Loyauté des Décisions Algorithmiques

Philippe Besse, Céline Castets-Renard, Aurélien Garivier

Loyauté des Décisions Algorithmiques

Contribution au débat public initié par la CNIL : Éthique et Numérique.

Philippe Besse[1], Céline Castets-Renard[2] & Aurélien Garivier[3]

Résumé

L'éthique des algorithmes s'inscrit dans des débats plus larges : éthique et nouvelles technologies, éthique et technologie du numérique, éthique et déontologie. Les retracer rapidement permet de préciser la notion de loyauté appliquée à celle de décision algorithmique. Le cadre juridique de la « décision individuelle prise sur le fondement d'un traitement algorithmique » est alors précisé. Il apparaît que, dans ce contexte, les notions de précision, transparence ou explicabilité des décisions, biais des algorithmes, sont centrales pour aborder la « loyauté » de ces derniers et les questions éthiques qu'ils soulèvent. Après avoir explicitées puis illustrées ces questions par des exemples (justice prédictive, biologie et santé), leurs propriétés sont ensuite confrontées aux textes juridiques en cours et à venir pour montrer le peu d'adéquation de ces derniers à la grande complexité des algorithmes concernés. Sont finalement envisagées les actions possibles : évolution des textes juridiques, contrôle ou audit des algorithmes mais pourquoi et par qui ? L'acceptabilité de ces nouvelles technologies par les utilisateurs, citoyens et consommateurs, reste encore la principale motivation ou la rare contrainte des industriels du secteur pour moraliser leur pratique.

Introduction

« *Big data* », « *datafication* » du quotidien, « d*igitalisation* » de la vie[4], « *gouvernementalité algorithmique* »[5]… autant de mots clefs qui suscitent enthousiasmes ou craintes quant aux conséquences économiques, politiques et sociales de l'exploitation des flux massifs de données. Si la gouvernance par les nombres[6] n'est pas nouvelle et prit traditionnellement la forme des méthodes statistiques ou probabilistes[7], force est de constater que la question prend aujourd'hui une toute autre ampleur, eu égard aux possibilités techniques nouvelles et aux risques qu'elles sont susceptibles d'engendrer. La réaction

[1] Université de Toulouse INSA, Institut de Mathématiques UMR CNRS 5219.

[2] Université Toulouse Capitole, Institut de Recherche en Droit Européen, International et Comparé (IRDEIC). Membre de l'Institut Universitaire de France (IUF).

[3] Université de Toulouse Paul Sabatier, Institut de Mathématiques UMR CNRS 5219.

[4] A. Rouvroy, La « digitalisation de la vie même » : enjeux épistémologiques et politiques de la mémoire digitale, Documentaliste – Sciences de l'information, 2010, vol. 47 n° 1, p. 63.

[5] A. Rouvroy et Th. Berns, Gouvernementalité algorithmique et perspectives d'émancipation : le disparate comme condition d'individuation par la relation ?, *in* Politique des algorithmes : les métriques du web, Réseaux, La Découverte, févr.-avr. 2013.

[6] A. Supiot, La gouvernance par les nombres, Cours au Collège de France (2012-2014), Fayard, coll. Poids et mesures du monde, 2015.

[7] D. Pontille et D. Torny, La manufacture de l'évaluation scientifique : algorithmes, jeux de données et outils bibliométriques, *in* Politique des algorithmes : les métriques du web, Réseaux, La Découverte, févr.-avr. 2013, spéc. pp. 25-61

législative ne s'est alors pas faite attendre. Parallèlement à l'adoption en avril 2016 du règlement n° 2016/679/UE *relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données* qui entrera en vigueur en mai 2018, le législateur français a consacré des dispositions plus spécifiques sur ces questions dans la loi n° 1321 « *Pour République Numérique* » du 7 octobre 2016. Dans ce cadre, la CNIL est chargée de conduire une réflexion sur les *enjeux éthiques* soulevés par l'évolution des technologies numériques et a ouvert un débat public sur *Éthique et Numérique*. Ce document en est une contribution.

L'éthique des algorithmes s'inscrit dans des questions plus larges : éthique et nouvelles technologies, éthique et technologies du numérique, éthique et déontologie scientifique. Les retracer rapidement permet de dresser le décor et préciser la notion de *loyauté* appliquée à un algorithme.

Éthique et nouvelles technologies

Du plus général au particulier, et avant d'aborder précisément les questions qui interpellent l'enseignant chercheur, des aspects éthiques peuvent être débattus comme pour toute nouvelle technologie. Les technologies permettant la valorisation de grosses données ne sont pas idéologiquement neutres, pour la simple raison que tout le monde ne part pas d'un pied d'égalité. Ces technologies ne sont accessibles qu'aux entreprises ou institutions disposant : d'un accès aux grands volumes de données, de moyens techniques de les traiter et, enfin, des compétences pour le faire. Il est par ailleurs important d'identifier le *domaine d'application* ou l'objectif recherché car ces outils et technologies s'apparentent, comme beaucoup d'autres, à un *pharmakon*[8], à la fois remèdes et poisons, comme l'était l'écriture pour Platon : *remède* lorsqu'il s'agit d'aider à la détection de biomarqueurs ou cibles thérapeutiques pour une approche personnalisée du cancer ; *poison* lorsque ces algorithmes servent à identifier de possibles terroristes[9] ou cibler un suspect via un drone dans des zones de conflits à l'étranger, sur la base de l'utilisation d'un téléphone cellulaire. Par ailleurs l'émergence de nouvelles technologies soulève des risques de rejet, dès lors qu'elles suscitent craintes et appréhension du public ; nanotechnologies et OGM en sont des exemples.

Éthique et technologies numériques

La croissance des moyens de calculs numériques alliée à celle des masses de données accessibles nous impactent individuellement, comme citoyen ou consommateur. Cette évolution soulève des questions éthiques spécifiques. Il ne s'agit pas ici d'en aborder tous les aspects dont certains, largement débattus par ailleurs et en marge de notre champ de compétences, ne seront qu'évoqués. Tel est le cas notamment des questions de gouvernementalité algorithmique et la fin de la politique, ainsi que de la propriété, confidentialité, ouverture des données, risque de ré-identification, *datafication* du quotidien. Est ici posé le risque de fin du *consentement libre et éclairé*. Enfin, ne seront pas non plus abordées les problèmes d'entraves à la concurrence et biais discriminatoires des algorithmes de prix qui posent la question de leur contrôle[10]. Il n'en demeure pas moins nécessaire, voire

[8] Voir B. Stiegler qui l'emprunte à Derrida qui l'emprunte lui-même à Platon : B. Stiegler, « Questions de pharmacologie générale. Il n'y a pas de simple pharmakon », *Psychotropes*, vol. vol. 13, no. 3, 2007, pp. 27-54.

[9] Cf. l'article proposé par le site arstechnica.

[10] Rapport conjoint de l'Autorité de la concurrence et du Bundeskartellamt, « Droit de la concurrence et données », 10 mai 2016.

indispensable, de les mentionner en les référençant, tant tous ces aspects sont largement interconnectés avec les questions sur lesquels nous porterons notre attention.

Éthique et déontologie scientifique

Cette évolution technologique nous impacte également en tant que chercheurs en modifiant les modes de construction de la connaissance scientifique. D'un point de vue épistémologique, le déluge de données et leur capacité d'analyse interrogent sur la fin de la théorie (Anderson 2008).

Éthique et cadre juridique

L'accent est mis sur ce que les textes juridiques (loi pour une République Numérique (dite LRN), Règlement Européen sur la protection des données dit RGPD) nomment « décision individuelle prise sur le fondement d'un traitement algorithmique » raccourci ici en « décision algorithmique ». L'important est alors de préciser les frontières de ces textes de loi afin de pouvoir les confronter aux technologies concernées. Plus précisément celles concernées par les procédures d'apprentissage statistique ou machine à partir des données (*data driven*).

L'objectif est de mettre en évidence comment les caractéristiques techniques et propriétés de ces procédures sont compatibles ou non avec la loi, comment des zones d'ombre ou de flou laissent des espaces de non droit ou encore de *disruption technologique*. C'est typiquement dans ces espaces que doit être *interrogée l'éthique* des pratiques en cours et à venir et impactant les utilisateurs, consommateurs ou citoyens.

Des exemples issus de différents domaines (santé, justice, administration, commerce) illustrent ces questions dont les réponses influencent en retour l'adhésion ou non du public (patient, citoyen ou contribuable), à une politique de recherche scientifique, à l'ouverture des bases de données ou encore à la généralisation de l'automatisation de certaines décisions administratives, commerciales ou politiques.

1. Technologies numériques et éthique

Les questions posées ci-dessous sortent du champ de compétences des auteurs. Il est néanmoins important de les signaler dans une vision plus large.

1.1 Gouvernementalité algorithmique et mort de la politique

La notion de *gouvernementalité* introduite par Michel Foucault est étendue à l'usage administratif des algorithmes par Antoinette Rouvroy dans différents articles (2011), Rouvroy et Berns (2013), ainsi que dans une contribution[11] aux travaux de la Commission Européenne et une autre annexée au rapport annuel du Conseil d'Etat (2014). Par ailleurs, Morozov (2013) dénonce l'accumulation des données opérée par Google (*data washing*) au détriment des citoyens et de l'état providence, au point d'annoncer la *mort de la politique* (Morozov ; 2014, 2015), conséquence de l'intermédiation algorithmique des géants technologiques ou de l'application automatique des lois promise par cette même technologie.

Ces considérations ont largement influencé les textes juridiques sur la protection des personnes récemment publiés et analysés en section 3.

[11] A. Rouvroy (2016). Des données et des hommes. Droits et libertés fondamentaux dans un monde de données massives. *Conseil de l'Europe*, T-PD-BUR(2015)09REV.

