Quelle responsabilité pour les algorithmes ?

Rodolphe Bigot

Maître de conférences, Université de Picardie Jules Verne

Arthur Charpentier

Professeur, Université du Québec à Montréal

Historiquement, les algorithmes se contentaient de fournir une aide à la décision, laissant à un être humain le rôle de prendre la décision, mais des expériences sont en cours, avec des systèmes autonomes prenant des décisions, que ce soient les systèmes de conduite de voiture ou les algorithmes de justice prédictive, comme le montre Huss et al. *[2018]. Cette autonomie, qui signifie fondamentalement la « faculté d'agir librement » désigne aussi l'idée « de se gouverner par ses propres lois ». Mais quelle est la responsabilité du décisionnaire dans le cas d'une prédiction qui entraînerait un préjudice ?*

Comprendre et prévoir

Dans Bigot et Charpentier [2019], nous avions questionné l'évolution de la notion de responsabilité au regard des évolutions des deux derniers siècles, mais un point essentiel est que, fondamentalement, un homme (1) est dans la majorité des cas responsable de ses actes. Pourquoi ? Probablement parce qu'un homme est censé pouvoir imaginer (pris dans le sens d'anticiper), comprendre et prévoir que ses actions auront des causes et des conséquences. Les animaux ne sont pas jugés responsables de leurs actes (mais leurs maîtres, propriétaires ou gardiens, le sont (2)). Dans ses expériences, Ivan Pavlov avait conditionné des chiens, qui se mettaient à saliver quand une sonnette retentissait, annonçant l'arrivée d'un repas. Ils avaient ainsi associé la sonnette au repas, mais il n'y avait pas de mécanisme causal, simplement une forme de compréhension instinctive dont disposent tous les animaux.

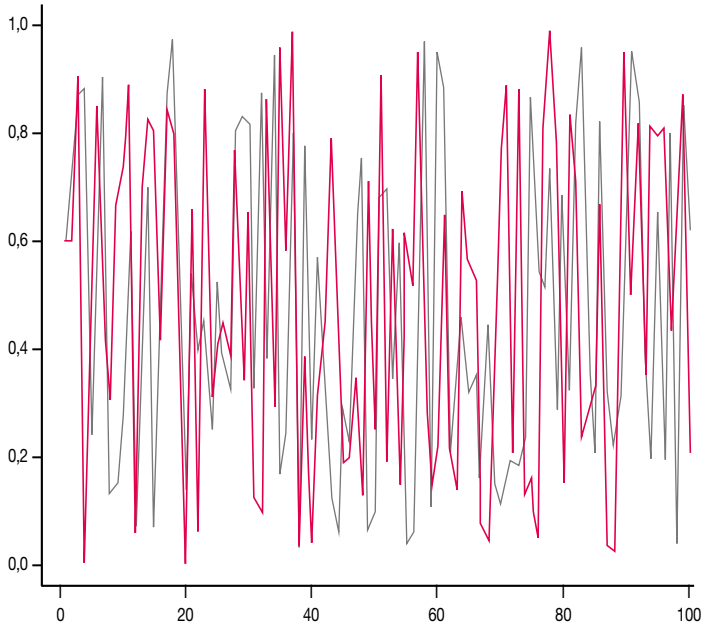
Comprendre, c'est connecter des connaissances et en déduire des formes de lois universelles, comme en physique. On cherche à construire une théorie qui explique des faits, les relie au reste des connaissances et permet alors d'anticiper. C'est le principe de l'abstraction. L'abstraction est un processus fondamental dans la compréhension d'un phénomène, l'observation suffit rarement. Ainsi, au XVIIe siècle Galilée énonce le principe d'inertie (postulant qu'en l'absence de force, les corps se déplacent en ligne droite à vitesse constante) contre toutes les expériences faites sur Terre. À l'époque – mais c'est probablement encore vrai

