Qu'est-ce qui échappe à l'intelligence artificielle ?
Les limites de la rationalité calculatoire : épistémologie et politique

Appel à communication - Colloque international
Organisé par François Levin (École polytechnique - philosophie) & Etienne Ollion (CNRS - sociologie)

Paris - 18 et 19 Juin 2020

Depuis plusieurs années, l'intelligence artificielle a fait l'objet d'analyses qui insistent sur son caractère contraignant. Elles peuvent viser à démontrer, à partir d'une relecture des textes cybernétiques, "l'essence totalitaire" du dispositif technologique, qui réduit les individus à devenir des éléments formatés et programmés, objets d'un calcul universel (Vioulac, 2018). L'analyse peut aussi porter sur les effets d'individuation nouveaux induits par les algorithmes, que ce soit via la production d'une nouvelle forme d'individualité constituée par "la société de ciblage" (Chamayou, 2015) ou encore via la mise en valeur des effets de "court-circuitage" algorithmique de l'attention et des volontés individuelles (Citton, 2017). Plusieurs travaux ont opéré une reprise conceptuelle de la notion foucaldienne de gouvernementalité pour l'appliquer aux usages algorithmiques de la statistique et aux nouvelles formes de normativité qui s'y expriment (Rouvroy et Berns, 2013). Les effets destructeurs des algorithmes sur les savoirs et les désirs ont enfin fait l'objet d'investigations, par exemple autour du concept de « société automatique » (Stiegler, 2015).

Analysée comme l'aboutissement de la rationalité calculatoire, l'IA est largement pensée sur le registre du calcul et de ce qu'il produit. Ce colloque interdisciplinaire (philosophie, sciences humaines et sociales, informatique, statistiques, mais aussi art*)* se propose de revenir sur ce présupposé. Par contraste avec ces travaux, son est de réfléchir sur les formes et sur les limites de ces calculs. En d'autres termes, il s'agit de s'interroger sur ce qui, dans l'intelligence artificielle, échappe au calcul.

Plusieurs pistes pourront alors être explorées :
Des communications pourront porter sur **ce que l'IA ne parvient pas, ou ne parvient qu'imparfaitement, à calculer**. Ainsi, pour les voitures autonomes, ce qui relève de la négociation non-verbale entre conducteurs.trices échappe encore largement aux algorithmes ; plus généralement l'impossibilité pour l'IA de saisir le contexte (Dreyfus, 1992), ou encore l'ensemble des "réflexes sémantiques" acquis durant l'expérience d'une vie (French, 1990) et même, plus globalement, les valeurs sémantique et non uniquement syntaxiques (Searle, 1980) sont des critiques traditionnelles qui sont faites aux algorithmes.Ce faisant, si les algorithmes d'intelligence artificielle relèvent bien d'une rationalité calculatoire, toute une partie du monde leur échapperait - soit qu'elle n'est pas mise en donnée, soit qu'elle n'est pas calculable. Le non-calculable constituerait alors la/une borne de l'intelligence artificielle. Dans cette perspective, les discours (enthousiastes ou inquiets) sur l'intelligence artificielle surestimeraient son champ d'application possible et donc la capacité à produire les effets qui lui sont attribuées. Une telle hypothèse, qui devrait être étayée, ouvre alors vers d'autres interrogations. Comment définir ce non-calculable, entendu comme limite ? L'est-il circonstanciellement - c'est-à-dire qu'il pourrait être sans cesse repoussé par les avancées en intelligence artificielle - ou absolument ? Peut-il être défini de manière normative (la loi informatique et libertés dispose ainsi qu'une décision administrative faisant grief ne peut être prise de manière exclusivement automatisée) et suivant quels critères ? Est-il

constitué positivement, comme une classe d'objets définis, ou bien négativement, simplement comme la conséquence des limites des dispositifs techniques (erreurs, bugs, puissance limitée) ? Poser ainsi la question permet de réinscrire cette interrogation dans celles sur la non-calculabilité qui a gouverné les mathématiques depuis leur fondements (Gödel, 1931 ; Turing, 1936 ; Chaitin, 2004).

À cette limite externe s'en ajoute une autre, cette fois liée à la méthode et que l'on peut qualifier d'interne. La critique classique du caractère "opaque" des algorithmes d'intelligence artificielle pourra ici être évoquée. Dans quelle mesure la critique de l'inexplicabilité, ou celle de l'absence de démonstration exacte (Boelaert et Ollion, 2018 ; Schubbach, 2019), est-elle toujours d'actualité ? En quoi cette inexplicabilité renvoie-t-elle à ce qui, dans l'IA, échappe au calcul ? Les développements récents destinés à développer l'explicabilité de l'IA, ceux destinés à favoriser l'identification causale (Athey, 2017) offrent-ils des moyens "d'ouvrir la boîte noire de l'IA", et si oui à quelles conditions, et pour quels résultats? Des communications pourront être proposées dans ce sens, qui pourront aussi rappeler certains des débats relatifs à l'histoire du domaine, où la question de l'explicabilité, et de la calculabilité, ont été centraux - par exemple dans les querelles entre les paradigmes connexionnistes et symboliques (Crevier, 1997; Cardon *et al*., 2018). Des présentations relatives à la place de la prédiction en sciences (en général, ou par rapport à l'explication) pourront utilement éclairer ces débats.

