Intelligence articielle, mégadonnées et gouvernance

François Pellegrini

Intelligence artificielle, mégadonnées et gouvernance

François Pellegrini

20 octobre 2017

Université de Bordeaux, LaBRI & Inria Bordeaux - Sud-Ouest, 351 cours de la Libération, 33405 Talence cedex, France

{francois.pellegrini@labri.fr}

La révolution numérique dans laquelle l'humanité est actuellement plongée se déploie dans un temps extrêmement rapide par rapport à celles qui l'ont précédée, du fait de l'accélération continue du progrès technique. En dépit du manque de recul, quatre âges peuvent déjà être distingués. Le premier est l'âge de l'ordinateur (1950-1970), qui a vu l'arrivée sur le marché des premiers ordinateurs commerciaux, au service des administrations et des grandes entreprises. Le deuxième est l'âge du logiciel (1970-1990), au cours duquel les « programmes informatiques » deviennent des « logiciels », biens non rivaux dotés de leur marché propre, découplé de celui des ordinateurs. Le troisième est l'âge des réseaux (1990-2010), au cours duquel le déploiement et la démocratisation de l'Internet et des autres réseaux numériques (tels que la téléphonie mobile) ont conduit à l'interconnexion forte des personnes et des organisations, permettant la création de biens communs numériques (logiciels libres, Wikipedia, etc.) mais aussi des grands silos de données (les « GAFA », pour « Google - Apple - Facebook - Amazon »). Le quatrième, enfin, est l'âge de la donnée (2010-...), concomitant de l'émergence des objets connectés, dans lequel le logiciel, autrefois leur maître, devient le serviteur des données puisque, dans le cas des logiciels auto-apprenants, son comportement est directement modifié par elles.

La possibilité de traiter efficacement des masses considérables de données accroit la tentation de faire reposer la gestion des sociétés humaines sur une gouvernance algorithmique. Dans cette vision, les traitements de données sont considérés comme immunisés contre les errements irrationnels de l'âme humaine, et leurs résultats assimilés à des vérités scientifiques qu'il serait absurde de contester. Par la capacité d'extraire des « signaux faibles » de masses de données inaccessibles à l'esprit humain, les « mégadonnées » et l'« intelligence artificielle » échapperaient à l'entendement et ne seraient plus contestables en pratique. Or, les traitements de données, comme tout artefact, s'inscrivent dans leur environnement social, économique et culturel, et sont de fait pétris de biais humains.

1 Les mégadonnées

Il n'existe pas de définition uniforme des « mégadonnées » (« *big data* »). Leur plus petit dénominateur commun est couramment désigné par les « xV », du fait de l'initiale « v » des caractéristiques qui sont supposées les définir. Quant à la valeur « x », couramment égale à 3, elle peut aller jusqu'à 6 chez certains auteurs.

Le premier critère en « v » est le « volume » des données considérées. La volumétrie est un phénomène déclencheur, mais n'est en fait pas discriminante. Le critère sous-jacent est l'impossibilité de traiter exhaustivement l'information considérée, qui va susciter l'emploi de méthodes heuristiques de recherche de corrélations, dont l'« intelligence artificielle » (voir *infra*). Le deuxième critère est la « vélocité ». Celle-ci caractérise le fait que les stocks de données considérés sont en perpétuelle modification, au point même même d'induire des incohérences au sein des données collectées. Le troisième critère est la « variété », indiquant par cela que les traitements de mégadonnées opèrent sur des données non structurées, dont la sémantique dépend du contexte de collecte. Sont également évoqués, comme autres « v » : la « véracité », relative à l'obligation de disposer des données les plus brutes et exactes possibles, tout filtrage étant susceptible de détruire les signaux recherchés ; la « visualisation », c'est-à-dire la capacité d'interprétation fine des résultats, afin de ne pas créer de biais en confondant corrélation et causalité ; et la « valorisation », signifiant que l'usage doit piloter la conception du traitement afin de tirer le meilleur parti des résultats de celui-ci.

