Le mythe de l'interprétabilité et de l'explicabilité des modèles

Arthur Charpentier

Professeur, Université du Québec à Montréal

Rubinstein [2012] affirmait que « dans la théorie économique, comme dans Harry Potter, Les habits neufs de l'empereur *ou les contes du roi Salomon, nous nous amusons dans des mondes imaginaires. La théorie économique invente des contes et les appelle des modèles. Un modèle économique se situe également entre la fantaisie et la réalité [...] Le mot "modèle" semble plus scientifique que le mot "fable" ou "conte", mais je pense que nous parlons de la même chose ». Aujourd'hui, bien souvent, les modèles d'apprentissage vont construire un modèle, sur la base de données d'apprentissage, et le travail de l'actuaire sera de lui donner sens, de trouver l'histoire – la fable – qu'il est possible de raconter.*

Expliquer un modèle ?

Le rapport Villani [2018] mentionnait l'importance de l'explicabilité des algorithmes d'apprentissage automatique, reprenant ainsi un néologisme, correspondant à une exigence que l'on observe depuis un demi-siècle dans les systèmes complexes. Dans les années 1980, comme le rappelle Swartout *et al.* [1991], la Strategic Computing Initiative du département de la défense américaine lançait l'acronyme EES (Explainable Expert Systems), l'adjectif « *explainable* » donnant le nom « *explainability* ». Plus récemment, en 2016, le règlement général sur la protection des données (RGPD) imposait l'obligation de fournir « des informations utiles concernant la logique sous-jacente » de toute décision automatique. A peu près à la même époque, la loi Lemaire (1) imposait à l'administration l'obligation de communiquer à l'individu les « règles définissant tout traitement [automatique] et les principales caractéristiques de sa mise en œuvre ». En 2018, la convention 108 (2) conférait aux personnes le droit d'obtenir connaissance « du raisonnement qui sous-tend le traitement » automatique. Si les termes d'interprétabilité et d'explicabilité ne sont pas mentionnés, on y retrouve clairement l'idée (à laquelle seront rapidement associés les concepts de transparence, d'« auditabilité » ou de responsabilité des algorithmes).

Récemment, Miller [2019] a tenté une définition de cette « explicabilité ». Comme il le souligne, définir le terme « explication » a mobilisé bon nombre de philosophes, et tous semblent souligner l'importance de la causalité dans l'explication, c'est-à-dire qu'une explication se réfère forcément à des causes. Toutes

sortes de termes semblent être utilisés, pour décrire cette idée, que ce soit « interprétation » ou « justification ». Bien souvent, l'explication est donc un mode par lequel un observateur peut obtenir la compréhension. Une justification explique pourquoi une décision est bonne, mais ne vise pas nécessairement à donner une explication du processus décisionnel.