aujourd'hui –, la perception est plus proche de ce qu'avait énoncé Aristote, à savoir que la force était nécessaire pour entretenir le mouvement. On peut penser à l'expérience de pensée proposée par Galilée sur la chute des corps, pour contredire la théorie aristotélicienne du mouvement, selon laquelle la vitesse d'un corps en chute libre est proportionnelle à son poids – il proposait de lâcher dans le vide deux corps de masse différente en les reliant entre eux par une corde. Lorsqu'il affirme que tous les corps tombent à la même vitesse, cette loi n'est pas une synthèse de faits empiriques connus, mais bien une compréhension abstraite des phénomènes. C'est d'ailleurs ce qu'affirmait Weber [1905] : « l'attribution des effets aux causes prend place à travers un processus de pensée qui inclut une série d'abstractions. La première et la plus décisive a lieu quand nous concevons que l'une ou plusieurs des composantes causales sont modifiées dans une certaine direction et que nous nous demandons si, dans les conditions ainsi modifiées, le même effet [...] ou un autre effet "serait attendu" ».

Mais il est possible de comprendre, sans pouvoir prévoir. Dans Charpentier [2018a], il était expliqué comment générer du chaos de manière déterministe. Sur la figure 1, on voit l'évolution de deux suites définies par récurrence, avec deux valeurs initiales différentes, avec un écart de 1 sur dix mille au point de départ. Très rapidement les deux séries divergent et sont alors considérées comme statistiquement indépendantes. Poincaré (1908) disait (en parlant des lois naturelles) que si « cela nous permet de prévoir la situation ultérieure avec la même approximation, c'est tout ce qu'il nous faut, nous disons que le phénomène a été prévu ; mais il n'en est pas toujours ainsi, il peut arriver que de petites différences dans les conditions initiales en engendrent de très grandes dans les phénomènes finaux [...] La prédiction devient impossible et nous avons un phénomène fortuit ».

De manière inverse, il est possible de prévoir sans comprendre. Comprendre, c'est souvent énoncer une loi générale, à partir du constat que les mêmes causes ont toujours les mêmes effets. Mais, comme le notait Maxwell [1876] : « *to make this maxim intelligible, we must define what we mean by the same causes and the same effects, since it is manifest that no event ever happens more than once, so that the causes and effects cannot be the same in all aspect* ». Et effectivement, face à une situation donnée, une voiture autonome cherchera des situations semblables qui auraient pu être vécues. On est souvent tenté de voir des liens causaux alors qu'il n'y a parfois que des corrélations, « *cum hoc ergo propter hoc* » (avec ceci, donc à cause de ceci), voire parfois de simples coïncidences. La corrélation tient du fait que si x cause y et x cause z, alors y et z seront corrélés, sans qu'aucun lien de causalité n'existe. L'exemple classique dans les écoles primaires est la corrélation entre un nombre de fautes dans une dictée (y) et la pointure des souliers (z), sa variable causale étant ici l'âge des élèves (x). La coïncidence est d'autant plus facile à obtenir en grande dimension : si on dispose d'une variable d'intérêt, et d'une centaine de variables indépendantes de cette variable, cinq de ces variables (en moyenne) seront

Figure 1 - Simulation de nombres (pseudo) aléatoires par la méthode de Sedgewick



Lecture : avec $u_n = x_n/m$ où $u_n = (ax_{n-1}+c)$ modulo m (avec ici $m = 10^8$), la première série (trait rouge) commençant avec x_1 valant 6 millions, ou $u_1 = 0{,}6$ et la seconde (trait gris) $u_1 = 0{,}60001$.

Source : auteurs.

« significativement corrélées » avec notre variable d'intérêt, avec un seuil de 95 %. « À dire vrai, *big data* signifie surtout le franchissement d'un seuil à partir duquel nous serions contraints (par la quantité, la complexité, la rapidité de prolifération des données) d'abandonner les ambitions de la rationalité moderne consistant à relier les phénomènes à leurs causes, au profit d'une rationalité que l'on pourrait dire post-moderne, indifférente à la causalité, purement statistique et inductive, se bornant à repérer des modèles, c'est-à-dire des motifs formés par les corrélations observées entre des données indépendamment de toute explication causale. La répétition de ces motifs au sein de grandes quantités de données leur confère une valeur prédictive » écrit Sauvé [2014].