Une troisième ligne d'interrogation pourrait porter sur les conditions de production des dispositifs d'IA, et ainsi interroger les discours totalisants qui les décrivent comme des systèmes automatisés fondés sur la seule puissance du calcul. La plupart des algorithmes sont en effet entraînés par des individus dont on oublie d'évoquer l'activité de préparation, de nettoyage et d'alimentation des données. L'apprentissage automatique est largement nourri par le travail invisible de milliers de personnes, ces "travailleurs du clic" (Casilli 2018) payés peu ou parfois pas à entraîner les algorithmes. Loin d'être un système entièrement automatisé, l'apprentissage machine actuel repose sur des interventions humaines récurrentes. Les communications pourront alors porter sur ce travail humain d'entraînement des machines. L'IA peut-elle y échapper (par exemple en développant des mécanismes de cumulativité via des procédés de *transfer learning*), ou est-elle condamnée à s'appuyer toujours sur ce travail de l'ombre? On pourra aussi s'interroger sur les effets de cet entraînement toujours particulier sur les résultats que proposent les algorithmes, et sur les situations qu'ils ne peuvent pas prendre en compte (invitant ainsi prolonger les réflexions classiques sur les biais et sur leurs origines afin de les mettre en discussion avec la thématique de la journée d'étude). On pourra encore se demander ce que les critères d'évaluation des algorithmes de *learning*, largement organisés autour de l'amélioration de performances prédictives sur quelques jeux de données devenus classiques (MNIST, ImageNet), a fait à la logique calculatoire de l'IA, à la fois en termes d'établissement d'un langage commun, mais aussi en termes de limites pour appréhender des cas qui différerait de l'étalon commun.

D'autres réflexions pourront être explorées. **La capacité auto-productive de certains algorithmes** (de type apprentissage par renforcement) **ne remet-elle pas en cause la vision de l'IA comme application stricte d'une rationalité algorithmique contenue dans ses données d'entraînement** – aussi biaisées soient-elles. Les exemples d'œuvres d'art produites à partir de réseaux de neurones de type generative adversarial networks ne sont-ils pas une invitation à reposer la question de l'automatisation, tout en s'interrogeant sur la conséquence de cette capacité productive ?

Parmi les intervenant.es qui ont déjà confirmé leur venue : Michèle Sebag, directrice du laboratoire de recherche en informatique ; Antonio Casilli, chercheur sur le digital labor, auteur de *En attendant les*

robots (2019) ; David Bates, chercheur à Berkeley, ancien directeur du Berkeley Center for New Media.

Les propositions de communication, qui peuvent émaner de disciplines aussi diverses que la philosophie, les SHS, l'économie, le droit ou l'art mais peuvent également provenir de spécialistes en machine learning seront à envoyer avant le 1er mars 2020 à l'adresse suivante : francoislevin01@gmail.com (une page maximum). Les réponses seront données dans le cours du mois de mars.

Bibliographie indicative

Susan Athey, "Beyond Prediction: Using Big Data for Policy Problems," Science, February 3, 2017
David Bates, "Automacity, Plasticity and the deviant Origins of artificial intelligence", In Plasticity and pathology : On the Formation of the Neural Subject, pp.194-218, ed. Fordham University, 2015
Julien Boelaert, Étienne Ollion, "The Great Regression. Machine Learning, Econometrics, and the Future of Quantitative Social Sciences", *Revue française de sociologie* 2018/3 (Vol. 59), p. 475-506.
Dominique Cardon, Jean-Philippe Cointet, Antoine Mazières, "La revanche des neurones. L'invention des machines inductives et la controverse de l'intelligence artificielle", *Réseaux*, 2018/5 (n° 211)
Antonio Casilli, *En Attendant les robots*, Seuil, 2018
Gregory Chaitin, "Leibniz, Randomness and the Halting Probability", 2004
Grégoire Chamayou, "Avant-propos sur les sociétés de ciblage", revue Jef Klak, 2015
Yves Citton, "Le court-circuitage néolibéral des volontés et des attentions", *Multitudes* n°68, 2017
Hubert Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason*, The MIT Press, 1992
Robert French, "Subcognition and The Limits of the Turing Test", Mind, 1990, 99, pp. 53-66.
Kurt Gödel, "Sur les propositions formellement indécidables des *Principia Mathematica* et desystèmes apparentés", 1931
Catherine Malabou, *Que faire de leur cerveau bleu ?,* Puf*,* 2017
Luciana Parisi, "La raison instrumentale, le capitalisme algorithmique et l'incomputable"*, Multitudes,* 2016
Antoinette Rouvroy et Thomas Berns, "Gouvernementalité algorithmique et perspectives d'émancipation", *Réseaux*, 2013
Arno Schubbach, "Judging machines: philosophical aspects of deep learning", *Synthese*, 2019
J. R. Searle, "Minds, Brains and programs", *The Behavioral and Brain Sciences*, vol. 3, Cambridge University Press, 1980
Bernard Stiegler, *La Société automatique*, tome I, Fayard, 2015
Alan Turing, "On Computable numbers", 1936
Jean Vioulac, *Approche de la criticité*, PUF, 2018