Le besoin d'explication, ou de confiance

« Et pourquoi ? » demandent presque sans fin les jeunes enfants [3]. Il existe de nombreuses raisons pour lesquelles les gens peuvent demander des explications. La curiosité est l'une des principales motivations, mais des arguments plus pragmatiques peuvent être invoqués. Pour illustrer le besoin (naturel) que nous avons d'explications, Johnston [2021] donne l'exemple simple d'une visite chez le garagiste : ma voiture est en panne, je l'amène chez le garagiste, qui la répare et me demande 900 euros. Mais, avant de remettre l'argent, je demande ce qui ne va pas et il explique que la soupape d'entrée du convertisseur catalytique était bloquée, ce qui entraînait l'usure des roulements du moteur de l'injecteur et permettait à la poussière de pénétrer dans la soupape de décharge, ce qui pouvait obstruer l'embout du tuyau de sortie, et donc, il a remplacé les roulements et mis un nouveau tuyau. Ce charabia technique a été inventé, mais le fait est que, bien souvent, l'explication n'était pas vraiment utile, car on n'a qu'une connaissance (très) limitée du fonctionnement des voitures. Ce qui a été utile, c'est que le mécanicien a réellement réparé la voiture. Et si on va plus loin, ces explications ont un pouvoir réconfortant : on se sent en confiance si l'explication semble pertinente (comme le rappellent Kästner *et al.* [2021]). Et si la réparation fonctionne, je n'ai pas envie forcément d'en savoir davantage. En revanche, si le problème persiste, je veux comprendre. C'est bien souvent l'analyse des erreurs qui est importante. En 2017, à la conférence NeurIPS (Neural Information Processing Systems), lors d'un des premiers débats à évoquer l'équité algorithmique, il avait été souligné que « *if we wish to make all systems deployed on self-driving cars safe, straightforward black-box models will not suffice, as we need methods of understanding their rare but costly mistakes* ». Lors de cette conférence, Yann LeCun soulignait que lorsqu'on présente deux modèles à des usagers (l'un extrêmement explicable et précis à 90 %, l'autre se comportant comme une boîte noire mais avec une précision supérieure de 99 %), ils choisissent toujours le plus précis. Autrement dit « *people don't really care about interpretability but just want some sort of reassurances from the working model* » ; ce qui signifie que l'interprétabilité n'est pas importante si on est convaincu que le modèle fonctionne bien, dans les conditions dans lesquelles il est censé fonctionner. N'est-ce pas ce qui se passe quand je monte dans un avion, ou que je subis une opération chirurgicale ?

La justice, un exemple à suivre ?

Il y a presque un siècle, Ernest Burgess commençait à suivre une cohorte de 3 000 condamnés, libérés sous condition, et il parvint à identifier vingt-deux paramètres permettant de distinguer ceux qui réussiraient leur probation et ceux qui échoueraient. Dans les années 1950, les époux Glueck et d'autres chercheurs ont poursuivi son analyse, multipliant les études consacrées aux facteurs de récidive pour construire de multiples échelles de prédiction des risques. Ces instruments s'appuient sur des méthodes statistiques inspirées des pratiques assurantielles, afin de déterminer les niveaux de risque associés à un groupe de délinquants présentant des caractéristiques communes, et, sur la base de ces corrélations, de prédire le comportement criminel futur d'un individu spécifique, comme le raconte Harcourt [2007]. Cette « justice actuarielle » va rendre possible le développement d'outils d'aide à la décision, permettant aux juges d'avoir des scores de récidive ou de dangerosité. Mais ces outils ne sont qu'une aide, et les juges se doivent ensuite de motiver leur décision, de fournir une explication.

Car en matière judiciaire, l'individualisation de la peine constitue l'un des principes fondamentaux du droit pénal français, comme le soulignait Saleilles [1897]. Pour Dadoun [2018], pour que les exigences d'un procès équitable soient respectées, « l'accusé doit être à même de comprendre le verdict qui a été rendu ». Cette motivation des jugements et arrêts peut être démonstrative, narrative ou péremptoire, comme le montrent Zerouki-Cottin *et al.* [2020]. Dans la majorité des cas, la motivation est une explication de la décision, sauf la motivation péremptoire qui relève plus de l'affirmation que de l'explication (4). La version narrative contient une explication du contexte de l'infraction, il s'agit de raconter l'histoire. C'est probablement l'approche qui s'apparente le plus à ce que nous avons en tête quand on demande une explication (5). La justice peut-elle servir d'exemple pour comprendre ce que devrait être l'explicabilité ?

En justice, la motivation se base sur des faits et des avis d'experts. Comme le souligne Coche [2011], « pour apprécier la dangerosité des personnes poursuivies ou condamnées, le législateur multiplie le recours aux expertises. Cependant, les expertises de dangerosité, non seulement ne sont pas fiables, mais elles ne peuvent pas le devenir. Elles créent donc l'illusion, sans cesse déçue, d'une appréciation qui serait scientifique de la dangerosité ». Cette évaluation par les experts est appelée « évaluation clinique », et comme le notent Dubourg et Gautron [2015], des centaines d'études considèrent que, dans le contexte d'évaluer un risque de récidive, ces évaluations cliniques non structurées présentent des estimations proches du hasard, historiquement aux Etats-Unis, reprochant aux cliniciens une surévaluation des risques de récidive.