1.2 Protection des Données

Certaines questions importantes relatives aux données personnelles ne seront pas abordées dans le cadre des débats menés par la CNIL car bien connues de cette dernière. Tel est le cas des problématiques liées à : la propriété, la vie privée, l'identité numérique, la protection et confidentialité des grandes bases de données, leur anonymisation pour ouverture à la recherche publique (*open data*) qui nécessite un contrôle du risque de ré-identification (confidentialité différentielle).

Noter, à titre d'exemple, les risques de remise en cause du principe de « *consentement libre et éclairé* », notamment pour des études cliniques, conséquence de l'enregistrement systématique, puis potentiellement l'analyse, de toutes les traces numériques diffusées, *volontairement ou non*, par les individus. Dans une conférence, Annick Guillevic-Alpérovitch (2016) cite l'exemple fort instructif du programme de care.data (annoncé en 2013) d'ouverture de toutes les données de la NHS au Royaume Uni ; projet abandonné en 2016 à la suite de nombreuses controverses. Elle fait référence au refus, par la société britannique, de transgresser implicitement (notion de *social licence)* certains principes dans l'espoir d'un gain collectif pour la santé publique. Elle insiste sur l'acceptabilité de tels projets conditionnée par leur :

- *trustworthiness*, véracité, crédibilité, pour mériter la confiance,
- *accountability*, responsabilité des décisions, capacité à en rendre compte.

Ces notions définissent les propriétés nécessaires à des technologies ou pratiques éthiques. Elles servent de base ou *référence* pour aborder l'étude plus technique des propriétés des méthodes et algorithmes utilisés, notamment ceux issus d'une procédure d'apprentissage (section 4).

Plus en retard et bénéficiant sans doute de cette expérience, le projet d'ouverture des données françaises de santé est en cours de réalisation après l'adoption de la loi n° 2016-41 du 26 janvier 2016 de modernisation du système de santé dite loi Touraine (art. 193) et la mise en place du Système National des Données de Santé (SNDS) (qui regroupe en particulier les données des établissements de santé, du SNIIRAM, du CépiDC, de la CNSA) et de l'Institut des Données de Santé (INDS) et un partenariat entre la CNAM et l'Ecole Polytechnique (*Data Science Initiative*) pour un accès exclusif à la base des prescriptions du Système National d'Information Inter-Régimes de l'Assurance Maladie (Sniiram). Ce fichier est unique au monde puisqu'il retrace exhaustivement le parcours de santé de 62 millions de personnes, bien plus que n'importe quel fichier d'une mutuelle américaine qui comporte au plus 8 millions de patients. Le décret n° 2016-1872 du 26 décembre 2016 précise les modalités de fonctionnement de l'INDS et du Comité d'Expertise pour les Recherches, les Etudes et les Evaluations dans le domaine de la Santé (CEREES). L'INDS sera en lien direct avec le CEREES, afin de fournir un avis à la CNIL sur la cohérence entre la finalité de l'étude proposée, la méthodologie présentée et le périmètre des données auxquelles il est demandé accès. Les possibilités d'analyse d'un tel volume de données et d'une telle complexité, notamment dans ses variétés de sources et formats, son flux permanent (vélocité), sont telles qu'elles soulèvent des *questions épistémologiques* (section 2) .

Des extractions de ces fichiers sont par ailleurs librement accessibles[12] en ligne mais dans des versions très anonymes pour rendre impossible l'identification des patients ni celle des personnels de santé. Seule est renseignée l'âge et la région administrative du patient

[12] Consulter le site https://www.data.gouv.fr/.

permettant, par exemple, des agrégations temporelles ou géographiques mais en perdant la possibilité de suivre un parcours de santé spécifique.

1.3 Entrave à la concurrence

D'autres questions concernent principalement le commerce en ligne et les possibles entraves à la concurrence avec des offres de prix automatiques, éventuellement entachées d'ententes ou de discriminations envers les consommateurs. Le livre d'Ezrachi et Stucke (2016) cité dans un blog du monde propose un tour d'horizon des problèmes émergeants avec l'algorithmisation de l'économie. Ceux-ci concernent notamment les stratégies automatiques de fixation des prix (*algorithmic pricing*) sur les sites de vente en ligne. Le projet INRIA de plateforme collaborative TransAlgo semble avoir pour objectif de regrouper des outils automatiques de contrôle mais son lancement semble dans l'attente de moyens suffisants dépendants eux-mêmes de futures volontés politiques.

Des compétences en Economie et Informatique sont préalablement nécessaires pour automatiser un recueil de données d'usage de ces plateformes avant qu'un statisticien puisse proposer des procédures de détection (test ?) automatique d'offres commerciales délictueuses : ententes illégales, comparaisons déloyales, prix discriminatoires.

2 Déontologie scientifique et épistémologie

Dès sa généralisation à partir de la fin du XIXème siècle, l'usage de la Statistique, comme aide à des *décisions politique*s ou comme élément de *preuve scientifique*, a suscité des critiques et remises en cause plus ou moins véhémentes. Il faut ajouter que les mauvais usages, intentionnels ou non, des outils de cette discipline, sans parler des fraudes manifestes par falsification des données, en politique comme en sciences expérimentales, ont largement contribué à ternir l'image et même le bienfondé de cette discipline. Wikistat (2017) propose un aperçu de ces questions de déontologie scientifique et donc d'éthique, avec des exemples empruntés à une actualité pas tout à fait récente mais indéfiniment renouvelée : scandales sanitaires liés à l'usage d'un médicament, innocuité ou toxicité de nouvelles technologies, causes du réchauffement climatique…

Le déluge des données et la nécessité d'automatiser leur traitement ne conduisent finalement qu'à considérablement amplifier, généraliser et médiatiser ces questions qu'il importe de replacer dans une perspective historique afin de mieux les appréhender.

2.1 Perspective historique

Le débat critique sur l'usage de la Statistique en lien avec l'évolution technologique se focalise ou peut être illustré par la remise en cause récurrente de l'utilisation d'un test statistique pour baser la démarche scientifique hypothético-déductive traditionnelle au sens de Popper. Cette pratique est bien établie depuis les années vingt avec les travaux fondateurs de Ronald Fisher : une *question* (*e.g.* ce médicament a-t-il un effet clinique ?), une *hypothèse* (*H0)*, dont l'*acceptation* ou la *réfutation* apporte la réponse à la question, une *expérimentation planifié*e pour recueillir au mieux les données sur un échantillon, un *modèle* supposé « vrai », l'*estimation* de ses paramètres à partir des données recueillies et sous l'hypothèse que *H0* est bien vérifiée, une *statistique de test*, une *p-valeur*, ou probabilité de dépasser cette statistique, qui conduit à *rejeter* ou non l'hypothèse *H0* et donc à répondre à la question initiale en contrôlant le *risque* (par défaut 5%), dit de première espèce, de *rejeter* ou réfuter *à tort l'hypothèse H0*. Tandis que le risque de deuxième espèce, ou puissance du test, d'*accepter à tort H0*, dépend de tout le contexte et ne peut être contrôlé que par la taille de l'échantillon.

Déjà, Tukey (1977) aux Etats Unis et le mouvement en France de l'*Analyse des Données* (Cailliez et Pages, 1976) remettent en cause l'usage systématique du modèle linéaire gaussien pour favoriser une approche descriptive et géométrique des données sans modèle probabiliste donc *sans inférence*.

Dans les années 90, des méthodes statistiques inférentielles classiques (modèles linéaires et régression logistique), descriptives (analyses factorielles) et d'apprentissage (classification supervisée ou non) sont associées à des outils de gestion de données (langage SQL de requêtage) dans des suites logicielles (SAS, Weka, Knime…) vendues pour fouiller ou prospecter les données (*data mining*). Ces données, bases comptables des banques, assurances, ventes par correspondance… ne sont pas acquises moyennant des expériences planifiées mais sont, *premier changement de paradigme*, préalables à l'analyse. Plus volumineuses (Méga octets), le discours commercial propose d'y chercher des diamants ou *pépites d'informations* sans se salir les mains. C'est la promotion d'une approche exhaustive, sans échantillonnage, afin d'éviter qu'une pépite ou *signal faible* ne passe au travers du crible d'un sondage. C'est aussi le risque déjà identifié du *data snooping* ou mise en exergue d'un simple *artefact* non reproductible sur d'autres données ; problème en relation avec les questions de *sur-apprentissage* des algorithmes ou *sur-ajustement* des modèles.

L'étape suivante ou *deuxième changement de paradigme* accompagne l'explosion des biotechnologies qui a suivi le premier séquençage d'un génome humain. Les données se stockent par Giga octets avec des tailles d'échantillons n toujours modestes (quelques dizaines) mais sur lesquelles un nombre considérable $p>>n$ de caractéristiques *omiques* sont observées (*single nucleotid polymorphisms*, expressions de gènes, protéines, métabolites…) pour chaque échantillon. Cette très grande ou hyper dimension provoque la remise en cause de la démarche classique : Benjamini et Hochberg (1995) proposent de contrôler le *taux de fausse découverte* (FDR) en complément de la *p-valeur* défaillante. Pour la même raison, les techniques de sélection de variables par pénalisation (Lasso) des modèles sont mis en compétition avec la panoplie des algorithmes issus de l'apprentissage machine ou statistique (*boosting*, SVM, *random forest*, réseaux de neurones…) ou leur sont adaptées. C'est dans ce contexte d'indétermination (fléau de la dimension) que sont publiées les critiques les plus véhémentes contre l'emploi bon ou mauvais de l'outil statistique, jugé responsable, sous la pression de publication (*publish or perish*), de la proportion considérable d'articles acceptés pour publication alors qu'ils exposent des résultats non reproductibles par la suite.

