Qualité, équité, transparence, vérification et explicabilité des décisions algorithmiques

Serge Abiteboul

Qualité, équité, transparence, vérification et explicabilité des décisions algorithmiques

Par Serge ABITEBOUL
Autorité de régulation des communications électroniques et de la poste (Arcep),
Institut national de recherche en sciences du numérique (Inria)

Nous considérons des aspects, surtout techniques, de la qualité, l'équité, la transparence, et l'explicabilité des décisions algorithmiques.

We consider aspects, especially technical, of the quality, fairness, transparency, and explainability of algorithmic decisions.

Introduction

L'informatique accélère les découvertes scientifiques et l'innovation, transforme nos vies et la société. De plus en plus, les logiciels sont amenés à prendre des décisions qui influencent nos vies[1] dans les domaines de la santé, la justice, la banque, etc. Ils choisissent les informations auxquelles nous sommes exposés sur les réseaux sociaux et les moteurs de recherche. Les algorithmes ont ainsi pris une forme d'autorité. En « assumant » ou pas la responsabilité qui devrait accompagner cette autorité, les algorithmes vont gagner la confiance des utilisateurs, ou leur défiance. Le triangle « autorité, responsabilité, confiance » est typiquement inhérent aux décisions algorithmiques. Par exemple, dans le cadre des réseaux sociaux :

• **Autorité**. Devant les comportements nocifs de certains de leurs utilisateurs et parfois d'États, certains réseaux sociaux bloquent des contenus, ferment des pages ou des groupes, bannissent des utilisateurs de la plateforme. Doit-on leur laisser ainsi la responsabilité de définir de fait des valeurs de notre société ? Quelle légitimité ont-ils à le faire ?

• **Responsabilité**. Les réseaux sociaux nous informent, nous permettent d'échanger, de tisser de nouveaux liens. Pourtant, pédopornographie, bobards (*fakenews*), messages de haine, harcèlements, etc., la liste de leurs dérives est longue. Comme participants à la vie de la cité, ne pouvons-nous exiger d'eux des comportements responsables ?

• **Confiance**. De plus en plus, les citoyens mettent en cause les réseaux sociaux. Comment faire pour qu'ils puissent en jouir sans avoir à en supporter les nuisances, et pour rétablir leur confiance dans ces réseaux ?

Ce triangle soulève des questions de qualité, d'équité, de transparence, de vérification, d'explicabilité des décisions algorithmiques, que nous considérons ici.

• **Éducation.** Le problème avec les algorithmes, c'est que nous ignorons souvent ce qu'ils sont réellement et comment ils fonctionnent. Un minimum de compétence pour entrouvrir les boîtes noires est un point de passage obligé pour établir la confiance dans les décisions algorithmiques. Cet aspect, bien qu'essentiel, ne sera pas

[1] ABITEBOUL S. & DOWEK G. (2017), *Le temps des algorithmes*, Paris, Le Pommier.

considéré dans cet article.

La qualité des décisions

Nous sommes régulièrement confrontés à des erreurs de logiciels. Comment pourrait-il en être autrement ? Un correcteur professionnel laisse encore des fautes d'orthographe dans un roman de quelques dizaines de milliers de caractères. Comment espérer la perfection d'un logiciel comme Windows XP contenant 40 millions de lignes de code, coécrites par une armée de programmeurs ? De plus, même si les résultats d'un logiciel sont le fruit de raisonnements logiques, celui-ci peut se tromper, parce que son raisonnement (son algorithme) est incorrect, que les données, sur lesquelles il se base, sont incomplètes, erronées, biaisées, qu'il est mal utilisé, ou que le problème est trop complexe, insuffisamment analysé.

Quand un logiciel est proposé pour prendre une décision importante à la place d'humains, la question de sa qualité se pose de façon aiguë. On est tenté de placer la barre très haut. Par exemple, aux yeux de certaines personnes, il serait inacceptable d'accorder à un véhicule autonome le droit d'emprunter nos routes s'il court le risque de causer un accident mortel. Comme les logiciels sont plus simples à analyser que les humains, plus simples à « corriger », il est légitime d'exiger que le risque soit statistiquement plus faible si le véhicule est autonome. Mais ne serait-il pas disproportionné d'exiger la perfection d'un véhicule autonome, quand nous acceptons de partager la route avec des conducteurs en état d'ébriété, avec d'autres à la vue déficiente, etc. ?

