Éthique des algorithmes

Christine Solnon

Éthique des algorithmes

Christine Solnon, INSA Lyon, Inria, CITI

Illustrations de Luc Damas

Les algorithmes que nous concevons sont souvent utilisés pour proposer des solutions à des décideurs, et parfois même prendre des décisions de façon autonome, dans des contextes très variés. Aussi sommes-nous amenés à nous interroger sur les applications de ces algorithmes. Cependant, si nous sommes plutôt bien placés pour évaluer les possibilités offertes par un nouvel algorithme, la question de savoir s'il est souhaitable ou non de l'utiliser pour une nouvelle application nous dépasse bien souvent.

Une première réponse à cette question consiste à s'appuyer sur la législation. En particulier, l'article 1 de la loi informatique et libertés de 1978 stipule :

> *L'informatique doit être au service de chaque citoyen. Son développement doit s'opérer dans le cadre de la coopération internationale. Elle ne doit porter atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques.*

Cette loi comporte également un certain nombre de principes qui ont été réaffirmés 40 ans plus tard par le RGPD (règlement général sur la protection des données) comme, par exemple, l'interdiction à une machine de prendre seule des décisions emportant des conséquences cruciales pour les personnes, ou le droit pour les personnes d'obtenir des informations sur la logique de fonctionnement de l'algorithme.

Si la loi interdit aux algorithmes de prendre des décisions cruciales de façon autonome, de tels algorithmes ne relèvent plus de la science fiction, et il existe des contextes où leur utilisation commence à être envisagée. Un exemple très médiatisé est celui des véhicules autonomes qui pourraient améliorer la sécurité de tous, mais posent également de nombreuses questions. Par exemple, quelle règle appliquer lorsque le véhicule doit éviter des piétons, mais que cela peut l'amener à blesser ses passagers ? Ou encore, qui est responsable en cas d'accident ? Ces questions ont fait l'objet de procès fictifs qui ont montré toute leur complexité [1].

Dans la mesure où les avancées technologiques ouvrent régulièrement de nouvelles possibilités sur lesquelles la loi ne s'est pas encore prononcée, le simple respect de la loi n'est pas suffisant, et il est nécessaire de suivre des principes éthiques garantissant le respect des droits fondamentaux de chaque être humain : dignité, liberté, égalité, solidarité, ou encore justice. Enfin, en plus d'être licite et éthique, un algorithme doit également être robuste afin de garantir qu'il ne peut avoir d'effets involontaires. Concrètement, ces principes se traduisent par des propriétés qui sont évoquées dans la suite de cet article.

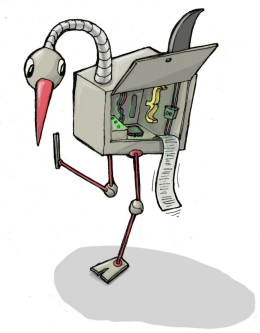
Le sujet est d'actualité, du fait des avancées spectaculaires récentes en Intelligence Artificielle (IA), et de nombreux groupes, associations, ou commissions ont rédigé des rapports, souvent

1. Voir, par exemple, le procès de l'IA organisé par la cour d'appel de Paris et l'association Jurisnaute en 2018.

largement relayés dans la presse. Cet article est notamment inspiré des rapports de Claude Castelluccia et Daniel Le Métayer [1], de la CNIL [2], et de la commission européenne [3], ainsi que du livre de Serge Abiteboul et Gilles Dowek [4].

Transparence et explicabilité

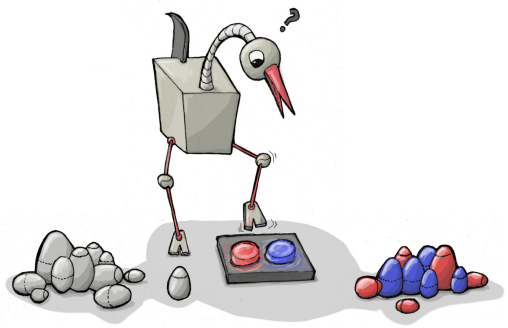
Pour pouvoir utiliser un algorithme en toute confiance, il faut comprendre comment il fonctionne, et donc avoir accès au code et/ou à des documents décrivant la logique de fonctionnement de l'algorithme. Dans le cas où des données ont été utilisées pour concevoir ou paramétrer l'algorithme, il faut également connaître les caractéristiques et l'origine de ces données.