Quels algorithmes, quelles machines ?

Un algorithme désigne simplement un ensemble de règles opératoires fini permettant de résoudre un problème. On peut penser à l'analogie avec les recettes de cuisine ou les procédures bureaucratiques, comme le montre Charpentier [2018b]. On se doit toutefois de distinguer entre un algorithme d'automatisation et un algorithme d'apprentissage, comme le rappelle Godefroy [2017]. Les scores des banques ou des assureurs sont (encore) souvent du premier type, ce qui permet d'expliquer à un client la raison du refus d'un crédit hypothécaire : un score est construit comme une moyenne pondérée (3) de différentes grandeurs (comme l'âge, le salaire, la durée d'emploi, etc.), que l'on compare à un seuil. Ces algorithmes, classiques en assurance, présentent l'avantage d'être compréhensibles, avec un bon pouvoir prédictif. La seconde classe d'algorithmes fait gagner en précision quant aux prédictions, mais le prix à payer est la construction de boîtes noires (ou de machines trop « intelligentes » pour être intelligibles).

Les algorithmes d'apprentissage « apprennent » par induction en cherchant des corrélations permettant d'améliorer la prévision, avec des allers-retours constants, réitératifs (on pourra penser à la validation croisée), ce qui rend difficile la compréhension du processus retenu. Cette approche inductive fait la force, mais aussi la faiblesse, de ces algorithmes. Comme le note Domingos [2012], « *induction is a vastly more powerful lever than deduction, requiring much less input knowledge to produce useful results, but it still needs more than zero input knowledge to work. And, as with any lever, the more we put in, the more we can get out [...] Machine learning is not magic; it cannot get something from nothing. What it does is get more from less.* » Dans les algorithmes d'apprentissage, on ne trouve pas d'arbres décisionnels figés (si... alors...), mais une construction évolutive, comme le rappelle Reigeluth [2016], qui les dote de trois facultés : la mémoire, l'adaptation, la généralisation. On peut penser aux algorithmes par renforcement, où on regarde dans le passé des situations (ou des états de la nature semblables), les actions qui ont été prises, et les conséquences qui ont été produites. On peut alors réessayer – ou explorer – et tenter autre chose (et apprendre davantage) (4). Ce sont les algorithmes que l'on voit arriver dans les machines dites autonomes.

Mais ces machines, si on peut les qualifier d'autonomes, n'ont pas de volonté propre, ou de libre arbitre : elles prennent des décisions qui vont maximiser une fonction dite « objective », tout en respectant un ensemble de contraintes. Si elle peut s'adapter à un nouvel inconnu, elle donne l'impression de comprendre, mais comme le dit la boutade, un algorithme qui peut identifier des objets sur une image peut reconnaître un chien, mais la machine ne sait pas ce qu'est un chien (5) (c'est « trop robot pour être vrai » aurait dit Jacques Prévert). C'est cette indétermination dans le processus autonome de prise de décision qui pose des questions quant à la responsabilité des machines dites autonomes. En 2016, le Parlement européen notait que « dans l'hypothèse où un robot puisse prendre des décisions de manière autonome, les règles habituelles ne suffiraient pas à établir la responsabilité juridique pour des dommages causés par un robot ».

De la responsabilité des machines, le cas des véhicules autonomes

Avant d'aller trop loin, peut-être convient-il de rappeler, qu'aujourd'hui, le véhicule autonome n'existe pas vraiment. À ce jour, seules diverses formes de délégation de conduite sont autorisées et expérimentées, laissant une place plus ou moins grande aux technologies, au passager ou à une personne à l'extérieur de l'habitacle. Le système de classification de la Society of Automotive Engineers (SAE) comporte six niveaux, représentés dans le tableau 1.