Un exemple permet d'illustrer les problèmes soulevés. De nombreuses équipes ont traqué et traquent toujours des gènes responsables de maladies multifactorielles complexes. Ainsi, de nombreux gènes ont successivement été identifiés comme possibles responsables de la dépression ; travaux démentis par une étude systématique (Hek et al. 2013) qui n'a pas réussi à retrouver le moindre gène potentiellement responsable d'une dépression. Voir également, parmi de nombreux autres, les articles de Ioannidis (2016) sur la reproductibilité des résultats et la remise en cause de la *p-valeur*, voire son *bannissement* (Trafimow et Marks, 2015) de journaux scientifiques.

Le *troisième changement de paradigme* intervient avec l'explosion de la volumétrie (Téra octets) des données, notamment dans le commerce en ligne et sur les réseaux sociaux, avec des tailles n d'échantillon multipliées par des facteurs 10^3, 10^6… Avec cette taille, tous les tests statistiques sont significatifs (Demidenko ; 2016) donc sans utilité. Le modèle explicatif ou confirmatoire est définitivement remplacé par un objectif de prévision d'une variable quantitative ou de la probabilité d'occurrence (score) d'une classe. Alors que le

statisticien devient un *data scientist*[13], l'analyse ou *Science* des grandes bases *de données* nécessitent encore plus de compétences, rigueur et déontologie dans l'usage de ces méthodes pour échapper, sous la pression de publication, aux sirènes du *data snooping*.

2.2 Science des Données

Dans un entretien publié par l'Obs, D. Patil (*LinkedIn*) explique comment il a « inventé » la *Science des Données* en 2008 avec J. Hammerbacher (*Facebook*) :

« Analyste, ça fait trop Wall Street ; statisticien, ça agace les économistes ; chercheur scientifique, ça fait trop académique. Pourquoi pas « data scientist »?

Besse et Laurent (2015) décrivent l'évolution d'une formation de statisticiens pour en faire des scientifiques de données massives et listent les compétences qu'il est nécessaire de transmettre *en plus* de celles déjà bien identifiées pour une pratique efficace et honnête de la Statistique.

Le chercheur, qu'il soit statisticien ou *data scientist*, utilisateur de la Statistique ou de la Science des données, est soumis aux mêmes règles déontologiques et sa pratique est interrogée par les mêmes questions éthiques (*accountabiliy*, *trustworthiness*). Il a une obligation de moyens et sa responsabilité est engagée quant à l'efficacité de son travail, sa pertinence, qualité, validité ou au moins l'honnêteté, la loyauté, des résultats qu'il publie, et encore sa capacité à *rendre compte* de leur utilité.

2.3 Epistémologie

Quel que soit le volume des données, les contraintes déontologiques restent les mêmes. En revanche, les aspects épistémologiques de l'analyse historique mettent en avant des changements de paradigmes et interroge sur le choix des moyens. Ce n'est pas parce que la stratégie classique d'acceptation / réfutation d'une hypothèse par un test statistique est devenue inutilisable face au déluge de données qu'il faut remettre en cause la démarche scientifique au sens de Popper, même si Anderson (2008) annonce la *fin de la théorie* et l'obsolescence de la démarche scientifique face aux données massives. Les hypothèses ne sont plus issues d'une théorie à valider ou invalider, mais de l'exploitation ou de la fouille de grandes masses de données à la recherche de corrélations (sans preuve de causalité), signaux faibles, cooccurrences ou motifs improbables qui prennent alors le statut d'hypothèse. Mais attention, ce n'est pas parce que ces hypothèses sont automatiquement générées par les données (*data driven*) que le chercheur est affranchi d'une vérification méthodique et scrupuleuse de leur validité.

Prenons l'exemple de la base Sniiram de l'assurance maladie. Très schématiquement, deux approches sont possibles :

- S'interroger sur les effets secondaires indésirables voire les risques causés par un médicament ou un *traitement spécifique* (*e.g.* la pilule de troisième génération). Il s'agit alors de définir précisément un protocole avec deux échantillons aléatoires (témoin et contrôle) de cette base afin de mettre en évidence une différence significative (test) entre les conséquences observées ou mieux en évaluant les capacités prédictives d'un modèle de risque sanitaire.
- Explorer (fouille) systématiquement d'un sous ensemble aléatoire de la base à la recherche de cooccurrences improbables entre des traitements susceptibles de

[13] Définition attribuée à Josh Willis (Cloudera) et reprise dans de nombreux exposés: "Data scientist (n): Person who is better at statistics than any software engineer and better at software than any statistician"

mettre en évidence des risques de certains d'entre eux. Pour éviter les pièges de ce *data snooping* et montrer que le signal faible détecté est réellement un signal et pas un artefact, il n'y a pas d'autres alternatives que de considérer ce résultat comme une hypothèse à valider sur un *autre sous-ensemble* de la base initiale avec la mise en place rigoureuse du protocole 1.

2.4 Ouverture des données et de la recherche

En Europe, la règlementation favorise voire impose l'ouverture des bases de données administratives et publiques dont les avancées sont relatées en France sur le site etalab.gouv.fr avec pour objectif plus de transparence dans les décisions et plus de facilités dans les analyses. La loi n° 1321 *Pour une république numérique* a ainsi consacré le principe de l'ouverture des données publiques par défaut. L'accès aux données offre des possibilités d'analyse à des institutions groupes ou individus pas nécessairement spécialistes ou professionnels des techniques ou méthodes concernées. Il est évident que cet amateurisme peut générer beaucoup de bruit autour des données. Considérer à titre d'exemple le nombre impressionnant de preuves amateurs et fausses du théorème de Fermat reçues par les collègues mathématiciens de l'Université Paul Sabatier avant la publication de la contribution définitive d'Andrew Wiles en 1994. L'ouverture des données peut aussi encourager ou faciliter la production malintentionnée de *fake (*pseudo*) sciences* au même titre que les *fake news* (faits « alternatifs » !) envahissent les réseaux sociaux. Certains auteurs comme le climato-sceptique Allègre (2010) n'ont d'ailleurs pas eu besoin de l'ouverture de données massives pour en produire.

Ce ne serait pas pour autant une raison pour chercher à limiter l'accès aux données et à contraindre leurs explorations tous azimuts. La comparaison avec la situation de l'Astronomie est justement très encourageante. Toute la communauté des astronomes amateurs et compétents qui observent le ciel sont amenés à des découvertes originales que ne peuvent faire les astronomes professionnels car ils sont bien moins nombreux et n'ont de toute façon pas le même objectif, la même focale d'observation. Ainsi, ces astronomes amateurs sont invités à rechercher et suivre de nouveaux astéroïdes (astrométrie). De façon analogue, des *data scientists* (ni scientistes, ni apprentis sorciers) amateurs, compétents et particulièrement sensibilisés aux risques du *data snoooping*, peuvent explorer l'univers des données à la recherche de signaux pertinents. Noter à ce sujet l'initiative de Paul Duan qui a créé en 2014 Bayes Impact avec le statut d'entreprise en Californie puis d'association 1901 en France. Noter également le succès des sites collaboratifs (type *Kaggle*) proposant des concours de prévisions. Ces réflexions rejoignent les conclusions d'un article[14] d'Henri Verdier (2016) sur son blog : « *Le seul choix raisonnable, il me semble, c'est d'entrer avec passion dans ce mouvement, de se réjouir que notre époque redécouvre le plaisir de chercher des réponses et de manipuler des données, et de contribuer de toutes nos forces à faire naître de nouvelles Lumières* ».

Pour résumer, la croissance du volume des données est donc à l'origine de changements de paradigmes méthodologiques mais ne provoque pas de rupture épistémologique. Kepler faisait déjà de la *Science des Données* pour trouver, puis prouver, l'ellipticité des trajectoires planétaires à partir de ses observations et *réfuter* ainsi l'hypothèse théorique posée *a priori* de leur circularité. Statistique et Science des Données sont confrontées aux mêmes contraintes déontologiques, posent les mêmes questions éthiques sur le rôle de la Science, sur ce qu'on veut lui faire dire ou tenter de prouver. Que la recherche

[14] Il était Directeur interministériel du numérique et du système d'information de l'Etat (*chief data officer*).

soit produite par des amateurs, des professionnels académiques ou du privé, le problème fondamental reste celui de la vérification des résultats et donc de l'accessibilité des données observées, des codes de calcul les analysant, afin de conférer et conserver le statut de *réfutabilité* aux hypothèses avancées. Ceci rejoint tous les mouvements et initiatives pour une *Science ouverte* et collaborative.

3 Textes juridiques

3.1. L'énoncé de droits liés aux décisions individuelles fondées sur un traitement automatisé de données

Textes

D'un point de vue juridique, la question de la loyauté des traitements et décisions algorithmiques a d'abord été appréhendée de manière indirecte et incomplète par la règlementation française et européenne sur la protection des données personnelles. L'article 10 de la loi n° 78-17 relative à l'informatique, aux fichiers et aux libertés du 6 janvier 1978 prévoit ainsi que « *Aucune décision produisant des effets juridiques à l'égard d'une personne ne peut être prise sur le seul fondement d'un traitement automatisé de données destiné à définir le profil de l'intéressé ou à évaluer certains aspects de sa personnalité* ». Autrement dit, une évaluation automatisée des caractéristiques d'une personne conduisant à une prise de décision ne peut être réalisée sur la seule base du traitement automatisé. Cela suppose donc que d'autres critères soient pris en compte ou encore que d'autres moyens soient utilisés. En particulier, les personnes concernées par la décision peuvent attendre que l'évaluation puisse être vérifiée par une intervention humaine. Si ce principe qui tend à contrôler les effets négatifs du profilage est consacré depuis longtemps, son énoncé n'a pu empêcher l'explosion de cette technique, parallèlement à l'émergence de la collecte massive des données sur internet. Autrement dit, beaucoup de techniques de profilage ont été développées, sans nécessairement prévoir des garde-fous techniques ou humains. Cette règle est donc peu respectée et sa violation ne donne pas lieu à sanction.

