Note de lecture, UE 902 EC 5*

Léo Vacher

2022

Dans le présent essai, nous discuterons l'article "Ethical Implications and Accountability of Algorithms" (Martin 2018), centré sur la question des conséquences éthiques associées aux algorithmes et les responsabilités qui en découlent. Nous commencerons par une présentation synthétique de l'argumentaire de l'article en section 1, puis nous en discuterons le contenu en section 2.

1 Présentation (879 mots)

Malgré une grande discrétion, les algorithmes structurent nos vies. C'est en particulier le cas lorsqu'ils assistent des prises de décisions. Ils couvrent alors un large éventail allant de l'automatisation du contenu et du placement des publicités en ligne ou l'ordre d'apparition des résultats des moteurs de recherche jusqu'à des décisions majeures qui peuvent être cruciales pour la vie des individus telles que les décisions de remises de peine, les candidatures aux emplois et universités ainsi que les prêts bancaires et demandes d'assurances. Malgré l'indéniable bénéfice apporté par l'automatisation de ces prises de décisions - notamment en les rendant ciblés et efficaces - cette démarche peut se révéler injuste ou biaisée. Dans ce dernier cas, il devient alors légitime de se demander qui tenir pour responsable des conséquences.

Dans Martin 2018, l'auteur propose d'aborder cette question comme suit : en un premier temps, il montre que les algorithmes ne sont pas éthiquement neutres mais bien chargés de valeur (value-laden) c'est-à-dire au coeur de décisions ayant de lourds impacts moraux. En un second temps, il discute comment l'automatisation du processus de décision redistribue le poids des responsabilités entre les individus impliqués. En un troisième temps, il argumente la thèse selon laquelle les entreprises créant les algorithmes doivent être tenues pour responsable des implications éthiques de ceux-ci.

Pour montrer que les algorithmes sont chargés de valeur, on peut mettre en évidence qu'ils (1) impliquent des conséquences morales (2) renforcent ou amoindrissent les principes éthiques de la décision, et (3) favorisent ou diminuent les droits et la dignité des individus concernés. Le logiciel COMPAS, assistant les décisions de remises de peines aux Etats-Unis, est pris comme exemple par l'auteur.

1. En se basant notamment sur l'analyse établie par Propublica (Julia Angwin and Kirchner 2016), on peut montrer que non seulement COMPAS est inefficace (seul 20% des récidives prévues pour des crimes graves ont eu lieu) mais il est de plus significativement et injustement biaisé contre les minorités en présentant de fortes disparités raciales défavorisant les individus de couleur (voir également e.g. Rudin, Wang, and Coker 2020). Il est alors clair que COMPAS a des conséquences morales fortes.

2. Par design, COMPAS utilise des facteurs inaltérables et incontrôlables par l'individu tel que l'historique criminel de ses parents ou sa première arrestation par la police. En cela, il utilise des informations hors propos et injustes. Il renforce ou amoindrit ainsi les principes éthiques mis en jeux dans le processus décisionnel.

*Pour toute requête, contacter vacher.leo.etu@gmail.com

3. Pour des raisons de droit de propriété, les détails de l'algorithme ainsi que les facteurs considérés et les poids associés sont gardés secrets. En cela il réduit fortement l'autonomie, les droits et la dignité des personnes jugés, qui se trouvent face à des décisions injustifiées et ne peuvent adapter leur comportement en conséquences pour encourager une décision qui leur serait favorable.

Il est facile d'argumenter que COMPAS n'est pas un cas à part, mais que la majorité des algorithmes de décision valide un ou plusieurs des trois points discutés ci-dessus. De plus, il est clair qu'un algorithme n'est pas neutre, car il hérite des biais du programmeur qui doit définir ses critères de réussite. On peut ainsi conclure que les algorithmes ne sont pas éthiquement neutres mais bien chargés de valeur.