La même critique est faite en France où « l'utilisation de concepts psychanalytiques demeure prédominante tant en psychiatrie générale que dans le contexte des expertises. Or, ces concepts ne possèdent aucun lien théorique avec les comportements délictuels à prédire. Ainsi, la méthode basée sur un jugement clinique non structuré est une évaluation subjective, non validée scientifiquement, et fondée sur des corrélations intuitives... Il n'est donc plus question de tenter de comprendre ou d'expliquer ».

Dans la majorité des expertises, la base scientifique de l'explication est qu'il n'y a pas de fumée sans feu. Dubourg et Gautron [2015] montrent que l'hypothèse d'une dangerosité et/ou d'un risque de récidive est très souvent validée ; 75 % des condamnés se voyant attribuer un pronostic défavorable par au moins un expert au fil du processus. Delacrausaz et Gasser [2012] observent que l'expert se contente « d'extraire tel ou tel élément d'observation pour en déduire toutes sortes de raisonnements, sans expliquer ni les motivations de ses choix, ni les fondements théoriques sur lesquels il les base ». Bien souvent, à partir des mêmes éléments factuels présents dans le dossier, l'avocat de la défense et le procureur vont avoir deux explications radicalement différentes.

Les deux cultures

Le monde de la justice, avec des juges, des procureurs, des avocats qui sont très majoritairement, en France en tout cas, des femmes et des hommes « de lettres », est très loin de la culture statistique des données. A la fin des années 1950, le baron Charles Percy Snow affirmait que la vie intellectuelle de la société occidentale se divise essentiellement en deux cultures distinctes, celle des sciences et celle des humanités, et que la culture partagée tend à disparaître. Le monde des chiffres et le monde des récits.

Glenn [2000] reprenait cette idée en expliquant que le processus de sélection des risques d'un assureur avait deux visages (comme le dieu romain Janus) : celui qui est présenté aux régulateurs et aux assurés, et celui qui est présenté aux souscripteurs. Il y a d'un côté le visage des chiffres, des statistiques et de l'objectivité. De l'autre, il y a le visage des récits, du caractère et du jugement subjectif. Paul Meehl, en 1954, parlait de « *mechanical prediction* » pour décrire les modèles actuariels. La rhétorique de l'exclusion de l'assurance (basée sur des chiffres objectifs), par exemple, forme

ce que Brian Glenn appelle « le mythe de l'actuaire », à savoir « une situation rhétorique puissante dans laquelle les décisions semblent être fondées sur des critères déterminés objectivement alors qu'elles sont aussi largement fondées sur des critères subjectifs ». Glenn [2003] allait plus loin, affirmant que « les assureurs peuvent évaluer les risques de nombreuses façons différentes en fonction des histoires qu'ils racontent sur les caractéristiques qui sont importantes et celles qui ne le sont pas [...] Le fait que la sélection des facteurs de risque soit subjective et dépendante des récits de risque et de responsabilité a joué dans le passé un rôle bien plus important que le fait que quelqu'un avec un poêle à bois se voit facturer des primes plus élevées [...] Pratiquement tous les aspects de l'industrie de l'assurance sont fondés sur des histoires d'abord et sur des chiffres ensuite ». Cette importance de la narration se retrouve dans la fameuse phrase de George Box « *all models are wrong but some models are useful* ». Autrement dit, les modèles sont, au mieux, une fiction intéressante.