Dans le même ordre d'idée, l'article 22§1 du règlement 2016/679/EU (règlement général sur la protection des données personnelles dit RGDP) porte également sur les décisions individuelles automatisées, y compris le profilage. Il dispose que « *La personne concernée a le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative* ». La règle est sensiblement identique à celle consacrée en France, si ce n'est que le règlement consacre un véritable droit subjectif de la personne de ne pas faire l'objet d'une décision automatisée, dès lors qu'elle serait négative. On passe donc d'une interdiction à la reconnaissance d'un véritable droit.

Par ailleurs, le paragraphe 3 vient préciser « le droit de ne pas faire l'objet d'un traitement automatisé » prévu à l'alinéa 1er, en imposant au responsable de traitement de mettre en œuvre des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée. Concrètement, cette obligation doit se traduire « au moins » par le respect du « *droit de la personne concernée d'obtenir une intervention humaine de la part du responsable du traitement, d'exprimer son point de vue et de contester la décision* ».

D'autres droits sont donc ici consacrés qui viennent clairement expliciter ce que la loi française sous-entend, en particulier le droit à une intervention humaine. En outre, une forme de principe de contradictoire est instauré au travers du droit d'exprimer son point de vue et contester la décision. Mais encore faut-il que la personne concernée sache qu'elle a fait l'objet

d'un traitement automatisé. Les droits ainsi consacrés sont donc incomplets en l'absence de consécration d'un droit de savoir (droit à l'information), prérequis nécessaire à l'exercice des autres droits. Cependant, l'article 15h) du règlement consacre le droit d'être informé de l'existence d'une prise de décision automatisée. Ce point vient donc combler cette lacune. On pourra toutefois regretter que le manque de lisibilité et de lien clair entre ces dispositions qui auraient pu être regroupées dans le même article. Enfin, l'alinéa 4 prévoit que les décisions ne peuvent être fondées sur les catégories particulières de données à caractère personnel prévues à l'article 9, c'est-à-dire les données sensibles que sont les données biométriques, génétiques, de santé, ethniques, d'orientation politique, syndicale, sexuelle, religieuse, philosophique. Le traitement de telles données peut en effet conduire à des décisions discriminantes. Un principe de non-discrimination est donc ici sous-entendu.

En résumé

Un certain nombre de **droits** sont consacrés en cas de décision individuelle prise sur le fondement d'un traitement automatisé de données :

- Le droit d'être informé de l'existence d'une prise de décision automatisée (RGDP, art. 15h) ;
- Le droit de ne pas faire l'objet d'un traitement automatisé produisant des effets juridiques ou affectant la personne concernée de manière significative (RGDP, art. 22§1) ;
- Le droit d'obtenir une intervention humaine de la part du responsable du traitement (RGDP, art. 22§3) ;
- Le droit d'exprimer son point de vue et de contester la décision (RGDP, art. 22§3).

Cependant, des exceptions sont aussi prévues.

L'article 22§2 prévoit toutefois des exceptions lorsque la décision :

- est nécessaire à **la conclusion ou à l'exécution d'un contrat** entre la personne concernée et un responsable du traitement ;
- est **autorisée par le droit de l'Union ou le droit de l'Etat membre** auquel le responsable du traitement est soumis et qui prévoit également des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée ;
- est fondée sur le **consentement explicite** de la personne concernée ».

Cette série d'exceptions est loin d'être anodine et appauvrit substantiellement la règle. S'agissant des activités économiques du numérique, de nombreux traitements automatisés peuvent en effet se prévaloir d'un fondement contractuel, dès lors que l'utilisation par les internautes des services des sites de e-commerce ou plateformes de mise en relation, telles celles des réseaux sociaux, est de fait considérée comme une acceptation des conditions générales d'utilisation et manifestant l'acceptation de l'offre contractuelle.

En outre, le point c) du paragraphe 2 prévoit l'hypothèse d'un consentement explicite de la personne concernée. Si un consentement peut effectivement être assez aisément recueilli en sa forme, on peut toutefois douter au fond de son caractère éclairé, tant l'accessibilité intellectuelle aux procédés de traitement automatisé est douteuse à l'endroit des profanes composant la grande majorité des personnes concernées.

Au final

Si ces dispositions ont pour objectif de renforcer les droits des personnes concernées, des lacunes sont constatables, liées, d'une part, aux exceptions et liées, d'autre part, au fait

que l'énoncé de ces droits n'obligent en aucun cas à une transparence ou loyauté algorithmique.

En effet, force est de constater que ces dispositions ne consacrent pas directement des principes éthiques de transparence ou loyauté des algorithmes ou décisions automatisées. Aucune autre disposition du règlement ne va par ailleurs en ce sens. Dès lors, contrairement à ce que certains chercheurs ont pu affirmer[15], le règlement général ne consacre ni de droit d'explication ni de rendre compte ni plus globalement de principe de transparence ou loyauté algorithmique[16].

La seule référence à un « droit à explication » peut être trouvée au considérant 71 ce qui pourrait à l'avenir fonder une interprétation extensive de l'article 22 mais paraît bien maigre pour le moment. Egalement, l'article 15 h) précité prévoit que la personne concernée a le droit d'obtenir du responsable de traitement des informations sur l'existence d'une prise de décision automatisée, y compris un profilage, visée à l'article 22, paragraphes 1 et 4, mais aussi « au moins en pareils cas, des informations utiles concernant la logique sous-jacente, ainsi que l'importance et les conséquences prévues de ce traitement pour la personne concernée ».

On peut donc dire que le règlement général sur la protection des données ne concerne pas directement ni véritablement indirectement le principe de transparence algorithmique

3.2. Principe de transparence et loyauté algorithmique : vers une « *accountability* » ?

Le principe de loyauté et transparence algorithmique est en revanche consacré dans la loi n° 2016-1321 pour une République numérique (LRN) du 7 octobre 2016. La loi pose deux catégories de règles à l'égard des plateformes numériques, d'une part, et des administrations, d'autre part.

Le principe de loyauté des plateformes numériques

S'agissant des plateformes numériques[17], l'article 49 de la LRN, codifié à l'article L. 111-7. I. du Code de la consommation, donne une définition des plateformes : « Est qualifiée d'opérateur de plateforme en ligne toute personne physique ou morale proposant, à titre professionnel, de manière rémunérée ou non, un service de communication au public en ligne reposant sur :

- 1° Le classement ou le référencement, au moyen d'algorithmes informatiques, de contenus, de biens ou de services proposés ou mis en ligne par des tiers ;
- 2° Ou la mise en relation de plusieurs parties en vue de la vente d'un bien, de la fourniture d'un service ou de l'échange ou du partage d'un contenu, d'un bien ou d'un service.

---

[15] B. Goodman and S. Flaxman, EU Regulations on Algorithmic Decision-Making and A « right to Explanation » (2016) : https://arxiv.org/abs/1606.08813; B. Goodman: A Step Towards Accountable Algorithms?: Algorithmic Discrimination and the European Union General Data Protection, 29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain ; M. Hildebrandt, « The New Imbroglio – Living with Machine Algorithms », in The Art of Ethics in the Information Society (2016)

[16] En ce sens : S. Wachter, B. Mittelstadt, L. Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation?, *International Data Privacy Law*, à paraître.

[17] J. Rochfeld et C. Zolynski, La « loyauté » des « plateformes ». Quelles plateformes ? Quelle loyauté ?, Dalloz IP/IT 2016. 520.

En conséquence, le statut « d'opérateur de plateformes » regroupe une pluralité d'activités variées et très disparates. Il y a en réalité peu de points communs entre une activité de moteur de recherche comme Google et de mise en relations sociales comme Facebook. On peut douter de la pertinence d'une qualification globale, lors même qu'elle devrait déboucher sur un régime juridique unique ce qui paraît peu pertinent. Cela étant dit, le régime juridique tel qu'élaboré par la LRN, est, à l'heure actuelle, relativement pauvre. Le point II de l'article L. 111-7 dispose simplement que « *tout opérateur de plateforme en ligne est tenu de délivrer au consommateur une information loyale, claire et transparente sur :*

- 1° Les conditions générales d'utilisation du service d'intermédiation qu'il propose et sur les modalités de référencement, de classement et de déréférencement des contenus, des biens ou des services auxquels ce service permet d'accéder ;
- 2° L'existence d'une relation contractuelle, d'un lien capitalistique ou d'une rémunération à son profit, dès lors qu'ils influencent le classement ou le référencement des contenus, des biens ou des services proposés ou mis en ligne ;
- 3° La qualité de l'annonceur et les droits et obligations des parties en matière civile et fiscale, lorsque des consommateurs sont mis en relation avec des professionnels ou des non-professionnels ».

Autrement dit, est simplement prévue à la charge des plateformes une obligation d'information au profit des utilisateurs des services, non seulement sur les conditions générales d'utilisation (CGU), mais aussi sur les modalités de référencement. Il faut d'abord noter que la première information portant sur les CGU est inutile puisque le code de la consommation prévoit déjà l'obligation pesant sur le professionnel de préciser les conditions contractuelles au consommateur. Cette disposition est tout à fait justifiée, s'agissant de relations par nature déséquilibrées. La nouveauté vient de la deuxième obligation d'information portant sur les modalités de référencement. Mais cette formulation est suffisamment large et vague pour laisser toute liberté aux opérateurs de plateformes de ne pas ou peu préciser leurs règles de référencement. Ceci est d'autant plus vrai qu'ils pourront le plus souvent se prévaloir du secret des affaires les autorisant à ne rien dévoiler. Enfin, cette obligation de préciser les modalités de référencement se traduira vraisemblablement par l'obligation des professionnels de respecter leurs propres règles de classement et de référencement faute de quoi le consommateur pourra leur reprocher de se contredire au détriment d'autrui et les accuser de pratiques commerciales trompeuses.