Tableau 1 - Système de classification de la SAE

		Direction, accélération, décélération	Surveillance de l'environnement de conduite	Manœuvres de conduite dynamique	Fonctions (modes de conduite)
0	Aucune automatisation	Conducteur	Conducteur	Conducteur	
1	Aide à la conduite	Conducteur	Conducteur	Conducteur	Certains modes de conduite
2	Automatisation partielle	Système	Conducteur	Conducteur	Certains modes de conduite
3	Automatisation conditionnelle	Système	Système	Conducteur	Certains modes de conduite
4	Automatisation élevée	Système	Système	Système	Certains modes de conduite
5	Automatisation complète	Système	Système	Système	**Tous** modes de conduite

Source : Society of Automotive Engineers (SAE) [2016].

On peut aussi rappeler une subtilité, évoquée dans Bigot et Charpentier [2019] : la causalité scientifique n'est pas la causalité juridique. En fait, la causalité juridique résulte de la qualification juridique des événements, pour reprendre Radé [2012]. La causalité scientifique suppose une succession automatique d'événements, sans intervention de la volonté, sans intention. L'interrogation des scientifiques sur l'interprétabilité des modèles n'est alors qu'un maillon de la chaîne. Et si les scientifiques sont perplexes, les juristes le seront davantage (6).

Les algorithmes d'apprentissage posent des soucis en raison de l'indétermination en matière d'imputation de la responsabilité en cas de dommage, s'il n'y a pas d'erreur de conception ou de mauvaise utilisation. En 2018, un procès (fictif) avait eu lieu en France, comme le rappelle Prévost [2018], posant la question de la responsabilité par suite d'un accident (imaginaire). Comme le rappelle le tableau 1, tous les systèmes laissent (ou imposent ?) un rôle au « conducteur » (car il reste une personne identifiée comme telle, ayant la possibilité de repasser à un mode de conduite dit « manuel »), et la responsabilité lui incomberait. Comme le note Noguéro [2019], la responsabilité du fait des choses est consacrée à l'article 1242 du Code civil qui énonce que l'on « est responsable non seulement du dommage que l'on cause par son propre fait, mais encore de celui qui est causé par le fait des personnes dont on doit répondre, ou des choses que l'on a sous sa garde. » Le souci est que le but même de ces voitures est de laisser à leurs utilisateurs la liberté de ne jamais devoir se préoccuper de leur conduite une fois la destination enregistrée. Il est alors difficile de comprendre, dans le même temps, qu'on les désignerait toujours comme ayant sur celles-ci un pouvoir d'usage, de contrôle et de direction.

Bensoussan [2015] note qu'aux États-Unis, dans certains États (par exemple au Nevada), les robots se sont vu reconnaître certains attributs de la personne morale, sans être toutefois visés comme tels. Ils sont alors immatriculés et répertoriés dans un fichier spécialement dédié, et ils se voient surtout assigner un capital, dont la fonction première consiste au fond à les assurer directement pour leur permettre de répondre des dommages qu'ils causeraient dans leurs interactions en environnement ouvert, comme le rappelle Coulon [2016]. Or qui, en amont, affecterait un capital susceptible de compenser les dommages d'un grave accident, pouvant représenter plusieurs millions – voire dizaines de millions – d'euros d'indemnité ? le fabricant ? le vendeur ? Dans tous les cas, on serait loin de la garantie apportée par l'assurance obligatoire de responsabilité qui est, en France, illimitée (7). Cette proposition génère notamment le problème « de déresponsabilisation des intervenants : quelle que soit

l'hypothèse, concepteurs, intégrateurs ou encore utilisateurs sauront que leur responsabilité ne sera jamais engagée et que, in fine, l'assurance paiera au moyen d'un fonds de garantie financé par les sociétés de robotique », selon Touati [2017].