L'apprentissage machine et les boîtes noires

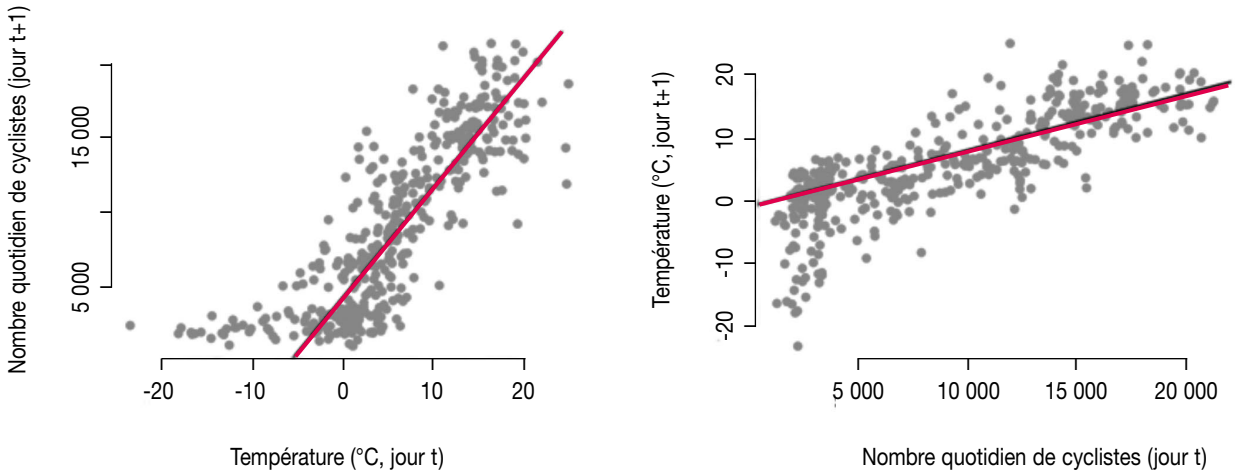
Mais il ne faut pas se tromper sur ce qu'on cherche à expliquer. Par exemple, expliquer ce que fait un algorithme d'apprentissage est assez simple : il tente de minimiser un objectif (l'écart entre ce qu'il prédit et ce qui est observé) sur la base d'algorithmes d'optimisation plus ou moins compliqués (par un algorithme de Newton-Raphson pour une simple régression logistique, ou un algorithme de rétro-propagation sur des portions de la base pour de l'apprentissage profond – réseau de neurones avec plusieurs couches cachées). Un algorithme dit « des plus proches voisins » consiste à dire que la fréquence de sinistres automobiles d'un individu sera la fréquence moyenne des personnes les plus proches de cet individu (en termes de caractéristiques : même expérience de conduite, conduisant le même type de véhicule, parcourant la même distance, etc). Cet algorithme est simple à expliquer. La partie ardue consiste à interpréter le modèle construit, à rendre intelligible les prédictions. Pasquale [2015] a souligné que les algorithmes d'apprentissage machine sont caractérisés par leur opacité et leur « incompréhensibilité », parfois appelée « propriétés de boîte noire ».

En réponse, ou de manière concomitante, une demande de transparence des algorithmes, voire des codes informatiques, a été formulée, comme Citron et Pasquale [2014] ou Mittelstadt *et al.* [2016]. Cela dit, l'algorithme « des plus proches voisins » est transparent et simple, ce sont les données qui sont importantes : il est impossible de faire une prévision sans avoir accès aux données (contrairement à une régression linéaire qui donne une fonction numérique). Burrell [2016] et Laat [2017] ont noté que le manque de transparence était en partie dû au comportement des entités de développement et d'utilisation qui refusent de divulguer les algorithmes, ou même simplement les règles et critères de décision programmés à des parties externes pour des raisons de protection des secrets commerciaux, de protection des droits d'auteur, de protection des données (lorsque les systèmes informatiques contiennent des données personnelles de tiers) ou par prudence contre des ajustements comportementaux ciblés par les personnes concernées. Ce dernier point avait été souligné dans le cas de l'algorithme de Facebook, comme raconté dans Charpentier [2021].

Le modèle linéaire comme référence de « boîte blanche » ?

Un des soucis est que le modèle le plus simple, le modèle linéaire, est souvent décrit comme un modèle « interprétable », mais par construction, cette interprétation est fallacieuse. En effet, pour estimer un modèle linéaire, pour lier deux variables x et y, c'est la corrélation entre les deux variables qui compte, et qui est souvent interprétée comme une relation causale, pour la simplicité de la narration (voir figure 1 p. 113).