Selon l'article L. 121-2. 2° b) du code de la consommation, est considérée comme trompeuse une pratique qui repose sur des allégations, indications ou présentations fausses ou de nature à induire en erreur et portant sur les caractéristiques essentielles du bien ou du service, à savoir : ses qualités substantielles, sa composition, ses accessoires, son origine, sa quantité, son mode et sa date de fabrication, les conditions de son utilisation et son aptitude à l'usage, ses propriétés et les résultats attendus de son utilisation, ainsi que les résultats et les principales caractéristiques des tests et contrôles effectués sur le bien ou le service. Si l'algorithme est trompeur ou si la méthode utilisée n'est pas celle annoncée, il sera donc aisé d'appliquer ces dispositions. Egalement, le paragraphe e) peut être utile si la tromperie concerne la portée des engagements de l'annonceur, la nature, le procédé ou le motif de la vente ou de la prestation de services ou bien encore le paragraphe g) qui concerne le traitement des réclamations et les droits du consommateur.

On peut en outre ajouter que l'article L. 121-3 du code de la consommation considère également une pratique commerciale comme étant trompeuse, « si, compte tenu des limites propres au moyen de communication utilisé et des circonstances qui l'entourent, elle omet, dissimule ou fournit de façon inintelligible, ambiguë ou à contretemps une information

substantielle ou lorsqu'elle n'indique pas sa véritable intention commerciale dès lors que celle-ci ne ressort pas déjà du contexte ». Ainsi, on pourrait également se servir de ces dispositions pour imposer que l'algorithme ne soit faussé et conduise à dissimuler une information essentielle pour le consommateur. Il ne doit pas non plus avoir une finalité commerciale sans qu'elle soit clairement indiquée.

Par ailleurs, l'article L. 121-4 17° du code de la consommation est également utile car il qualifie également de pratiques commerciales trompeuses, le fait « de communiquer des informations matériellement inexactes sur les conditions de marché ou sur les possibilités de trouver un produit ou un service, dans le but d'inciter le consommateur à acquérir celui-ci à des conditions moins favorables que les conditions normales de marché ». Ainsi, l'algorithme qui serait faussé et programmé pour augmenter artificiellement les prix pour inciter à l'achat immédiat pourrait donner lieu à sanction sur ce fondement. Ainsi, dans un arrêt rendu par la Cour de cassation le 4 décembre 2012, cette dernière a estimé que *« l'absence d'identification claire du référencement prioritaire est susceptible d'altérer de manière substantielle le comportement économique du consommateur qui est orienté d'abord vers les produits et offres des e-marchands "payants" et ne dispose pas ainsi de critères objectifs de choix »*. La cour d'appel a donc légalement justifié sa décision qui concluait à l'existence d'une pratique commerciale déloyale et trompeuse[18].

Au plan de la sanction, l'article L. 132-1 du code de la consommation précise que le délit de pratique commerciale trompeuse est constitué dès lors que la pratique est mise en œuvre ou qu'elle produit ses effets en France. Les pratiques commerciales déloyales sont passibles d'un emprisonnement de deux ans et d'une amende de 300 000 euros. Plus encore, le montant de l'amende peut être porté, de manière proportionnée aux avantages tirés du délit, à 10 % du chiffre d'affaires moyen annuel, calculé sur les trois derniers chiffres d'affaires annuels connus à la date des faits, ou à 50 % des dépenses engagées pour la réalisation de la publicité ou de la pratique constituant ce délit (C. consom., art. L. 132-1).

En résumé, la loi *pour une république numérique* a essentiellement pour objet d'imposer une obligation d'information (loyauté) sur les modalités de référencement des algorithmes, laquelle s'ajoute aux autres obligations d'information du code de la consommation. Surtout, cette obligation est utilement complétée par les dispositions préexistantes dans le code de consommation relatives aux pratiques commerciales trompeuses dont les énoncés sont suffisamment larges pour viser et sanctionner les comportements déviants qui pourraient être fondés sur des traitements algorithmiques déloyaux ou faussés.

Enfin, la loi pour une république numérique ajoute à l'article 50 que « Les opérateurs de plateformes en ligne dont l'activité dépasse un seuil de nombre de connexions défini par décret élaborent et diffusent aux consommateurs des bonnes pratiques visant à renforcer les obligations de clarté, de transparence et de loyauté ».

Le décret d'application n'a pas encore été pris.

La transparence algorithmique des décisions prises par les administrations

Par ailleurs, l'article 6 de la loi pour une république numérique prévoit que « Sous réserve des secrets protégés, les administrations ... publient en ligne les règles définissant les principaux traitements algorithmiques utilisés dans l'accomplissement de leurs missions lorsqu'ils fondent des décisions individuelles ». Un décret n° 2017-330 relatif aux droits des personnes faisant l'objet de décisions individuelles prises sur le fondement d'un traitement

[18] Cass. Com., 4 déc. 2012, Leguide.com c/ Pewterpassion.com, N° de pourvoi : 11-27729.

algorithmique a été pris le 14 mars 2017[19]. Il précise désormais à l'article R. 311-3-1-2 du code des relations entre le public et l'administration (CRPA) que :

« L'administration communique à la personne faisant l'objet d'une décision individuelle prise sur le fondement d'un traitement algorithmique, à la demande de celle-ci, sous une forme intelligible et sous réserve de ne pas porter atteinte à des secrets protégés par la loi, les informations suivantes :

- Le degré et le mode de contribution du traitement algorithmique à la prise de décision ;
- Les données traitées et leurs sources ;
- Les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé ;
- Les opérations effectuées par le traitement. »

Ce décret entrera en vigueur « le 1er jour du sixième mois suivant celui de sa publication », soit le 1er septembre 2017, afin de permettre aux administrations de s'organiser. Ce droit d'accès peut s'exercer auprès de toute administration, y compris des collectivités territoriales, « sous réserve de ne pas porter atteinte à des secrets protégés par la loi » mais aussi dans les limites des restrictions et secrets énumérés au 2° de l'article L. 311-5 du CRPA. Enfin, le silence gardé par l'administration au terme du délai d'un mois vaut décision de rejet (CRPA, art. R. 311-12 et R. 311-13).

La possibilité ainsi laissée de se prévaloir des secrets risque de vider de sa substance le principe de la diffusion de l'information. Dans son avis sur le projet de loi (avis du 3 déc. 2015, n° 390741), le Conseil d'Etat avait d'ailleurs mis en garde contre une trop grande précision des informations données dans ce cadre à même de « permettre à des usagers de se constituer un profil permettant de contourner les prescriptions qui seraient applicables aux opérateurs ».

Au-delà, l'encadrement technique du traitement algorithmique mérite une analyse à part entière.

4 Loyauté d'une Décision algorithmique (point de vue technique)

Comme le montre l'analyse fine des textes juridiques développée dans la section précédente, les obligations légales sont relativement peu contraignantes en matière de transparence des algorithmes. En l'absence de texte plus précis, les principes de loyauté des algorithmes deviennent des questions éthiques. Il importe de pouvoir définir comment cette notion peut se traduire en termes techniques.

4.1 Objectif

Il s'agit ici de focaliser le débat éthique sur les aspects les plus techniques des *décisions basées sur l'exécution d'algorithmes*, eux-mêmes issus de *modèles estimés* sur un corpus de données observées ; par opposition à des algorithmes procéduraux décrivant un enchaînement de propositions logiques comme celui d'admission post bac (APB). Une décision, comme par exemple le choix d'un traitement thérapeutique, d'une action commerciale en faveur d'un client pour éviter la rupture d'un contrat, d'une opération de maintenance préventive, du refus de crédit, de la mise sous surveillance d'un individu, de la recommandation d'un produit en ligne…) devient le résultat d'une *prévision* : probabilité de déclencher une maladie, de départ, de défaillance d'un équipement, de faillite, de

[19] J.-M. Pastor, Accès aux traitements algorithmiques utilisés par l'administration, AJDA 2017. 604.

radicalisation, d'intérêt d'un client… sous la forme d'un score calculé par un modèle estimé ou *algorithme appris* sur un *échantillon d'apprentissage* et évalué (taux d'erreur) sur un *échantillon* indépendant de *test*.

Se pose alors une question délicate : Comment définir ou par quelles caractéristiques « mesurables » traduire les notions de *trustworthiness, accountability,* appliquées à de telles décisions algorithmiques lorsqu'elles sont la conséquence ou le résultat d'une prévision ?

La réponse se décline en trois points. Même statistique ou probabiliste, cette décision doit pouvoir être *attribuée* à un humain qui en assume la responsabilité et

- la plus *juste* au sens de l'intérêt de la personne concernée et / ou globalement de la communauté ; donc issue d'une meilleure prévision.
- Il faut pouvoir en *rendre compte* et donc, pouvoir l'*expliquer* de façon compréhensible (*e.g.* médecin à son patient).
- Enfin, elle doit éviter tout biais *discriminatoire* vis-à-vis de minorités et groupes sensibles protégés par la loi.

4.2 Justesse et qualité de prévision

La *justesse* de la décision dépend de la qualité d'une prévision et donc de la qualité d'un modèle. Cette dernière dépend de la *représentativité* ou biais des données initiales, de l'adéquation du modèle au problème posé et à la quantité (variance) de bruit résiduel. Elle est évaluée sur un échantillon test indépendant ou par validation croisée (*Monte Carlo*) mais reste indicative sous la forme d'un risque *probabiliste d'erreur*. Alors que la loi n'en fait absolument pas mention, une décision algorithmique doit ou devrait être accompagnée par une évaluation de ce risque comme la loi oblige les instituts de sondage à produire des marges d'incertitude.

Les méthodes de prévisions sont estimées sur les données d'apprentissage, c'est donc la *qualité* de celle-ci qui est en premier lieu déterminante selon le vieil adage : *garbage in garbage out*. Leur volume peut être un facteur utile de qualité mais seulement si les données sont bien représentatives de l'objectif et pas biaisées. Dans le cas contraire des Téra octets n'y font rien. C'est l'exemple de Google Flu Trend (2008) qui visait à suivre en temps réel et prédire le déroulement d'une épidémie de grippe à partir du nombre de recherches de certains mots clefs et connaissant la localisation (adresse IP) du questionneur. L'outil a été abandonné par Google (2015) car source de lourdes erreurs de prévision. C'était le battage médiatique de la grippe qui était suivi, pas l'épidémie elle-même. Les données ont été reprises avec de meilleurs résultats par une équipe de Boston (Yanga et al ; 2015) en estimant un modèle autorégressif intégrant une chaîne de Markov cachée et corrigée sur la base des tendances des recherches sur Google.