Les mégadonnées visent à représenter un système ouvert, et non pas fermé comme dans le cas de la simulation numérique traditionnelle, par exemple (telle qu'en aérodynamique). Il ne s'agit pas de modéliser l'évolution d'un système en fonction de ses caractéristiques prédéterminées (telles que les lois déjà connues de la mécanique des fluides), mais d'abstraire le comportement du système en fonction des données qui décrivent son évolution. Parce que cette abstraction s'appuie sur un raisonnement inductif et non pas déductif, il est de fait impossible d'obtenir des certitudes mathématiques. Ainsi, s'il est possible de déduire, du fait que Socrate est mort, le fait que Socrate était bien mortel, on ne peut qu'induire, du fait que Socrate était un homme, que les hommes sont vraisemblablement mortels, parce que rien ne dit que l'avancée du progrès scientifique ne va pas un jour rendre les hommes immortels.

De par la nature inductive des traitements considérés, conserver plus de données n'apporte pas plus de précision dans l'analyse du phénomène considéré. Les données comportementales en sont un parfait exemple, du fait de leur obsolescence rapide : pour prévoir le comportement des acheteurs d'électro-ménager de 2018, disposer de données sur les habitudes d'achat de l'année 1930 n'apportera rien par rapport à celles de 2017. Ainsi, il faut être capable d'oublier pour continuer à agir efficacement.

Les mégadonnées sont d'une nature différente des traitements statistiques. En effet, ces dernières permettent de capturer l'information relative à un comportement « moyen », supposé partagé, avec de faibles différences, par un grand nombre d'acteurs. À l'inverse, l'objectif des mégadonnées est de capturer les cas différenciateurs [1], « signaux faibles » précurseurs de « nouvelles tendances », qu'il est parfois tentant dans certains cas d'assimiler à des « déviances » suspectes. Afin de gérer la multiplicité des comportements, les mégadonnées s'appuient sur l'abstraction, c'est-à-dire le fait de remplacer des informations par des informations plus compactes, et la connexité, visant à déterminer les liens pouvant exister entre les données.

L'un des principaux arguments des promoteurs des mégadonnées est ainsi la personnalisation des résultats, issue de la prise en compte des individualités. Par l'étude des traces comportementales, il s'agit de prévoir comment chacun va raisonner par rapport à son référentiel de pensée, c'est-à-dire sa rationalité et ses biais cognitifs. Il s'agit de mieux connaître les individus afin de mieux les servir, mais aussi jouer sur leurs sensibilités et anticiper leurs réactions. Ceci pose la question de la loyauté et de l'éthique du responsable de traitement vis-à-vis des personnes concernées.

2 Algorithmes auto-apprenants et biais de traitement

Le modèle traditionnel de programmation des ordinateurs est celui de la « programmation impérative ». Dans ce paradigme, le programmeur doit spécifier explicitement chacune des opérations que le logiciel doit effectuer jusqu'à obtenir le résultat désiré. Dans le cas des logiciels devant opérer au sein d'un espace de problème très vaste, comme par exemple celui de la marche robotisée en environnement naturel (dénivelés, escaliers, fossés, obstacles divers), l'approche impérative conduirait à des programmes extrêmement lourds et complexes et, malgré cela, inaptes à prendre en compte toutes les situations particulières. De ce constat a émergé l'idée de construire des programmes « auto-apprenants » : il ne s'agit plus alors de spécifier chaque action du programme, mais de concevoir un programme capable d'adapter son comportement à son environnement. Dans ce modèle, les résultats produits en sortie par le logiciel à un instant donné dépendent d'un calcul appliqué aux valeurs d'entrée, ce calcul dépendant de paramètres évoluant dans le temps en fonction de l'« apprentissage » du logiciel. Cet apprentissage consiste à renforcer les valeurs de paramétrage conduisant à des résultats conformes à l'objectif attendu, et à modifier celles ne conduisant pas à un résultat pertinent.