Il faut alors comprendre le processus de décision comme un réseau complexe impliquant de nombreux acteurs dans lequel les algorithmes occupent une nouvelle place. L'auteur cherche à établir où se situent les algorithmes dans ce réseau et quel est leur impact sur le rôle et les responsabilités des individus mis en jeu en se basant sur Akrich 1992 et Latour 1992.

D'après Akrich 1992, toute technologie est conçue dans un cadre qui la met en relation avec de nombreux autres objets et individus. Son design hérite alors des biais et des croyances du concepteur sur chacun des éléments contenus dans le réseau. Une fois la conception achevée et le produit inséré dans la société, il devient très difficile de la changer. Les biais sont alors perpétués par les concepteurs suivants de manière implicite.

D'après Latour 1992, la mise en place d'une technologie déresponsabilise les individus (e.g. des portes automatiques enlèvent au portier la responsabilité de ne pas blesser les clients). Mais les responsabilités n'ont pas pour autant disparu (par analogie avec la matière noire, Latour parle de "masse manquante"). Elles sont en fait transférées dans les choix qui sont pris explicitement ou le plus souvent implicitement lors de la conception de la technologie en question. En s'insérant dans le processus décisionnel, les algorithmes ne changent donc pas uniquement la manière dont la décision est prise, mais s'insèrent dans un réseau complexe dont ils modifient le rôle des différents acteurs à travers un transfert de responsabilité.

Comme elles sont responsable de leur conception et des choix moraux implicitement ou explicitement faits (masses manquantes), l'auteur conclut donc que les entreprises doivent être tenues responsables de leurs algorithmes et leurs implications. En particuliers, elles sont tenues de rendre leur code le plus accessible et ouvert possible en faisant en sorte d'identifier et d'expliciter tous les choix effectués lors de la conception. Le fait qu'un algorithme soit opaque ne doit pas être un argument légitime pour se déresponsabiliser, car en vendant un algorithme, l'entreprise accepte de faire part du processus de décision et n'est pas un acteur extérieur et indépendant.

2 Discussion (597 mots)

Nous argumenterons à travers trois exemples (données biaisées, métriques de l'équité et opacité) que (dans l'état actuel des choses) les algorithmes ne sont pas seulement chargés de valeur, mais fondamentalement et intrinsèquement inéquitables, rendant impossible l'attribution de la responsabilité de leurs conséquences éthiques aux entreprises seules.

Comme Martin 2018 le pointe déjà, outre les biais induis directement par les programmeurs, des biais peuvent être induits par les données utilisées par les algorithmes. En effet, même en éliminant tous les critères de détection directe, de nombreux proxy apparaissent corrélés aux caractéristiques des individus apparaissent dans les données. Ce problème est d'autant plus grave et pernicieux pour les modèles de machine learning qui s'entraînent sur des données biaisées. Il semble fortement improbable que l'existence de tels proxy puisse être totalement maîtrisés et on peut difficilement attribuer la responsabilité de leur existence aux entreprises, car ils sont des conséquences de la société en elle même.

Comme discuté dans Kleinberg, Mullainathan, and Raghavan 2016, demander l'équité d'un algorithme tel que COMPAS est impossible, car tout modèle prédictif réaliste (classificateur

imparfait et taux de base des variables différent selon les populations) ne peut pas satisfaire simultanément tous les critères de l'équité tels que calibration (le résultat le même sens pour toutes les populations) et égalité des taux d'erreur (taux d'erreur défavorable et favorable sont les mêmes pour toutes les populations). Dans l'état actuel des choses, les algorithmes sont donc intrinsèquement injustes, quel que soit l'effort effectué par l'entreprise.