Par ailleurs, pour pouvoir contester une décision prise par un algorithme (ou bien corriger des dysfonctionnements constatés), il faut pouvoir identifier les raisons qui ont amené à la décision, et expliquer ces raisons de façon intelligible pour un être humain. Ces explications peuvent parfois être dérivées par analyse de l'algorithme, mais cela n'est pas toujours possible (par exemple si le code n'est pas accessible, ou si l'algorithme a des paramètres dont les valeurs ont été fixées par apprentissage à partir de très nombreuses données). Dans ce cas, l'explication doit être construite (sous la forme d'une relation entre les données en entrée et la décision, par exemple) en observant des exécutions de l'algorithme. L'explicabilité des algorithmes d'IA (et notamment des algorithmes "boites noires") est un domaine de recherche très actif appelé *Explainable AI (XAI)*.

L'exemple des applications Admission Post-Bac (APB) et ParcoursSup illustre bien les besoins de transparence et explicabilité : d'une part les utilisateurs de ces applications ont besoin de comprendre le fonctionnement de l'algorithme *a priori* pour faire des choix éclairés, et d'autre part ils doivent pouvoir demander des explications *a posteriori* dans le cas où ils souhaitent contester la décision. Ces deux besoins sont affirmés dans la loi informatique et liberté de 1978, et la CNIL a souligné qu'APB ne respectait pas cette loi en 2017.

Equité et absence ce biais



Un algorithme produisant un classement ou encore une sélection de personnes (candidates à un crédit, par exemple) est par nature discriminant. Le fait d'effectuer cette sélection à l'aide d'un algorithme permet de garantir que les mêmes règles sont appliquées à tous les candidats, contrairement à une sélection faite par des humains. Cependant, il est nécessaire d'assurer que les règles appliquées par l'algorithme sont justes et équitables. En particulier, l'algorithme ne doit pas exploiter d'information non pertinente (telle que, par exemple, les convictions politiques ou la religion). Notons qu'il n'est pas toujours suffisant de simplement supprimer une information non pertinente des données utilisées par l'algorithme pour assurer l'équité car cette information peut être fortement corrélée à d'autres de sorte que l'algorithme pourra l'inférer à partir de ses données.

Les règles de sélection ne sont pas toujours explicitement définies dans l'algorithme, et il est possible d'apprendre un modèle de sélection à partir de données d'entrainement. Dans ce cas, le modèle appris peut ne pas être équitable lorsque les données d'entrainement ont été produites ou sélectionnées par des humains qui ne sont pas équitables eux-mêmes. Cela est illustré dans [5] sur un algorithme de *Word Embedding* qui calcule une représentation vectorielle des mots permettant d'inférer des analogies telles que *king - man + woman = queen*. Quand l'algorithme est entrainé sur des textes provenant d'un site d'information en ligne, le modèle appris reproduit les biais de genre présents dans ces textes : la réponse à *doctor - man + woman* est *nurse*, et la réponse à *computer programmer - man + woman* est *homemaker*. Si ce modèle était utilisé par un *chatbot* dédié à l'orientation professionnelle, il contribuerait à renforcer la sur-représentation des femmes dans les écoles d'infirmières, et leur raréfaction dans les filières d'informatique.

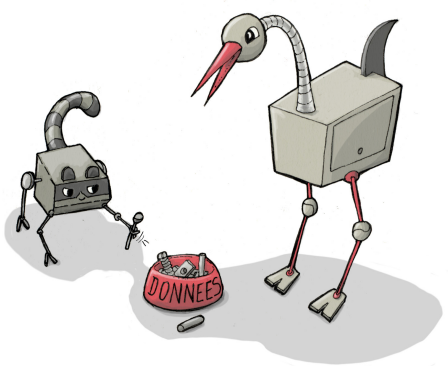
Selon la façon de collecter les données, d'autres biais peuvent apparaître. Typiquement, si certaines catégories de personnes sont peu ou pas représentées dans les données, alors les modèles appris seront moins pertinents pour ces personnes que pour les personnes fortement représentées. Par exemple, près de la moitié des images de la base *ImageNet* proviennent des États-Unis, et près de 17% du Royaume Uni, de l'Italie ou du Canada. Par conséquent, un algorithme entrainé sur cette base aura tendance à fournir des réponses plus pertinentes aux requêtes d'un américain qu'à celles d'un indien [6].

Certains algorithmes construisent un profil de chaque utilisateur à partir de données initiales et de traces d'interaction entre l'utilisateur et l'algorithme. Ces profils permettent à l'algorithme de personnaliser ses réponses (informations, recommandations ou publicités, par exemple). Si l'utilisateur ne sait pas que les réponses sont filtrées et ne connait pas les critères utilisés pour cela, l'algorithme biaise sa perception du monde et réduit notamment la diversité de pensée (volontairement ou pas).