Il existe donc une convergence naturelle entre traitements auto-apprenants et mégadonnées, celles-ci étant justement caractérisées par le besoin d'inférer des corrélations au sein de jeux de données massifs. Les algorithmes auto-apprenants sont donc les principaux outils de « l'intelligence artificielle faible », c'est-à-dire l'assistance informatisée à des tâches spécialisées (assistance à la conduite automobile, détection de tumeurs dans les images radiographiques, etc.). Ces traitements opèrent dans un contexte borné, en termes de données fournies et de réponses attendues, bien loin des performances attendues des « intelligences artificielles fortes », intelligences synthétiques généralistes dont la production caractérisera l'atteinte de la « singularité [2] ».

2.1 Réseaux de neurones et apprentissage profond

Les « réseaux de neurones » sont l'une des formes les plus populaires de traitement auto-apprenant. Il s'agit de simuler le fonctionnement d'un ensemble de neurones organiques, les valeurs de sortie du logiciel étant les valeurs calculées par certains neurones, ces valeurs dépendant de l'action inhibitrice ou stimulatrice d'autres neurones qui leur sont connectés, et qui sont eux-mêmes activés par les valeurs d'entrée fournies au logiciel. On reproduit ainsi le comportement d'un organisme vivant, qui agit dans son environnement sur la base des stimuli issus de celui-ci et de sa propre histoire, mémorisée et traités au sein de ses structures neurales [3].

Les systèmes auto-apprenants par réseaux de neurones ont pu apporter des solutions pratiques relativement efficaces à certains problèmes, tels que la reconnaissance des adresses manuscrites pour le tri postal, mais leurs capacités plafonnaient à des taux de réussite modérés (de l'ordre de 60 à 70 % des jeux de test). Leurs performances ont été révolutionnées par l'« apprentissage profond » (« *deep learning* »), technique consistant à structurer les réseaux de neurones en couches successives, à l'image de certaines structures du système nerveux des êtres vivants (rétine, etc.) [4].

1. C'est ainsi qu'on peut parfois dire que « Les mégadonnées commencent là où la statistique s'arrête ».

2. On appelle ainsi le moment hypothétique où les capacités de l'intelligence synthétique dépasseraient celle de l'intelligence humaine, conduisant par accélération exponentielle à l'émergence d'une « supra-intelligence » susceptible de conduire à la perte du pouvoir de l'humanité sur son destin. Voir par exemple : Bill JOY, « Why the future doesn't need us », *Wired Magazine*, avril 2000, https://www.wired.com/2000/04/joy-2/ .

3. Dans le domaine de la locomotion robotisée évoquée ci-dessus, voir par exemple : Randall D. BEER, Hillel J. CHIEL, Roger D. QUINN, Kenneth S. ESPENSCHIED et Patrik LARSSON, « A Distributed Neural Network Architecture for Hexapod Robot Locomotion », *Neural Computation*, vol. 4(3), 1992, pp. 356-365, 1992, DOI 10.1162/neco.1992.4.3.356.

4. Yann LE CUN, Léon BOTTOU, Yoshua BENGIO et Patrick HAFFNER, « Gradient-based learning applied to document recognition », *Proceedings of the IEEE*, vol. 86(11), pp. 2278-2324, novembre 1998, DOI 10.1109/5.726791, http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf .

Cette structuration permet l'extraction de caractéristiques de plus en plus « abstraites » des jeux de données proposés [5], par les niveaux successifs de neurones, jusqu'à pouvoir « capturer » les éléments stylistiques d'un tableau pour les transposer dans une autre image [6].

2.2 Biais des traitements algorithmiques

Les traitements informatisés, en tant que produits de l'activité humaine, ne sont pas exempts de biais. Ceux-ci sont de plusieurs ordres.

En premier lieu, toute tentative de modélisation conduit par nature à un filtrage, puisqu'il s'agit de ne conserver du phénomène considéré que les paramètres qui semblent significatifs. Cette simplification est opérée *a priori* par la personne en charge de concevoir le traitement, en fonction de ses préjugés propres. Par exemple, réduire la définition du genre d'une personne au seul choix « M. / Mme » dans un formulaire traduit le fait que, pour la personne qui a conçu le formulaire, ces deux choix sont les seuls possibles, de façon exclusive l'un de l'autre. Ce biais peut être propagé silencieusement lorsqu'un jeu de données est utilisé pour une finalité autre que celle pour laquelle il avait été initialement recueilli. Par exemple, une fois ces données transférées dans un système dans lequel d'autres choix sont possibles, l'opérateur du nouveau système peut penser que les personnes avaient le choix de répondre autrement et ne l'ont pas souhaité, alors qu'initialement cette liberté ne leur était pas offerte.