Les principaux fournisseurs ou vendeurs d'Intelligence Artificielle (IA) : Google, Facebook, IBM, Microsoft… ont intérêt à mettre en évidence, amplifiés par les médias, les résultats les plus spectaculaires de l'IA : reconnaissance d'images, traduction automatique, compétition de jeu de go… avec des taux de succès exceptionnels, meilleurs que l'expert humain. Malheureusement ou heureusement (?) les taux d'erreurs attachés à la prévision de comportements humains *e.g* : score de récidive d'un détenu, détection de commentaires injurieux, de fausses nouvelles, d'un comportement à risque, sont nettement plus, voire tristement, pessimistes.

La loi ne codifie pas une obligation de résultat mais l'éthique des concepteurs y fait réfléchir dans le *partage des responsabilités*. L'obligation de publication ou de notification serait, comme pour les sondages un facteur important de responsabilisation de l'utilisateur.

4.3 Explicabilité vs. interprétabilité

Une décision algorithmique est dite *explicable* s'il est possible d'en rendre compte explicitement à partir de données et caractéristiques connues de la situation. Autrement dit, s'il est possible de mettre en relation les valeurs prises par certaines variables (les caractéristiques) et leurs conséquences sur la prévision, par exemple d'un score, et ainsi sur la décision.

Une décision algorithmique est dite *interprétable* s'il est possible d'identifier les caractéristiques ou variables qui participent le plus à la décision, voire même d'en quantifier l'importance.

Les algorithmes concernés par la construction de décisions sont basés sur des méthodes à l'interface entre disciplines Mathématiques (Statistique) et Informatique (Intelligence Artificielle) et sont d'une très grande diversité. Les méthodes d'apprentissage peuvent être structurées en deux groupes selon qu'elles conduisent, par construction, à un modèle explicite ou à une boîte noire.

- Dans le cas d'un modèle explicite[20] de type modèle gaussien, binomial, arbre binaire de décision (sauf si celui-ci est trop complexe : trop de variables et donc de paramètres), la décision qui en découle est explicable.
- La grande majorité des autres méthodes et algorithmes d'apprentissage : *k*-plus proches voisins, réseaux de neurones, machines à vecteurs supports, agrégation de modèles… sont des boîtes noires avec néanmoins la possibilité de construire des indicateurs d'*importance* des variables.

Dans le premier cas, le choix *a priori* d'un type de méthode ou de modèle conduit à une possibilité d'explication de la décision. Dans le deuxième cas, une fois déterminée la méthode conduisant à la meilleure prévision, des indicateurs d'importance sont calculés, *a posteriori*.

Par construction, une décision explicable est interprétable. Mais, dans le cas d'un *algorithme opaque*, il devient impossible de mettre simplement en relation des valeurs ou des caractéristiques avec le résultat de la décision, notamment en cas de modèle non linéaire ou avec interactions. Telle valeur élevée d'une variable peut conduire à une décision dans un sens ou dans un autre selon la valeur prise par une autre variable non identifiable, voire même une combinaison complexe d'autres variables.

Entre les deux stratégies, *explicabilité vs. interprétabilité*, tout est question d'objectif et de recherche d'un meilleur compromis entre niveau de compréhension et qualité de prévision. Si l'explication est privilégiée, la stratégie consiste à recherche le modèle explicable dégradant le moins possible la qualité de prévision ; voir par exemple les efforts de Zeng et al. (2016).

Si une qualité nécessaire de prévision n'est atteinte que par un algorithme opaque, il reste à quantifier l'importance des variables ou caractéristiques. C'était déjà une proposition de Breiman (2001) pour l'algorithme des forêts aléatoires, elle est reprise pour l'algorithme plus récent d'*extreme gradient boosting* (Chen et Guestrin, 2016). Data et al. (2016) proposent un autre critère applicable à toute méthode d'apprentissage, en vue du même objectif d'aide à l'interprétation. Quel est le niveau d'interprétabilité ou d'explicabilité demandé à une décision algorithmique en fonction du contexte ?

[20] Consulter le site wikistat.fr pour des présentations détaillées de chacune de ces méthodes.

A noter que le partenariat (Partnership on IA) entre les principaux acteurs (*Google, Facebook, Microsoft, IBM, Apple*…) pour une IA au service de l'homme est très sensible à ce besoin d'interprétation même si celle-ci est impossible en apprentissage profond (réseaux de neurones et *deep learning)*. Un article de leur charte précise :

7. We believe that it is important for the operation of AI systems to be understandable and interpretable by people, for purposes of explaining the technology.

4.4. Vers « un droit à explication » ?

Les dispositions de la loi pour une République numérique vont dans le sens d'une « explicabilité » même si un « droit subjectif à explication » n'est pas explicitement consacré. Le décret d'application pose ainsi des conditions à l'administration de devoir expliquer précisément ses décisions. Rappelons que doivent être communiquées « *à la personne faisant l'objet d'une décision individuelle prise sur le fondement d'un traitement algorithmique, à la demande de celle-ci, sous une forme intelligible et sous réserve de ne pas porter atteinte à des secrets protégés par la loi, les informations suivantes :*

1. *Le degré et le mode de contribution du traitement algorithmique à la prise de décision ;*
2. *Les données traitées et leurs sources ;*
3. *Les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé ;*
4. *Les opérations effectuées par le traitement. »*

Différentes situations peuvent être schématiquement considérées pour l'application de ce décret. Dans le cas d'un algorithme procédural de type APB, les règles de fonctionnement doivent être clairement explicitées ; le Ministère concerné s'y prépare pour septembre 2017. Dans le cas d'un algorithme d'apprentissage explicable, les coefficients d'un modèle linéaire ou logistique peuvent et doivent être explicités pour l'individu concerné de même que la séquence de règles définissant un arbre de décision. Le dernier cas d'un algorithme opaque ou seulement « interprétable » semble difficilement concerné ou pris en compte par le décret. Que peut faire un individu confronté à un ensemble de quelques centaines d'arbre de décision, d'un réseau de neurones définis par des milliers voire des millions de paramètres appris sur un gros volume de données ?

Dans beaucoup de domaines d'application et notamment en médecine, un modèle opaque qui ne permet pas de s'expliquer facilement, par exemple face un patient, et qui aboutirait à une *forme de déresponsabilisation* du décideur, ne serait que difficilement acceptable, à moins d'apporter une qualité de prévision nettement supérieure dans la recherche d'un meilleur *compromis entre qualité et explicabilité.*

Notons par ailleurs que le « droit à explication » peut faire l'objet de deux approches différentes[21] :

- le droit d'avoir une explication sur le fonctionnement général du système mettant en œuvre des décisions algorithmiques ;
- le droit d'avoir une explication sur une décision spécifique.

Au demeurant, l'explication peut être *ex ante* ou *ex post*[22]. S'il s'agit de donner une explication spécifique sur une décision individuelle, l'explication ne pourra être donnée que

[21] Article S. Watcher et alii., op. cit., p. 5.

[22] Ibid., p. 6.

ex post, alors que si elle porte sur le fonctionnement général, elle pourra l'être *ex ante* ou *ex post*.

Ainsi, le législateur pourrait se montrer plus précis en indiquant la temporalité et le mode d'explication attendu et en adaptant le décret pour concerner également les algorithmes rendus opaques par leur extrême complexité. Une approche graduée pourrait s'envisager, suivant la priorité donnée à l'explication ou la qualité de la prévision, à supposer qu'un algorithme plus opaque permette *a contrario* de meilleurs résultats. La réponse serait alors potentiellement différente selon les activités concernées car il n'est peut-être plus pertinent de traiter de la même façon un algorithme utilisé en médecine ou en techniques de commercialisation. Cela conduirait alors à encourager une règlementation sectorielle. En tout état de cause, il paraît indispensable de pouvoir faire un choix social sur ce qui est préférable dans une balance d'intérêts circonstanciée entre la qualité de l'explication et la qualité de la prévision, au moins dans les hypothèses où les caractéristiques des algorithmes sont réductibles à ces deux principales qualités.

4.4 Biais et discrimination

Définition et sanction de la discrimination

Rappelons que la discrimination est définie par le code pénal. Selon l'article 225-1 :

- « Constitue une discrimination toute distinction opérée entre les **personnes physiques** sur le fondement de leur origine, de leur sexe, de leur situation de famille, de leur grossesse, de leur apparence physique, de la particulière vulnérabilité résultant de leur situation économique, apparente ou connue de son auteur, de leur patronyme, de leur lieu de résidence, de leur état de santé, de leur perte d'autonomie, de leur handicap, de leurs caractéristiques génétiques, de leurs mœurs, de leur orientation sexuelle, de leur identité de genre, de leur âge, de leurs opinions politiques, de leurs activités syndicales, de leur capacité à s'exprimer dans une langue autre que le français, de leur appartenance ou de leur non-appartenance, vraie ou supposée, à une ethnie, une Nation, une prétendue race ou une religion déterminée. »
- « Constitue également une discrimination toute distinction opérée entre les **personnes morales** sur le fondement de l'origine, du sexe, de la situation de famille, de la grossesse, de l'apparence physique, de la particulière vulnérabilité résultant de la situation économique, apparente ou connue de son auteur, du patronyme, du lieu de résidence, de l'état de santé, de la perte d'autonomie, du handicap, des caractéristiques génétiques, des mœurs, de l'orientation sexuelle, de l'identité de genre, de l'âge, des opinions politiques, des activités syndicales, de la capacité à s'exprimer dans une langue autre que le français, de l'appartenance ou de la non-appartenance, vraie ou supposée, à une ethnie, une Nation, une prétendue race ou une religion déterminée des membres ou de certains membres de ces personnes morales. »

L'article 225-2 C. pén. ajoute que : *« La discrimination définie aux articles 225-1 à 225-1-2, commise à l'égard d'une personne physique ou morale, est punie de trois ans d'emprisonnement et de 45 000 euros d'amende lorsqu'elle consiste : 1° A refuser la fourniture d'un bien ou d'un service ; 2° A entraver l'exercice normal d'une activité économique quelconque ; 3° A refuser d'embaucher, à sanctionner ou à licencier une personne* ».