Figure 1 - Cyclistes à Stockholm, avec la température moyenne et le nombre de cyclistes, par jour



Source : auteur.

La figure 1 présente un couple de données tirées d'une même base, où x_t est le nombre de cyclistes dans une rue à Stockholm, le jour t (de 2014), et y_t est la température moyenne le jour t, à Stockholm. Sur la partie gauche, on a représenté (y_{t-1}, x_t) et un modèle linéaire ($x = \alpha_0 + \alpha_1 y + \eta$), supposé pertinent quand la température est au-dessus de 0°C. La pente est significativement non nulle, on a un R^2 excédant 75 %, et l'interprétation serait « le nombre de cyclistes sur la route croît avec la température, un degré de plus apportant 750 cyclistes de plus par jour sur la route ». Ou une explication plus succincte serait « les habitants de Stockholm préfèrent faire du vélo quand il fait chaud ». Sur la partie droite, on a représenté (x_{t-1}, y_t) et là encore, un modèle linéaire ($y = \beta_0 + \beta_1 x + \xi$) semble avoir du sens, en excluant la partie la plus à gauche de la courbe (quand x est faible). Là encore, la pente est significativement non nulle, on a un R^2 légèrement inférieur à 75 %, et l'interprétation serait « la température croît avec le nombre de cyclistes sur la route, chaque millier de cyclistes en plus faisant augmenter la température d'1°C ». Là encore, en simplifiant, « on peut lutter contre le réchauffement climatique en limitant le nombre de vélos sur la route ». A partir des données, et des données seulement, puis-je affirmer qu'une des explications données est plus valide que l'autre ?

Un vœu pieu ?

On demande de pouvoir comprendre et interpréter toute prévision algorithmique, mais n'est-ce pas trop ambitieux ? Car comme l'aurait dit saint Augustin, « si personne ne me demande ce qu'est le temps, je sais ce qu'il est ; et si on me le demande et que je veuille l'expliquer, je ne le sais plus ». C'est aussi ce que notait plus récemment Kahneman [2012], introduisant les notions de système 1 / système 2 (les deux vitesses de la pensée). Le système 1 est utilisé pour la prise de décisions rapides : il nous permet de reconnaître les gens et les objets, nous aide à orienter notre attention, et nous encourage à craindre les araignées. Il est basé sur des connaissances stockées en mémoire et accessibles sans intention, et sans effort. On peut l'opposer au système 2, qui permet une prise de décision plus complexe, exigeant de la discipline et une réflexion séquentielle. Dans la majorité des cas, on prend des décisions sans vraiment pouvoir les expliquer, et sans que ce soit préoccupant. Car c'est probablement autre chose que l'on cherche dans l'explicabilité. Nous avions mentionné la confiance, mais il y a aussi l'importance de l'équité, partiellement évoquée dans

le contexte légal : l'explication importe peu si la décision semble juste.

Notes

1. Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique.

2. Convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel.

3. Selon la littérature pédiatrique, entre trois et quatre ans (davantage si on a la chance d'avoir des enfants à la curiosité insatiable).

4. « Ces affirmations péremptoires ne constituent pas une motivation véritable, mais sont tout au contraire la négation de la motivation » affirmaient Zerouki-Cottin et al. *[2020].*

5. La pratique est bien entendu plus complexe : dans de nombreux cas, la motivation n'est pas donnée, à moins que l'une des parties ne fasse appel. Ce qui n'est pas sans faire penser à Manguel [2020]. « Alice sait d'instinct que la logique est pour nous le moyen de donner du sens à ce qui n'en a pas et à en découvrir les règles secrètes, et elle l'applique impitoyablement, même chez ses aînés et supérieurs, qu'elle se trouve face à la duchesse ou au chapelier fou. Et quand les arguments s'avèrent inopérants, elle insiste pour, à tout le moins, rendre évidente l'absurdité de la situation. Quand la reine de cœur exige que la cour rende "la condamnation d'abord... et le jugement ensuite", Alice répond très justement : "Mais c'est de la bêtise !" C'est bien la seule réponse que méritent la plupart des absurdités dans notre monde. »

Bibliographie

Burrell J., "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms", *Big Data & Society*, 2016.