Une deuxième catégorie de biais découle des jeux de données utilisés pour entraîner les algorithmes auto-apprenants. La sélection de ces données par l'opérateur exprime les préjugés de celui-ci quant à leur représentativité et à leur exhaustivité. Or, du fait de la grande taille de l'espace des problèmes et de la diversité des inter-relations recherchées, il est très difficile de prétendre que cela est bien le cas [7].

Même en l'absence de toute action explicite de filtrage par l'opérateur, les jeux de données disponibles peuvent être porteurs de biais découlant de causalités cachées. Ainsi, dans le cadre de la recherche de critères permettant la sortie anticipée en traitement ambulatoire de patients atteints de pneumonie, un traitement a-t-il recommandé la sortie anticipée des patients connus comme asthmatiques, alors que cette affection est à l'opposé un risque majeur de complications. Cela tenait au fait que l'hôpital mettait déjà en œuvre une politique de suivi en soins intensifs des patients atteints d'asthme afin d'éviter la survenue de complications, ce qui transparaissait dans les données fournies au traitement comme le fait que les personnes asthmatiques étaient celles qui développaient le moins de complications à l'hôpital et semblaient donc les moins susceptibles d'en développer à l'extérieur [8]. Il aura fallu que des médecins humains étudient les recommandations produites par le traitement pour mettre en évidence ce biais aux conséquences potentiellement mortelles.

L'apparition d'un biais est encore moins évitable lorsque le jeu de données est issu d'un flot d'informations non contrôlé, soumis aux aléa statistiques des faibles nombres. Par exemple, si les deux dernières affaires de meurtres en série étaient par pure coïncidence le fait de deux personnes prénommées Anatole, le traitement pourrait pour sa part induire que tous les Anatole sont des tueurs redoutables.

Enfin, une troisième catégorie de biais tient à la convergence de l'algorithme lui-même, c'est-à-dire à la nature des corrélations qu'il pourra construire en fonction des jeux de données qui lui auront été fournis. Un exemple édifiant en la matière concerne un traitement destiné à reconnaître les photos de chambres à coucher par auto-apprentissage guidé. Ce terme désigne le fait que l'on a préalablement conditionné l'algorithme en lui présentant un ensemble de photographies représentant ou non des chambres à coucher, en lui indiquant à chaque fois s'il s'agissait bien d'une chambre à coucher ou non. Une fois cet apprentissage effectué, le traitement a été testé sur un autre jeu de photographies. Le taux de succès du traitement n'a alors été que d'environ 60 %, ce qui, comme nous l'avons dit, est très faible pour un traitement d'apprentissage profond. Intrigués, les concepteurs du traitement se sont alors livrés à une analyse en profondeur, qui leur a permis de déterminer que le critère sur lequel l'algorithme avait convergé, le considérant comme étant le plus discriminant, était la présence de rideaux aux fenêtres, et non la présence d'une surface plane pouvant servir de lit.

2.3 Re-jeu et preuve

Dans le cas de reconnaissance de chambres à coucher évoqué ci-dessus, il aura fallu que l'équipe de conception du traitement investisse un temps et une énergie significatifs pour prouver l'existence d'un biais et en déterminer la nature. Cela était possible parce que, le traitement fonctionnant en apprentissage guidé, son paramétrage était

5. C'est ainsi qu'un traitement auto-apprenant a pu se conditionner à reconnaître la forme d'une tête de chat dans un ensemble d'images. Voir : Quoc V. LE, Marc'Aurelio RANZATO, Rajat MONGA, Matthieu DEVIN, Kai CHEN, Greg S. CORRADO, Jeff DEAN et Andrew Y. NG, « Building high-level features using large scale unsupervised learning », Actes de la 29ème *International Conference on Machine Learning* (ICML), Édimbourg, décembre 2011, https://static.googleusercontent.com/media/research.google.com/en//archive/unsupervised_icml2012.pdf .