Définir une mesure de discrimination

Le règlement européen encadre strictement la collecte de données personnelles sensibles (orientation religieuse, politique, sexuelle, origine ethnique, …) et demande aux responsables de décision algorithmiques de s'assurer que celles-ci ne présentent pas de caractères discriminatoires vis à vis de ces caractéristiques (GDPR, art. 22§4). Par opposition à *discriminatoire*, une décision est dite *loyale* si elle ne se base pas sur l'appartenance d'une personne à une minorité protégée ou la connaissance *explicite* ou *implicite* d'une donnée personnelle sensible.

Ce point est sans doute le plus difficile à clarifier. En effet, il ne suffit pas que la variable « sensible » soit inconnue ou supprimée des données d'apprentissage pour que la décision soit sans biais vis-à-vis de ses modalités. L'information sensible peut être contenue implicitement, même sans intention de la rechercher, dans les informations non sensibles et ainsi participer au biais de la décision. Ainsi, des habitudes de consommation, des avis sur les réseaux sociaux, des données de géolocalisation… renseignent sur les orientations de la personne.

Les questions posées et difficultés rencontrées lors de la construction d'algorithmes avec un objectif de loyauté sont directement liées aux conditions d'apprentissage des décisions. En quoi l'échantillon est-il le reflet des biais de la société ? Et quoi ce biais est-il pris en compte, appris par l'algorithme ? Voire même renforcé lorsque, par exemple, une estimation (trop) élevée d'un risque de crédit génère un taux, donc des remboursements, plus élevés qui renforcent le risque de défaut de paiement. Cathy O'Neil (2016) développe en détail la perversité des effets de bord de ce type d'outils.

La première difficulté repose dans le choix d'une mesure de discrimination alors que la littérature propose beaucoup de façons de mesurer le *biais d'une décision* (positive ou négative) vis-à-vis de personnes appartenant ou non à un *groupe,* généralement une minorité, protégée par la loi. Un type de mesure *individuelle* s'intéresse au voisinage au sens des *k* plus proches voisins d'un individu afin de détecter une situation atypique. Néanmoins cet individu peut être entouré de ceux appartenant au même groupe protégé et tous ne bénéficiant pas à tort d'une décision positive. Il est plus informatif de considérer une mesure *collective* ou statistique de la discrimination basée sur une *table de contingence* (tableau 1) croisant deux variables : l'appartenance à un groupe protégé (Oui ou Non) par la loi et l'obtention d'une décision Positive (crédit, emploi, bourse…) ou Négative.

Tableau 1. Table de contingence entre appartenance au groupe et nature de la décision.
Proportions : *p1=a/n1, p2=c/n2, p=m1/n*

Groupe	**Décision**		
Protégé	**Positive**	**Négative**	**Marge**
Oui	*a*	*b*	*n1=a+b*
Non	*c*	*d*	*n2=c+d*
Marge	*m1*	*m2*	*n=n1+n2*

Des mesures simples de discrimination sont définies à partir de cette table (Pedreschi et al. 2012) :

- Différence de risque: *DR=p1-p2*
- Risque relatif: *RR=p1/p2*
- Chance relative : *CR=(1-p1)/(1-p2)*
- Rapport de cote (*odds ratio*): *RR/CR*

La loi du Royaume-Uni mentionne le *DR*, la Court de Justice européenne le *RR* tandis que les cours de justice des USA s'intéressent aux taux de sélection *(1-p1)* et *(1-p2).* D'autres mesures sont définies en comparant un groupe protégé par rapport à l'échantillon total :

- Différence de risque étendu: *(p1-p)*
- Rapport de risque étendu ou *extended lift* : *(p1/p)*
- Chance étendue: *(1-p1)/(1-p)*

Beaucoup d'autres mesures sont proposées. Consulter par exemple Žliobaité (2015) : différences de moyennes, de coefficients de régression, tests de rangs, information mutuelle, comparaison de prévisions.

N.B. Ces mesures ne prennent pas en compte la qualité de prévision qui est en fait au cœur de la controverse entre la société Northpointe (maintenant *equivant*) et le site ProPublica résumée ci-après.

Exemple : biais, score de récidive et discrimination

Les difficultés de prises en compte de biais potentiels ou implicites sont très bien illustrées par la controverse entre le site ProPublica (prix Pulitzer 2011) et l'ex-société Northpointe, maintenant equivant[23], qui commercialise l'application COMPAS (*Correctional Offender Management Profile for Alternative Sanction*) produisant un score de risque de récidive pour les détenus ou accusés lors d'un procès. ProPublica accuse ce score d'être biaisé et donc raciste. Cette controverse a suscité de très nombreux articles venant renforcer une bibliographie déjà présente sur le sujet depuis une dizaine d'années. Ces études proposent des pistes et mettent en évidence des contradictions rédhibitoires.

Le score est estimé sur la base d'un questionnaire détaillé et à partir d'un modèle de durée de vie (modèle de Cox). La qualité de ce score est optimisée, mesurée, par le *coefficient AUC* (aire sous la courbe ROC) approximativement autour de 0.7, valeur plutôt faible correspondant aux taux d'erreurs élevés observés. La société Northpointe défend l'*impartialité* de ce score en assurant que

- les *distributions* de ses valeurs (donc les taux de sélection) sont analogues selon l'origine (afro-américaine, caucasienne) des accusés,
- le *taux d'erreur* sur la prévision d'une récidive (matrice de confusion) qui en découle est analogue selon l'origine, autour de 40%.

Angwin et al. (2016) du site ProPublica dénoncent un biais du score COMPAS en considérant une cohorte de détenus libérés pour lesquels sont connus le score de récidive, ainsi que l'observation, ou non, d'une arrestation sur une période de deux ans. Ils montrent alors que le *taux de faux positifs* : score élevé mais sans récidive observée, est beaucoup plus importants pour les libérés d'origine afro-américaine que pour ceux d'origine caucasienne.

Pour expliquer l'impasse de cette controverse, Chouldechova (2016) montre que sous les contraintes de « loyauté » contrôlées par Northpointe et sachant que le taux de récidive des afro-américains est effectivement plus élevé alors, nécessairement, les taux de faux positifs / négatifs ne peuvent être que déséquilibrés au détriment des afro-américains et c'est d'autant plus manifeste que le taux d'erreur (40%) est élevé. En résumé, *ce n'est pas le modèle qui est*

[23] Les sociétés *CourtView Justice Solutions, Constellation Justice Systems et Northpointe* ont fusionné en janvier 2017 pour constituer la société *equivant*. L'URL précédente de NorthpointeInc pointe directement vers le nouveau site. Noter que toutes les informations et articles présentant la mise en place du score de récidive ont disparu du nouveau site. Le logiciel COMPAS existe toujours mais la présentation du score de récidive est noyée dans la proposition commerciale sans référence à la façon dont il est construit.

biaisé mais l'échantillon d'apprentissage, reflet des biais sociaux avec le risque, dénoncé par O'Neil (2016), de les renforcer.

Les applications sociales de police ou justice prédictive ne sont pas les seules susceptibles de conduire à des décisions discriminatoires car biaisées. Les primes d'assurance, offres d'emploi, taux de crédits… sont autant de champs d'applications de décisions algorithmiques pour lesquels les risques de biais sont important sauf dans les cas, comme l'assurance non vie, où la loi interdit explicitement de moduler le montant de la prime en fonction du client et des informations susceptibles d'être recueillies à son encontre. Ce serait en effet une remise en cause du principe d'asymétrie de l'information conduisant à une segmentation, selon le risque estimé, de la clientèle au détriment du principe de mutualisation et donc de solidarité.

Exemple : biais en Biologie et Santé[24]

Un autre domaine soumis à des biais expérimentaux concerne la construction des échantillons pour les expérimentations en Biologie et Santé. Ces biais, déjà bien connus avant l'invention de la Science des Données, sont d'autant plus gênants qu'il est largement question dans les médias de promouvoir une *médecine* de précision ou *personnalisée* par opposition à une médecine de population ; une médecine pour tous, appliquée avec les mêmes standards et objectifs *vs*. une médecine dépendant de la génétique d'un individu. L'objectif se schématise sous la forme de la question suivante. Connaissant les traitements déterminés par la médecine de populations, est-il possible de personnaliser ce traitement : médicament le plus adapté, dosage particulier, à un individu X, sur la base de connaissances génétiques ou génomiques le concernant.

Ces connaissances sont répertoriées dans des bases issues d'études d'association pangénomique (*genome-wide association study*, GWA *study*, ou GWAS) c'est-à-dire du génome complet. Il s'agit des analyses des variations génétiques (*singular nucleotid polymorphism* ou SNP) chez de nombreux individus, afin d'étudier leurs corrélations avec des traits phénotypiques, par exemple des maladies. L'étude d'association pangénomique « CoLaus » (cohorte lausannoise) cherche des associations entre les données médicales et génétiques d'une cohorte de plus de 6000 personnes.

La question éthique sous-jacente est alors : *sommes-nous égaux vis-à-vis de cette stratégie thérapeutique* ? La réponse est non principalement à cause de la façon dont ont été constitués les échantillons et donc des biais engendrés.