Enfin, les logiciels sont de plus en plus utilisés dans des domaines (typiquement de sciences humaines et sociales) où les concepts sont complexes et difficiles à spécifier de façon précise, des domaines où les meilleurs experts humains peuvent hésiter, ne pas être d'accord. Dans de tels cadres, on a recours à des techniques[2] comme l'analyse massive de données (*big data*) ou l'apprentissage automatique (*machine learning*) qui sont souvent les seules capables aujourd'hui d'apporter des réponses. La qualité des décisions ne dépend plus alors uniquement du code informatique, mais aussi des données qui vont guider ce code. Par exemple, les analyses du logiciel COMPAS, utilisé pour aider les juges états-uniens à décider de mises en liberté sous condition, ont mis en évidence la très piètre qualité de ses résultats[3].

L'équité

Les logiciels, quand ils proposent des décisions, doivent évidemment respecter nos lois. Mais, au-delà, nous attendons d'eux que, participant à la vie en société, ils respectent aussi nos valeurs éthiques, ainsi, qu'ils se montrent « équitables », qu'ils ne présentent pas de biais. Dans *Weapons of Math Destruction*[4], Cathy O'Neil

---

[2] ABITEBOUL S. & PEUGEOT V. (2017), *Terra data : qu'allons-nous faire des données numériques ?* Paris, Le Pommier.

[3] YONG E. (2018), "A popular algorithm is no better at predicting crimes than random people", *The Atlantic*.

[4] O'NEIL C. (2016), *Weapons of math destruction: How big data increases inequality and threatens democracy*, Crown Books.

explique parfaitement les mécanismes qui peuvent introduire des biais dans l'analyse de données.

La tentation d'introduire discrètement des biais est forte pour ceux qui contrôlent la conception des algorithmes[5]. Mais des biais peuvent aussi être introduits involontairement. Sans que l'entreprise en soit consciente, les tarifs des agrafeuses de Stapples.com étaient plus élevés dans les quartiers défavorisés[6], parce que le logiciel corrélait le prix avec la distance à un magasin vendant des agrafeuses, et que de tels magasins sont rares dans les quartiers défavorisés.

L'apprentissage automatique peut en particulier conduire à introduire des biais. De tels biais sont parfois causés par des insuffisances dans les données utilisées pour « entraîner » le logiciel, comme cela a été observé pour la reconnaissance faciale[7]. Ils proviennent quelquefois aussi du fait qu'un monde numérique peut se contenter d'être le miroir d'une certaine réalité : il a été démontré que le logiciel COMPAS[8] (pour la remise en liberté conditionnelle) défavorisait des minorités ethniques, en reproduisant les biais de certains juges états-uniens. Il faudrait aussi parler des algorithmes de classement[9], basés sur la popularité, que l'on trouve dans des applications de rencontre comme Match.com ou de financement comme Kickstarter. Le biais introduit par la mise en avant selon la popularité résulte, suivant les cas, en l'absence de diversité, la discrimination, voire l'exclusion de certains. *The rich get richer, the poor get poorer.*

Si l'équité aborde le rapport entre un individu et un groupe social, on peut aussi s'intéresser au rapport entre un individu et le logiciel lui-même. Un logiciel est typiquement proposé avec des fonctionnalités, des promesses envers ses utilisateurs. On s'attend à ce que le logiciel obéisse aux règles telles qu'elles ont été édictées pour lui-même. Si une plateforme de recommandation de restaurants annonce que les notes qu'elle décerne ne s'appuient que sur les avis de ses utilisateurs, il serait déloyal de sa part de favoriser des restaurants qui seraient clients de certains de ses services.

Nous n'avons fait ici qu'esquisser les grandes lignes de la responsabilité des décisions algorithmiques. Cette responsabilité est un composant essentiel pour construire la confiance. Elle doit s'accompagner de transparence et de vérification.

Transparence et vérification

La confiance dans les algorithmes peut se construire à partir de la compréhension de ce qu'ils font, de leur transparence. Forcés à rendre leurs algorithmes transparents, les entreprises ou les États sont contraints à rendre visibles leurs choix, peut-être à révéler leurs erreurs, leurs errements. Prenons le cas de l'affectation des futurs

[5] DUCOURTIEUX C. & PIQUARD A. (2019), « Concurrence : l'Europe inflige à Google une troisième amende, d'un montant de 1,49 milliard d'euros », *Le Monde*.

[6] VALENTINO-DEVRIES J. (2012), "Websites vary prices, deal based on users' information", *The Wall Street Journal*.

[7] SIGNORET P. (2018), « Une étude démontre les biais de la reconnaissance faciale, plus efficace sur les hommes blancs », *Le Monde*.

[8] LARSON J. *et al.* (2016), "How we analyzed the COMPAS recidivism algorithm", *Pro Publica*.