D'aucuns, comme Harari [2018], sont convaincus qu' « un chauffeur qui prédit les intentions d'un piéton, un banquier qui évalue la crédibilité d'un emprunteur potentiel et un avocat qui jauge de l'état d'esprit à la table des négociations ne s'en remettent pas à la sorcellerie. À leur insu, leurs cerveaux reconnaissent plutôt des configurations biochimiques en analysant les expressions du visage, les tons de la voix, les mouvements de main et même les odeurs corporelles. Une intelligence artificielle (IA) équipée de bons capteurs pourrait faire tout cela de manière bien plus précise et fiable qu'un être humain ». En faisant ainsi entrer « Mozart dans la machine », les véhicules autonomes élimineraient les principaux facteurs de risques à l'origine des accidents mortels (abus d'alcool, excès de vitesse et distraction). Il est alors avancé que « bien qu'ils puissent souffrir de leurs problèmes et limites propres, et que certains accidents soient inévitables, le remplacement de tous les conducteurs humains par des ordinateurs devrait réduire d'environ 90 % le nombre de morts et de blessés sur la route. Autrement dit, le passage aux véhicules autonomes est susceptible de sauver un million de vies chaque année » selon le comptage d'Harari [2018]. Dès lors, se dirigerait-on vers une responsabilité résiduelle ? En contrepoint, l'IA pourrait aussi apprendre le concept de mérite, qui peut s'inscrire contre la justice. Duru-Bellat [2019] rappelle que « le mérite a la cote. Avec lui, l'idée que chacun est responsable de ce qui lui arrive, de ses succès comme de ses échecs, et l'espérance qu'en récompensant talents et efforts, on produira une société juste et efficace. La mise en exergue constante du mérite, sans tenir compte des inégalités (sociales, de genre, d'origine, etc.) est pourtant tout sauf anodine. Elle engendre de nombreux effets pervers ». Pire, l'IA pourrait encore apprendre le mensonge humain comme l'imagine McEwan [2020], et commettre des fautes artificiellement volontaires, voire intentionnelles, sans compter, naturellement, les bugs, piratages et actes criminels qui pourraient alors générer des dommages de masse intensifiés par la mise en réseau.

Les accidents survenus lors des tests sont souvent révélateurs de l'ambiguïté qui existe : lors de l'un des accidents d'une *google car*, le problème venait de ce que le passager du véhicule semi-autonome, doutant de l'efficacité de celui-ci, avait lui-même pris une mauvaise décision en appuyant soudainement sur la pédale de frein. Le robot s'était donc arrêté plus tôt que prévu, puisque l'algorithme, contrôlant l'intensité et la distance de freinage, avait été modifié. C'est bien souvent l'interaction entre l'homme et la machine qui pose un problème. Et il n'est pas étonnant de voir le législateur vouloir intervenir dans ce débat, et le débat prévu par la Commission européenne dans les semaines à venir est grandement attendu par tous les acteurs.

Notes

1. Dans tout cet article « homme » désigne un être humain : il ne s'oppose pas à « femme » mais à « machine » ou « robot ».

2. Article 1243 du Code civil : « Le propriétaire d'un animal, ou celui qui s'en sert, pendant qu'il est à son usage, est responsable du dommage que l'animal a causé, soit que l'animal fût sous sa garde, soit qu'il fût égaré ou échappé ».

3. Les poids sont fixes du point de vue du banquier ; ils ont été estimés à partir de modèles de régression sur des historiques de données selon la terminologie statistique (l'informaticien parlera d'entraînement de l'algorithme sur une base d'apprentissage).

4. Conceptuellement, ces algorithmes n'ont rien de nouveau puisque le formalisme a été établi à la fin des années 1980. La puissance de ces algorithmes a toutefois été révélée lorsque des machines ont battu les meilleurs joueurs de go avec cette stratégie, et ont gagné à des jeux vidéo sans en avoir appris la règle (Minh et al. *[2015]).*

5. On peut penser à l'expérience de Ribeiro et al. *[2016] qui visait à construire un algorithme distinguant un chien d'un loup sur des photos, et qui avait un pouvoir prédictif élevé, mais dont la stratégie semblait être relativement simple : s'il y a de la neige sur la photo c'est un loup (les photos d'apprentissage montraient toutes des loups dans la neige). On peut s'interroger sur les problèmes qui pourraient se poser en justice prédictive avec l'utilisation de photos.*

6. Même si Brun [2007] note que la causalité juridique cherche la cause « la plus raisonnable » afin de « rendre la décision la plus juste », ce qui pourrait être en un sens plus simple que l'objectif de compréhension que se fixent les scientifiques.