Ainsi, quand un algorithme utilise des données (d'entrainement ou de personnalisation), il faudrait non seulement préciser l'origine des données mais également identifier et décrire les biais possibles et les conséquences que ces biais peuvent avoir sur le comportement de l'algorithme.

Robustesse, sécurité et confidentialité

Un algorithme doit bien évidemment être correct et conforme à sa spécification. Il existe cependant des cas où l'erreur fait partie de la spécification : typiquement, un algorithme de classification peut se tromper, et des mesures (telles que la précision et le rappel, par exemple) sont utilisées pour évaluer sa fiabilité. Dans ce cas, il est important d'informer l'utilisateur sur le niveau de fiabilité de l'algorithme ainsi que sur les conséquences d'une erreur.



Un algorithme doit aussi conserver ses propriétés en cas d'attaque malveillante. En particulier, il ne doit pas être possible de modifier son comportement, ou encore d'interrompre son exécution. Un point important concerne la protection des données personnelles (qu'elles soient fournies directement par une personne, ou bien générées à partir des traces d'utilisation de l'algorithme par une personne) qui ne doivent pas pouvoir être accessibles aux personnes non autorisées. Un

exemple très médiatisé est celui de *Cambridge Analytica* qui a utilisé des données personnelles collectées par *Facebook* lors de la campagne présidentielle de 2016 au États-Unis, montrant la finesse de la ligne rouge séparant la persuasion de la manipulation [7].

Notons que l'anonymisation de données personnelles est un problème difficile : il ne s'agit pas simplement de supprimer les informations permettant d'identifier directement une personne, mais de garantir qu'il n'est pas possible d'inférer cette identité à partir des données accessibles, ce qui est beaucoup plus complexe.

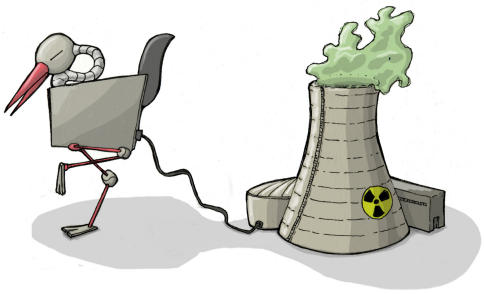
Pour garantir la robustesse, la sécurité et la confidentialité, il est possible d'adopter une démarche spécifique dès la conception de l'algorithme (*security/privacy by design*). Bien souvent, ces propriétés sont vérifiées *a posteriori* par des tests dont les résultats devraient être publics. Par exemple, les utilisateurs d'un algorithme de classification devraient être informés sur les taux de précision et de rappel mesurés par ces tests, et ces taux devraient correspondre à ceux observés lors de l'exploitation (ce qui n'est pas évident dès lors que l'algorithme est évalué sur des cas différents de ceux qui lui sont proposés lors de l'exploitation). Plus généralement, les expériences menées pour évaluer les propriétés d'un algorithme devraient être reproductibles. Cette notion de reproductibilité est également importante pour un algorithme décrit dans un article de recherche. Différents niveaux sont distingués par l'ACM [8] :

- la *repeatability* garantit que les auteurs sont capables de reproduire leurs propres résultats avec des conditions identiques à celles décrites dans l'article (même environnement, mêmes jeux de test, etc) ;
- la *replicability* garantit qu'une personne différente des auteurs peut reproduire les résultats publiés avec des conditions identiques à celles décrites dans l'article ;
- la *reproducibility* garantit qu'une personne différente peut reproduire les résultats même en changeant certaines conditions (par exemple, un autre environnement, ou d'autres jeux de test).

Soutenabilité

Les algorithmes de recherche opérationnelle peuvent être utilisés pour minimiser la consommation de ressources ou, de façon plus générale, l'impact environnemental : par exemple, minimiser la distance parcourue par des camions de livraison, ou encore minimiser la consommation électrique de *data centres*.

Malheureusement, ces applications consomment également des ressources que ce soit au moment de leur conception ou au moment de leur exploitation. Selon le rapport du *Shift Project* [9], la part du numérique dans la consommation mondiale d'énergie (elle-même en croissance de 1,5 % par an) est passée de 1,9% en 2013 à 3,3% en 2020, et la part du numérique dans les émissions de gaz à effet de serre est passée de 2,5% en 2013 à 4 % en 2020.