[9] YANG K, GKATZELIS V. & STOYANOVICH J. (2019), "Balanced ranking with diversity constraints", *Proceedings* of the Twenty-Eighth International Joint Conference on Artificial Intelligence.

étudiants de l'enseignement supérieur français dans les formations disponibles. Pendant longtemps, cette affectation a été réalisée en toute opacité par des humains. Ces humains ont été remplacés par des logiciels, d'abord APB puis Parcoursup, ce qui a introduit davantage de justice dans l'affectation. De plus, quand le code de ces logiciels a été mis en accès libre, la procédure est devenue transparente. Un avantage considérable est qu'il devient alors possible d'étudier les règles qui guident ses choix d'affectation, et de les contester. La transparence ouvre la porte au débat.

La transparence des algorithmes gagne du terrain en Europe. Elle devrait se trouver affirmée notamment par les futurs Digital Service Act et Digital Market Act. Il est d'ailleurs intéressant d'observer le contraste entre la transparence attendue de plus en plus de la part des grandes entreprises du *web* ou des États, et une forme d'opacité de la vie privée des citoyens prônée par des règlements comme le RGPD (règlement général sur la protection des données, règlement de l'Union européenne). Ce traitement différencié est destiné à corriger, au moins en partie, la dissymétrie d'information actuelle entre ces acteurs, et pourrait participer au rétablissement de la confiance des citoyens.

Pourtant, la transparence seule ne suffit pas. Par exemple, la « librairie » SSL de Debian (service d'encodage et de certification SSL) comportait une erreur qui affaiblissait son système cryptographique. La transparence (le logiciel était ouvert donc le code en accès libre) a permis à quelqu'un de vérifier le code et de trouver l'erreur, mais il a fallu attendre deux ans que quelqu'un se penche sur le code et trouve le *bug*. La transparence facilite la vérification, mais ne la remplace pas. Ceux qui proposent des algorithmes peuvent toujours déclarer leurs bonnes intentions. Mais, seule la vérification de ces algorithmes permet de se convaincre de la qualité de leurs réponses, de leur respect des valeurs éthiques et de leur loyauté.

Dans de nombreux cas, la spécification formelle de propriétés, faisant intervenir des humains, est souvent déjà une barrière. Dans le cas du logiciel COMPAS pour la remise en liberté conditionnelle, il a pu être montré que ce logiciel était équitable s'agissant d'une spécification particulière d'« équitable », injuste pour une autre, et que ces deux spécifications (toutes deux pourtant apparemment convaincantes) ne pouvaient être en même temps satisfaites[10].

Pour vérifier le comportement d'un traitement algorithmique, on peut :

• analyser son code et les données qu'il utilise, ce qui s'apparente à faire la preuve d'un théorème mathématique, ou
• analyser ses effets, ce qui tient plutôt à l'étude d'un phénomène physique comme le climat ou biologique comme le cœur humain.

La première approche nécessite d'avoir accès au code et aux données (par exemple, d'entrainement) qui guident ce code. Cela peut être le cas s'ils sont en accès ouvert. Sinon, on doit recourir à des mécanismes lourds d'audit. Il faut bien insister sur la complexité de ce problème de vérification. Si de nombreux résultats ont déjà été obtenus sur la vérification de logiciels dans des domaines tels que la sûreté (*e.g.,* la garantie d'absence de panne pour un logiciel de pilotage de métro automatique), la vérification de propriétés liées à l'équité de décisions a encore été peu étudiée.

L'analyse des effets s'appuie sur des observations, souvent statistiques, du

[10] KLEINBERG J. (2018), "On algorithms and fairness", Collège de France (site web).

comportement du logiciel. Le logiciel est alors vu comme une boîte noire, ce qui ne simplifie pas le problème. Le logiciel AdFisher[11] a ainsi permis de détecter que les publicités de Google Ads s'appuyaient sur le genre, un attribut protégé : les hommes voyaient des annonces pour des emplois mieux rémunérés beaucoup plus souvent que les femmes. Pour arriver à discerner ce biais, AdFisher a construit des profils d'internautes identiques si ce n'était pour leur genre, et analysé les publicités qu'ils recevaient.

Explicabilité et contestation

La transparence des algorithmes donne des indications sur leur comportement « en général ». Un individu confronté au choix d'un algorithme peut vouloir une explication pour « son cas particulier ».

La loi du 7 octobre 2016 pour une République numérique, comme le RGPD en avril de la même année, a commencé à introduire des exigences d'explicabilité. Cela semble naturel : quand des décisions sont prises qui nous concernent, nous voulons savoir pourquoi. Pour des décisions prises par des humains, on doit se satisfaire de peu d'explications souvent impossibles à vérifier. Pour des décisions algorithmiques, il est techniquement possible d'exiger bien davantage d'explications.

