies » IA ?

Stuart Russell trace un chemin conduisant à des techniques de respect des valeurs humaines. Il préempte les inquiétudes que nous pourrions avoir avant que nous les exprimions à notre manière.

Mais au fond, même si ces questions sont loin d'être résolues, nous partageons le sentiment de Stuart Russell qu'il n'y a pas de difficulté essentielle dès lors que nous parlons d'IA « faible ».

En revanche tout chercheur en IA partage le même horizon problématique (dont nul ne sait, à part quelques gourous, s'il peut être atteint) : une machine capable de réflexivité, de s'améliorer seule, plus vite et bien mieux que n'aurions pu le faire, au point de nous dépasser.

> **Stuart Russell** : Pouvons-nous démontrer que nos machines ne pourront jamais, quel que soit leur niveau d'intelligence, outrepasser les buts que nous, les hommes, leur avons fixé ? […] c'est un véritable problème lorsque ces machines ont une capacité d'action dans le monde réel. […] Elles pourraient se réécrire elles-mêmes à partir de leur programme existant et de leurs expériences dans le monde réel. Quel sont les effets possibles sur ce redesign de l'interaction avec le monde réel ? C'est ce que nous ignorons.

Le chercheur en IA atteint ici une limite qu'il ne pourra jamais étudier seul. Retenons que, quand nous entendons Elon Musk pousser des cris d'alarme tous les 15 jours, au-delà qu'il le fait dans son propre intérêt, c'est cette limite qu'il évoque, cet horizon de la machine qui se reprogramme dans le monde réel. Comme un être vivant...

Retour de la philosophie morale

Voici donc comment un chercheur en IA nous parle aujourd'hui des « valeurs humaines » et de la façon de les faire respecter par des machines « intelligentes ». Les articles à ce sujet vont foisonner au fur et à mesure que les techniques d'IA pénétreront notre quotidien, et tous les Stuart Russell seront sommés de s'exprimer. Autant essayer de les comprendre...

Il est clair que certaines injonctions morales, l'injonction « zéro » étant « tu ne tueras point », restent à être interprétées au cas par cas (« tu ne tueras point » sauf en cas de légitime défense...). La morale n'est pas figée : elle s'ajuste à nos progrès et se sédimente par jurisprudence. Le monde modifié par les machines nous posera inévitablement de nouvelles questions morales, nous invitera à proposer de nouvelles « valeurs humaines », qui ne sont pas encore présentes dans les films et les livres existants.

Mais déjà, la poussée de l'IA nous oblige à nous interroger sur nos valeurs, sur ce qu'elles sont, sur la façon dont les acquérons, sur ce que nous faisons lorsqu'elles sont bafouées, etc. Il n'y a aucun doute : ainsi que l'écrit Joshua Greene, psychologue à l'université de Harvard :

> Avant que nous puissions programmer nos valeurs morales dans des machines, nous devons nous efforcer de les clarifier et de les rendre cohérentes, ce qui pourrait être l'heure de vérité pour la philosophie morale du 21ème siècle.

Même « Being Stuart Russell », nous sommes entièrement d'accord.