6. Voir par exemple : Ahmed ELGAMAL, Bingchen LIU, Mohamed ELHOSEINY, Marian MAZZONE, « CAN: Creative Adversarial Networks Generating "Art" by Learning About Styles and Deviating from Style Norms », https://arxiv.org/pdf/1706.07068 .

7. Jesse EMSPAK, « How a Machine Learns Prejudice: artificial intelligence picks up bias from human creators - not from hard, cold logic », *Scientific American*, 29 décembre 2016, https://www.scientificamerican.com/article/how-a-machine-learns-prejudice/ .

8. Rich CARUANA, Yin LOU, Johannes GEHRKE, Paul KOCH, Marc STURM et Noémie ELHADAD, « Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission », Actes de la 21ème *International Conference on Knowledge Discovery and Data Mining* (ACM SIGKDD), 2015, pp. 1721-1730, DOI 10.1145/2783258.2788613, http://people.dbmi.columbia.edu/noemie/papers/15kdd.pdf

resté figé. Plusieurs travaux récents visent à augmenter la traçabilité de la prise de décisions par les traitements auto-apprenants (leur « explicabilité »), afin de pouvoir éventuellement exposer plus facilement leurs biais [9].

Dans le cas des traitements « à apprentissage dynamique », pour lesquels le logiciel modifie de façon permanente son paramétrage en fonction des nouveaux jeux de données qui lui sont fournis, il est bien plus difficile de prouver l'existence d'un biais, car le paramétrage qui sera étudié une fois l'expertise ordonnée n'est potentiellement plus le même que celui ayant donné lieu à la décision contestée. Ceci pose la question de la conservation de la preuve à fin de re-jeu et d'étude. Faut-il que les différents jeux de paramètres soient conservés au cours du temps jusqu'à la prescription du délai de recours ?

À l'opposé, quel sens la notion juridique d'appel peut-elle avoir lorsque le logiciel assistant le juge d'appel est le même que celui utilisé par le juge de première instance ?

3 Gouvernance algorithmique

Les sociétés humaines sont en fait déjà gérées selon les principes des mégadonnées. Comme le disait Jean ROSTAND : « Attendre d'en savoir assez pour agir en toute lumière, c'est se condamner à l'inaction ». Les êtres humains sont capables de décider par induction à partir de données incomplètes, et le dressage pavlovien vise justement à corréler, à force de répétition, un événement à un effet, c'est-à-dire à entraîner les réseaux de neurones des sujets en apprentissage guidé.

Le progrès technologique permet aux traitements informatiques de se rapprocher de plus en plus du modèle humain, mais peuvent-ils l'assister de façon pertinente ? Au niveau individuel, on attend de tels traitements un meilleur service à l'usager et, au niveau collectif, une influence positive sur la définition et la mise en œuvre des politiques publiques. Ainsi, parmi les domaines dans lesquels il est souvent considéré que les mégadonnées auront un fort bénéfice à s'appliquer, on cite la santé et le secteur police-justice.

3.1 Encadrement juridique

L'usage de traitements informatisés pour assister la prise de décisions à l'endroit des personnes n'est pas nouveau. C'est ainsi que l'article 10 de la loi « Informatique et Libertés » dispose déjà que nul ne peut être l'objet d'un traitement algorithmique sans qu'un humain ne prenne *in fine* la décision à son compte en bout de chaîne. La loi « CADA », pour sa part, autorise la communication aux personnes concernées du code source des logiciels mis en œuvre par l'administration publique, assimilés à des documents administratifs.

Cependant, la complexification de ces traitements a fait naître un besoin nouveau, celui de leur « transparence », c'est-à-dire de leur intelligibilité par le public et non par les seuls spécialistes, qui ne peut exister que dans la mesure de leur « explicabilité ». Ainsi la loi « République numérique » a-t-elle créé l'obligation pour la puissance publique d'informer les personnes concernées sur la nature des traitements auxquels ils sont soumis par elle [10] ; les modalités que pourrait prendre cette obligation dans le cas des traitements auto-apprenants restent inconnues. L'extension de cette obligation aux traitement mis en œuvre par les acteurs privés, dans le respect du secret industriel, semble inévitable à terme.