Si un médecin s'appuie sur un système informatique pour le diagnostic général d'un patient, il ne peut se contenter d'obtenir comme réponse : « appendicite ». Le médecin a besoin de la justification du diagnostic proposé, du raisonnement qui conduit à faire ce diagnostic, des statistiques sur lesquelles ce raisonnement s'appuie, etc. Le médecin a besoin de telles explications, pour adopter ce diagnostic, informer le patient, proposer éventuellement d'autres examens, demander l'avis d'un confrère, etc.

On retrouve dans de nombreux cas de décisions algorithmiques ce même besoin de rendre compréhensible une décision, et de la rendre ainsi acceptable. Il n'est pas toujours simple d'obtenir de telles explications, en particulier, les résultats d'algorithmes d'apprentissage automatique, qui résultent d'énormes volumes d'opérations et sont le fruit de la mise au point de nombreux paramètres, sont souvent difficilement explicables. Cette technique est utilisée en médecine, par exemple, pour détecter des tumeurs. Mais dans des cas comme le diagnostic général, des techniques à bases de règles qui proposent des explications sont préférées.

Comme on peut contester la décision d'un juge, « faire appel », on doit pouvoir contester celle d'un algorithme, car lui aussi peut se tromper. C'est ce que l'on trouve sur certains réseaux sociaux : on peut contester le rejet d'un contenu, bloqué par la modération. Une possibilité qu'offre par exemple Facebook. Devant la critique du fait que cette contestation était traitée en interne, ce réseau social est allé jusqu'à créer un « comité de supervision » indépendant doté des moyens de changer les grandes lignes de la modération. Ce comité agit comme une « Cour suprême », ce qui tend à réduire un peu plus la distance entre un réseau social et un État.

[11] DATTA A., TSCHANTZ M. C. & DATTA A. (2015), "Automated experiments on ad privacy settings", *Proceedings on Privacy Enhancing Technologies*.

Conclusion

Les réticences envers la prise de décision par les algorithmes s'expliquent parfois par le manque de robustesse et de transparence de ces derniers, par leur biais, par de nombreuses raisons qui tiennent souvent de la relative jeunesse de l'informatique et de ses usages. Nous avons vu que la transparence, la vérification, l'explication peuvent ouvrir la voie à une instauration de plus de confiance en ces algorithmes. Mais il ne faut pas être naïf : les intérêts en jeu sont tels qu'on imagine mal les citoyens seuls pouvoir imposer des comportements éthiques de leurs logiciels aux entreprises toutes puissantes, qui les proposent parfois. L'État doit apporter sa contribution par les lois et la régulation. Pour ce qui est des réseaux sociaux, la régulation proposée par la « Mission Facebook »[12, 13] illustre ce que pourraient être de telles lois.

Tant vis-à-vis l'individu que de la société, le contrôle des algorithmes pose de façon concrète, pour chaque contexte, la question de l'autorité que nous voulons leur accorder, celle de leur responsabilité. Mais aucune transparence, aucune perfection des algorithmes, ne pourraient seules éteindre les réticences des humains envers eux. Ces questions se rapportent aussi au pré carré décisionnel que les humains veulent conserver.

Nous disposerons bientôt de voitures autonomes qui se passeront de conducteurs humains. Certaines personnes refusent cette éventualité, voyant dans le fait de ne plus conduire sa voiture une perte de liberté (la personne devient dépendante de la machine), une forme d'infantilisation (les enfants ne savent pas conduire). Toutefois, pour des personnes incapables de conduire en toute sécurité, pour des questions d'âge ou autres, cela représentera surtout un progrès, une forme de mobilité en toute indépendance. Les avis sont contrastés.

Voulons-nous de robots-aide-soignants ? De robots-juges ? De robots-policiers… ? Dans chaque cas, les réponses ne sont pas simples, car c'est notre humanité qui est mise en question. Et les réponses évoluent dans le temps, en même temps que nous apprenons à vivre avec des logiciels de plus en plus sophistiqués. Dans ce cadre, il est important de répéter que nous ne pouvons céder de prises de décision à des algorithmes par commodité ou par paresse, que cet acte doit résulter de choix qu'il nous faut prendre collectivement.

[12] ABITEBOUL S. & POTIER F. *et al.* (2019), « Créer un cadre français de responsabilisation des réseaux sociaux : agir en France avec une ambition européenne », www.economie.gouv.fr

[13] ABITEBOUL S. & CATTAN J. (2020), « Nos réseaux sociaux, notre régulation », *Le Grand Continent*.