Un exemple bien connu pour être particulièrement gourmand en calcul est le *Bitcoin*, dont la consommation énergétique annuelle est comparable à celle de pays tels que l'Autriche ou le Vénezuela selon la plateforme *Digiconomist*[2]. Les algorithmes de *deep learning* sont également

2. https ://digiconomist.net/

connus pour avoir besoin d'une grande puissance de calcul au moment de leur entrainement. Cette phase d'entrainement n'est que la partie visible de l'iceberg car il est généralement nécessaire de faire un grand nombre d'expérimentations avant de trouver la bonne configuration des hyper-paramètres de l'algorithme (*hyperparameter grid search*) [10].

Cette phase de paramétrage est également incontournable pour les algorithmes basés sur des méta-heuristiques, et il existe des outils de configuration automatique (tels que *paramILS*, par exemple) pour chercher le meilleur paramétrage étant donné un ensemble d'instances. Ces outils permettent une comparaison équitable en assurant que la même attention a été portée au paramétrage de chaque algorithme. Cependant, ils sont également de gros consommateurs de calcul. Ainsi, au moment d'évaluer les performances d'un algorithme, il serait bon de préciser le temps de calcul qui a été nécessaire pour fixer ses paramètres ainsi que la sensibilité de l'algorithme aux variations des paramètres.

Au delà de l'énergie consommée par l'exécution d'un algorithme, il faut également considérer l'impact environnemental de l'infrastructure matérielle : énergie consommée pour extraire les matières premières, fabriquer et transporter le matériel, mais aussi consommation de matières premières non renouvelables et très mal recyclables telles que le gallium ou le lithium, par exemple. Si cet impact est actuellement difficile à quantifier exactement, il devient urgent d'intégrer ces considérations avant de déployer de nouvelles applications, et le GDS EcoInfo [3] du CNRS étudie ce sujet.

Et toutes les autres propriétés

Il est difficile d'être exhaustif sur un tel sujet, et il existe bien d'autres propriétés importantes. Par exemple, les nombreuses plateformes d'intermédiation, qui mettent en relation offres et demandes (pour se déplacer, ou se loger, par exemple) transforment notre société et posent des questions éthiques [11]. De même les applications permettant de partager les avis des internautes tendent à réduire des services ou des personnes à des notes qui peuvent avoir des effets dévastateurs. L'échelle de déploiement des algorithmes, qui peuvent toucher quasi instantanément la totalité de la planète, amplifie les rétroactions et doit nous rendre particulièrement prudents.

Face à ces nombreuses questions, il est important de continuer la réflexion (indépendamment des entreprises qui exploitent ces algorithmes) et de faire évoluer la législation. À notre niveau, au moment de choisir nos sujets de recherche, nous pouvons nous demander s'ils peuvent contribuer à rendre notre monde meilleur.

Références

[1] Claude Castelluccia and Daniel Le Métayer. Understanding algorithmic decision-making : Opportunities and challenges. Technical report, European Parliamentary Research Service, 2019.

[2] CNIL. Comment permettre à l'homme de garder la main ? les enjeux éthiques des algorithmes et de l'intelligence artificielle. Synthèse du débat public animé par la cnil dans le cadre de la mission de réflexion éthique confiée par la loi pour une république numérique, Commission Nationale Informatique & Libertés (CNIL), 2017.

3. https ://ecoinfo.cnrs.fr

[3] GEHNIA (Groupe d'experts de haut niveau sur l'Intelligence Artificielle). Lignes directrices en matière d'éthique pour une ia digne de confiance. Technical report, Commission Européenne, 2019.

[4] Gilles Dowek and Serge Abiteboul. *Le temps des algorithmes*. Essais & Documents. LE POMMIER, 2017.

[5] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. Man is to computer programmer as woman is to homemaker ? debiasing word embeddings. In *Conference on Neural Information Processing Systems*, pages 4349–4357, 2016.

[6] James Zou and Londa Schiebinger. Ai can be sexist and racist — it's time to make it fair. *Nature*, 559 :324–326, 07 2018.

[7] H. Berghel. Malice domestic : The cambridge analytica dystopia. *Computer*, 51(05) :84–89, 2018.

[8] ACM (American Computing Machinery). Artifact review and badging. *ACM publications policies and procedures*, 2016.

[9] The Shift Project. Pour une sobriété numérique. Technical report, 2018.

[10] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. *CoRR*, 2019.

[11] Stéphane Grumbach. Qu'est-ce que l'intermédiation algorithmique. *Bulletin de la Société Informatique de France*, 7 :93–111, 2015.