3.2 Limites intrinsèques

L'un des cas les plus médiatisés de gouvernance algorithmique est celui du logiciel PredPol®, présenté par ses promoteurs comme permettant de « faire baisser la criminalité » dans les lieux où il est mis en œuvre. L'analyse des principes et des modalités de mise en œuvre de ce logiciel révèle un ensemble de biais significatifs, qui conduit à la réfutation des bénéfices espérés [11].

La principale critique concerne l'usage du logiciel en tant qu'oracle conduisant à une prédiction auto-réalisatrice. La fonction du logiciel est de rationaliser la liste des lieux à patrouiller, en se basant sur les données de victimation déjà collectées. Or, il ressort des études criminalistiques que la criminalité suit une loi de Pareto : 80 % des délits sont concentrés sur 20 % du territoire. Qui plus est, le crime est contagieux spatio-temporellement : une victime non protégée se fera cambrioler à nouveau. De fait, il suffit de prédire toujours les mêmes lieux « à risque » pour être aussi performant que le logiciel. Si les lieux patrouillés sont l'objet de crimes, on pourra dire que le logiciel les avait bien prédits ; si la criminalité y baisse, on pourra dire que c'est grâce aux prédictions du logiciel que les criminels auront été dissuadés et que ce résultat aura été atteint.

Plus généralement, ce type de traitement, tout comme celui destiné à la « prédiction » des récidives, ne fait que mettre en évidence le rôle du contexte social sous-jacent, sans le traiter. En revanche, si celui-ci est traité, il est évident que le problème sera déplacé, sans que le logiciel puisse déterminer où, puisqu'il fournit des réponses selon un modèle

9. Voir par exemple : Abhishek DAS, Harsh AGRAWAL, C. Lawrence ZITNICK, Devi PARIKH et Dhruv BATRA, « Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? », https://arxiv.org/abs/1606.03556 .

10. Article L. 311-3-1 du code des relations entre le public et l'administration, créé par l'article 4 de la loi n° 2016-1321 du 7 octobre 2016.

11. Ismaël Benslimane, « Predpol : prédire des crimes ou des banalités ? », 10 décembre 2014, https://cortecs.org/mathematiques/predpol-predire-des-crimes-ou-des-banalites/ .

inductif dans lequel ces paramètres de niveau supérieur n'entrent pas. Les logiciels de ce type mettent en œuvre des modèles mathématiques inspirés de ceux de la sismologie qui, s'ils ne savent pas prédire les séismes, peuvent être efficaces pour prédire les répliques. Ils s'appuient pour cela sur un principe de localité, c'est-à-dire, transposé dans le domaine policier, en se concentrant sur la répétition des victimations. Il s'agit d'un modèle « stationnaire », stable entre deux « catastrophes », tout comme les policiers de terrain apprennent à connaître leur quartier, pour autant que de grands projets urbanistiques ne viennent pas en bouleverser brutalement la sociologie et l'économie.

4 Conclusion

L'intérêt général n'est pas réductible à la somme des intérêts particuliers. De fait, peut-il être compatible avec le modèle inductif, fondé sur l'agrégation de caractéristiques individuelles ? Les traitements algorithmiques inductifs peuvent être utiles pour identifier des motifs au sein de masses de données et, potentiellement, évaluer l'évolution desdits motifs en réaction à un changement de l'environnement, tel qu'un changement de politique publique. En revanche, pouvoir modifier l'environnement lui-même dans une direction donnée suppose la capacité de s'abstraire et « sortir » du modèle, tâche intellectuelle qui n'est absolument pas à la portée des intelligences artificielles faibles. Les personnes ne sont pas réductibles à leurs données, aussi précises fussent-elles. L'ère de la gouvernance algorithmique n'est pas encore advenue, pour autant que l'humain ne se défausse pas de la responsabilité qui est la sienne.