Charpentier A., « Une mesure ne peut pas être un objectif », *Risques,* n° 125, 2021, pp. 120-125.

Charpentier A., « L'intelligence artificielle dilue-t-elle la responsabilité ? », *Risques*, n° 114, 2018, pp. 145-150.

Citron D. K. ; Pasquale F. ; "The Scored Society: Due Process for Automated Predictions", *Washington Law Review*, vol. 89, 2014.

Coche A., « Faut-il supprimer les expertises de dangerosité », *Revue de science criminelle et de droit pénal comparé*, n° 1, 2011, pp. 21-35.

Dadoun A., « L'obligation constitutionnelle de motivation des peines », *Revue de science criminelle et de droit pénal comparé*, n° 4, 2018, pp. 805-827.

Delacrausaz Ph. ; Gasser J., « La place des instruments d'évaluation du risque de récidive dans la pratique de l'expertise psychiatrique pénale : l'exemple lausannois », *L'information psychiatrique*, vol. 88, n° 6, 2012, pp. 439-443.

Dubourg E. ; Gautron V., « La rationalisation des outils et méthodes d'évaluation : de l'approche clinique au jugement actuariel », Criminocorpus, *Revue hypermedia, Histoire de la justice, des crimes et des peines*, 2015.

Glenn B. J., "Postmodernism: the Basis of Insurance", *Risk Management and Insurance Review*, vol. 6, n° 2, 2003, pp. 131-143.

Glenn B. J., "The Shifting Rhetoric of Insurance Denial", *Law and Society Review*, vol. 34, 2000, pp. 779-808.

Harcourt B. E., *Against Prediction*, University of Chicago Press, 2007.

Johnston D., "Explainable Models Are Overrated", LinkedIn, 25 janvier 2021. *https://www.linkedin.com/pulse/explainable-models-overrated-david-johnston/*

Kahneman D., *Système 1 / Système 2 : les deux vitesses de la pensée*, Flammarion, 2012.

Kästner L. ; Langer M. ; Lazar V. ; Schomäcker A. ; Speith T. ; Sterz S., "On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness", *IEEE*, 29th International Requirements Engineering Conference Workshops (REW), 2021, pp. 169-175.

Laat P. B. (de), "Big Data and Algorithmic Decision-Making: can Transparency Restore Accountability?",

ACM Computers and Society, vol. 47, n° 3, 2017, pp. 39-53.

Manguel A., *Monstres fabuleux*, Actes Sud, 2020.

Miller T., "Explanation in Artificial Intelligence: Insights from the Social Sciences", *Artificial Intelligence*, vol. 267, 2019, pp. 1-38.

Mittelstadt B. D. ; Allo P. ; Taddeo M. ; Wachter S. ; Floridi L., "The Ethics of Algorithms: Mapping the Debate", *Big Data & Society*, 2016.

Pasquale F., *The Black Box Society. The Secret Algorithms that Control Money and Information*, Harvard University Press, 2015.

Rubinstein A., *Economic Fables*, Open book publishers, 2012.

Saleilles R., *L'individualisation de la peine : étude de criminalité sociale*, F. Alcan, 1897.

Swartout W. ; Paris C. ; Moore J., "Explanations in Knowledge Systems: Design for Explainable Expert Systems", *IEEE Expert*, vol. 6, n° 3, 1991, pp. 58-64.

Villani C., *Donner du sens à l'intelligence artificielle*, mission parlementaire confiée par le Premier ministre Edouard Philippe, 2018.

Zerouki-Cottin D. ; Perrocheau V. ; Milburn P., « L'obligation de motivation des décisions criminelles en France : de la loi aux pratiques. Analyse empirique de la motivation des décisions des cours d'assises », *Revue juridique Thémis*, vol. 54, n° 1, 2020.