Martin 2018 s'accorde également à dire que l'opacité d'un algorithme ne peut pas justifier une déresponsabilisation. Ce point reste cependant discutable. Comme Burrell 2016, il convient de distinguer l'opacité due à la restriction d'accès et l'incompétence, sur lequel l'entreprise a un impact direct et doit un certain niveau de transparence en publiant et expliquant en tout ou en partie ses algorithmes et opacité scientifique intrinsèque inhérente aux modèles de machine learning, sur laquelle l'entreprise ne peut parfois pas agir. Même si il existe des compromis important à effectuer entre opacité et performance, tout algorithme ne peut pas actuellement être rendu localement interprétable. Il semble important de distinguer également opacité globale et opacité locale. Pour effectuer des décisions éthiques, chaque individu pris dans le processus de décision devrait pouvoir exiger un retro-engineering permettant d'obtenir une justification (voir e.g. Wachter, Mittelstadt, and Russell 2017) et c'est donc cette dernière, bien plus difficile voir impossible en pratique, qui doit être contournée.

Les algorithmes sont ainsi intrinsèquement inéquitables et par conséquence les entreprises ne peuvent pas toujours être tenues pour responsable de leurs implications éthiques. En cas d'impact majeur sur la vie des individus ou de la société, l'utilisation d'un outil intrinsèquement inéquitable n'est peut-être tout simplement pas désirable. Dans le cas contraire, il est nécessaire qu'un arbitrage soit effectué pour encadrer fortement son utilisation en toute conscience des limites éthiques d'une telle technologie, incombant une part de responsabilité à l'état et au système législatif (comme cela peut-être le cas pour certaines technologies relative au nucléaire ou à la génétique). Dans certains cadres, on pourra par exemple demander l'obligation que l'individu concerné puisse avoir accès à une justification concernant la décision qui le concerne, quitte à restreindre l'usage légal aux algorithmes complètement ou partiellement localement interprétables et à demander qu'un second parti puisse prendre connaissance de cette justification. Il est également fondamental que la prévention soit effectuée pour que les individus concernés (e.g. les juges) soient formés aux limitations éthiques intrinsèque des outils qu'ils utilisent. Seulement une fois ce cadre fixé, les entreprises pourront être tenues pour responsable des conséquences morales de leurs algorithmes comme l'argumente Martin 2018.

References

Akrich, Madeleine (1992). "The De-scription of Technical Objects". In: *Shaping Technology/Building Society. Studies in Sociotechnical Change*. Ed. by Bijker, W. & Law, and J. MIT Press, pp. 205–224. URL: https://halshs.archives-ouvertes.fr/halshs-00081744.

Burrell, Jenna (2016). "How the machine 'thinks': Understanding opacity in machine learning algorithms". In: *Big Data & Society* 3.1, p. 2053951715622512. DOI: 10.1177/2053951715622512. eprint: https://doi.org/10.1177/2053951715622512. URL: https://doi.org/10.1177/2053951715622512.

Julia Angwin Jeff Larson, Surya Mattu and Lauren Kirchner (2016). "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks." In: *ProPublica*. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing#disqus_thread.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (Sept. 2016). "Inherent Trade-Offs in the Fair Determination of Risk Scores". In: *arXiv e-prints*, arXiv:1609.05807, arXiv:1609.05807. arXiv: 1609.05807 [cs.LG].

Latour, Bruno (1992). "Where are the missing masses? The sociology of a few mundane artifacts". In: *Shaping Technology / Building Society: Studies in Sociotechnical Change*. Ed. by Wiebe E. Bijker and John Law. Cambridge/MA: The MIT Press, pp. 225–258.

Martin, Kirsten (2018). "Ethical Implications and Accountability of Algorithms". In: *Journal of Business Ethics* 160.4, pp. 835–850. DOI: 10.1007/s10551-018-3921-3.

Rudin, Cynthia, Caroline Wang, and Beau Coker (Mar. 2020). "The Age of Secrecy and Unfairness in Recidivism Prediction". In: *Harvard Data Science Review* 2.1. URL: https://hdsr.mitpress.mit.edu/pub/7z10o269.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell (Nov. 2017). "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR". In: *arXiv e-prints*, arXiv:1711.00399, arXiv:1711.00399. arXiv: 1711.00399 [cs.AI].