7. Selon l'article R. 211-7 du Code des assurances : « L'assurance doit être souscrite sans limitation de somme en ce qui concerne les dommages corporels et pour une somme au moins égale à celle fixée par arrêté du ministre chargé de l'Économie, laquelle ne pourra être inférieure à 1 million d'euros, par sinistre et quel que soit le nombre de victimes, en ce qui concerne les dommages aux biens ».

Bibliographie

Bensoussan A., « Le droit de la robotique : aux confins du droit des biens et du droit des personnes », *Revue des juristes de Sciences Po*, n° 10, hiver 2015.

Bigot R. ; Charpentier A., « Repenser la responsabilité, et la causalité », *Risques*, n° 120, 2019.

Brun P., « Causalité juridique et causalité scientifique », *Revue Lamy droit civil*, vol. 40, n° 2630, 2007.

Charpentier A., « Histoire du hasard et de la simulation », *Risques*, n° 116, 2018a.

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

Coulon C., « Du robot en droit de la responsabilité civile : à propos des dommages causés par les choses intelligentes », *Responsabilité civile et assurances*, Ed. LexisNexis, avril 2016.

Domingos P., "A Few Useful Things to Know about Machine Learning", *Communications of the ACM*, vol. 55, n° 10, 2012.

Duru-Bellat M., *Le mérite contre la justice*, Sciences Po Les Presses, réédition, 2019.

Godefroy L., « Les algorithmes : quel statut juridique pour quelles responsabilités ? », *Communication commerce électronique*, n° 11, 2017, pp.16-20.

Harari Y. N., *21 leçons pour le XXIe siècle*, Albin Michel, 2018.

Huss J.-V. ; Legrand L. ; Sentis T., « Les enjeux éthiques de la justice prédictive », Livre blanc, École de droit de Sciences Po, 2018. *https://bit.ly/2sNfOhv*

McEwan I., *Une machine comme moi*, Gallimard, coll. « Du monde entier », 2020.

Maxwell J.C., *Matter and motion*, 1876.

Minh V. ; Kavukcuoglu K. ; Silver D. ; Rusu A., "Human-Level Control through Deep through Deep Reinforcement Learning", *Nature*, 2015, 518:7540.

Noguéro D., « Assurance et véhicules connectés. Regard de l'universitaire français », Dalloz IP/IT, n° 11, 2019, pp. 16-21.

Prévert J., *Fatras*, Gallimard, 1966.

Prévost S., « Procès de la voiture autonome : l'humain innocenté, l'IA condamnée », Dalloz IP/IT, 2018.

Radé C., « Causalité juridique et causalité scientifique : de la distinction à la dialectique », *Revue générale de droit médical*, n° 16, 2012, pp. 45-56.

Reigeluth T., « L'algorithme a ses comportements que le comportement ne connaît pas », *Multitudes*, n° 62, 2016, pp. 112-123.

Ribeiro M.T. ; Singh S. ; Guestrin C., "Why Should I Trust You?: Explaining the Predictions of Any Classifier", KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, août 2016. *ArXiv:1602.04938*

SAUVÉ J.-M., « Le numérique et les droits fondamentaux », Conseil d'État, intervention du 9 septembre 2014 lors de la présentation de l'étude annuelle, 2014. *https://bit.ly/2ra4xrn*

Society of Automotive Engineers (SAE), « Les véhicules automatisés au Canada », 2016. *https://bit.ly/2Q72Weo*

TOUATI A., « Il n'existe pas, à l'heure actuelle, de régime adapté pour gérer les dommages causés par des robots », *Revue Lamy droit civil*, n° 145, 2017.

WEBER M. (1905), *in* J. Reiss, "Counterfactuals Thought Experiments and Singular Causal Analysis in History", *Philosophy of Science*, vol. 76, n° 5, 2009, p. 712-723.