D'un point de vue ethnique, la grande majorité des bases GWAS ont été faites sur des populations d'ascendance blanche/européenne (cf. Figure 1, Popejoy et Fullerton, 2016). Les facteurs de risque estimés par des modèles statistiques classiques (régression logistique) ou par des algorithmes d'apprentissage machine seront donc très probablement différents pour un patient d'ascendance africaine ou asiatique.

Ces mêmes bases sont transversales et il y a comparativement très peu d'études longitudinales (Lee et al ; 2014). Une base de *n* patients adultes atteints d'obésité à un temps *t* prend en compte le facteur génétique mais masque la part environnementale impliquée dans le développement de cette maladie. Prenons le cas d'un enfant de dix ans dont le génome, une fois séquencé et comparé à une étude d'adultes obèses, présente un variant associé à l'obésité.

---

[24] Les auteurs remercient Aurèle Besse Patin de l'Institut de Recherche Clinique de Montréal, pour sa contribution à la rédaction de cette section.

La décision basée sur la génétique sera de le faire courir pour éviter une prise de poids; mais cela peut ne pas être la bonne solution.

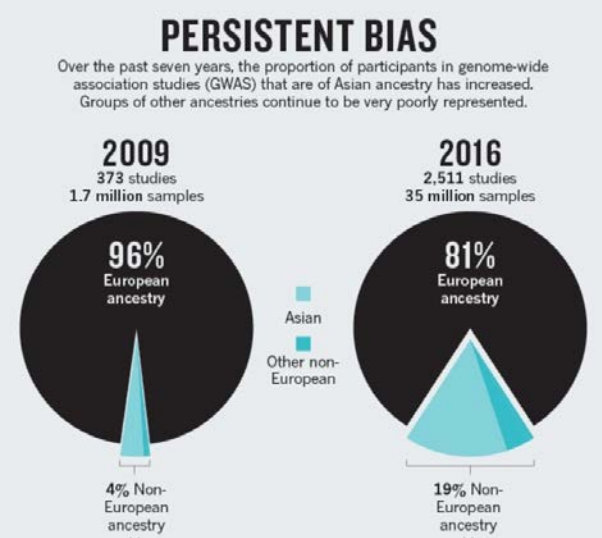

Figure 1 : Panpejoy et Fullerton (2016) : Biais de composition des échantillons des études d'associations pangénomiques (GWAS).

Ce variant peut n'être impliqué dans l'obésité qu'associé à la consommation d'un litre de soda sucré par jour, facteur non pris en compte dans l'étude de référence. Mais cet enfant ne boit pas de soda et donc ne sera pas à risque d'obésité. En revanche, l'obésité est fortement associée aux maladies cardiovasculaires, masquant tout effet de la génétique. L'enfant en question peut posséder ainsi un variant associé à une intolérance à l'exercice physique mais non pris en compte dans la prise de décision le concernant. Dans son cas, l'algorithme se basant sur des données biaisées, met à risque cet enfant d'avoir une crise cardiaque alors que la décision attendue aurait été d'éviter absolument les sodas.

Une troisième source de biais, sans doute la plus importante, est associée au genre. Génétiquement un homme blanc est plus proche d'un homme noir que d'une femme blanche. D'autre part même si les cohortes des GWAS incluent des femmes, la prise en compte de ce facteur n'est pas systématique. Plus curieusement et comme le font remarquer par exemple Chang et al. (2014), beaucoup d'études suggèrent une contribution du chromosome X au risque de maladies humaines complexes. Pulit et al. (2017) considèrent par exemple l'effet du genre et des facteurs sexués sur l'obésité. Malheureusement, ce chromosome est négligé ou insuffisamment analysé dans beaucoup de GWAS (Chang et al. 2014).

En résumé, les échantillons et donc les décisions qui en découlent sont largement biaisés ; une femme, jeune et d'ascendance africaine ou asiatique n'a actuellement pas grand-chose à attendre de la médecine personnalisée occidentale.

Ajouter à ces considérations de bais, l'opacité des modèles (pas d'interprétation), l'absence de marge d'erreur et l'absence d'intermédiation humaine pour apporter des explications et mises en garde ; nous pouvons sérieusement nous interroger sur la pertinence, et l'éthique, des analyses proposées par des sites comme 23andme.com. Elles sont interdites en France alors que la *Food and Drug Administration* (FDA) autorise[25] la vente sur internet de quantifications de risque de maladies multifactorielles sur la base d'une analyse génétique réalisée sur un échantillon de salive.

[25] Voir à ce sujet un article du New-York Times.

5 Contrôle des algorithmes

Lorsque des algorithmes conduisent à des décisions susceptibles d'impacter des personnes physiques il est nécessaire de pouvoir s'assurer a priori, par construction, ou a postériori par des contrôles que ces algorithmes respectent les clauses de loyauté des textes juridiques ou en respectent les valeurs éthiques lorsque ceux-ci sont trop imprécis. C'est un préalable indispensable à leur acceptabilité dans la sphère publique commerciale ou administrative.

La loyauté d'un algorithme peut être prise en compte *a priori* lors de sa conception ou contrôlé *a postériori*.

5.1 Conception loyale

L'apprentissage d'un algorithme conduisant à une décision sinon *loyale*, tout du moins pas trop déloyale, apparaît comme la recherche d'un meilleur *compromis* entre la *qualité de prévision* (justesse de décision), l'explicabilité et le *biais*. A cette fin, plusieurs stratégies sont proposées dont les deux principales sont résumées ci-dessous.

Débiaiser l'échantillon d'apprentissage

La première stratégie consiste à redresser l'échantillon d'apprentissage comme c'est pratiquer pour un sondage. Il s'agit de *transformer* les données pour assurer une forme d'indépendance entre les différents descripteurs des données et la variable d'appartenance à un groupe afin, finalement, que l'échantillon ne reflète pas les biais connus de la société ou plus généralement du domaine étudié. Le problème est plus ou moins complexe selon que la variable sensible d'appartenance au groupe « protégé » est observée ou non, soumise ou non à une clause de confidentialité.

Dans le cas le plus simple, lorsque l'appartenance au groupe est autorisée et connue, Kamiran et Calders (2011) comparent différentes stratégies poursuivant le même objectif: supprimer les *k* variables les plus liées à la variables groupe, changer les labels des observations proches de la frontière, repondérer les observations, supprimer ou sur-échantillonner certaines observations. Dans la même situation, Feldman et al. (2015) proposent une solution plus élaborées : transformer chaque variable de sorte que leurs distributions marginales, conditionnellement à l'appartenance au groupe, coïncident en préservant les rangs. Ce critère est basé sur la distance de Wasserstein (*earthmover distance*) entre les distributions.

Dans le cas le plus complexe, c'est-à-dire dans des situations où la connaissance du groupe n'est pas explicite, Hajian et al. (2013), Zliobaité et al. (2011), décrivent des procédures adaptées mais sans doute peu performantes car pas reprises ensuite dans la littérature.

Plus récemment, Ruggiery (2014), Hajian et al. (2014), combinent (meilleur compromis) des contraintes de confidentialité différentielle (contrôle du risque de ré-identification) lorsque la variable sensible est anonymisée, avec la recherche de décisions pas ou peu discriminatoires.

Apprentissage débiaisé

De nombreuses références proposent d'adapter ou contraindre des méthodes d'apprentissage à construire une règle de décision sans biais ou de biais réduit tout en préservant au mieux la qualité de prévision.

- Kamiran et al. (2010) proposent d'adapter la recherche des règles de division pour la définition des nœuds d'un arbre de décision.
- Calders et Verwer (2010) le font pour des classifieurs bayésiens naïfs.
- Tout récemment, Zafar et al. (2017) définissent des contraintes de loyauté qui peuvent être intégrées aux algorithmes d'estimation de méthodes classiques comme la régression logistique ou les SVM tout en contrôlant la perte de précision de la prévision.
- Zemel et al. (2013) associent un individu à une distribution dans un espace de représentation, qui « oublie » l'appartenance à un groupe protégé tout en conservant le « plus d'informations ». Il s'agit de la recherche d'une meilleure représentation masquant l'information sensible.
- Hajian et al. (2014) adoptent une autre stratégie en débiaisant une décision *après* le processus d'apprentissage.

La multitude des solutions proposées est aussi un reflet de la complexité du problème qui n'admet pas de solution élémentaire et consensuelle pour la communauté concernée ; la recherche est en cours mais pas aboutie. Du temps et des expérimentations seront nécessaires pour faire émerger les meilleures solutions en fonction du contexte et des données.

5.3 Contrôle et détection a postériori

L'objectif est de détecter des situations ou pratiques discriminatoires (un ensemble de décisions) au sein d'une base de données relatant un historique de décisions, sur la base d'un critère ou mesure de discrimination et d'un ensemble ou groupes de personnes potentiellement discriminées (Pedreschi et al. 2008).

Le problème est difficile, dépendant du choix de critère définissant la discrimination ou le biais, de la grande dimension et des contraintes de confidentialités qui ont pu conduire à anonymiser les données ou même supprimer la variable définissant le groupe protégé.

Plusieurs stratégies de détection basées sur des approches de type « fouille de données » sont proposées par Ruggieri et al. (2010). Elles conduisent à des détections collectives, vis-à-vis d'un groupe par règles d'association ou approche prédictive (impact décisif) en répondant à une question de rétro-ingénierie : est-il possible de prévoir l'appartenance à un groupe à partir des données et de la décision ? Des détections individuelles sont également proposées en utilisant les *k* plus poches voisins, un réseau bayésien…

5.4 Preuve du biais et de la discrimination

Il convient non seulement de contrôler les dispositifs en amont, pour vérifier la nature des données traitées, mais aussi et surtout en aval, en considération des résultats obtenus. Le contrôle est difficile dans cette deuxième hypothèse dans la mesure où les données utilisées ne sont par définition pas des données sensibles mais dont le traitement peut aboutir à des discriminations dont la preuve sera particulièrement difficile à rapporter pour la personne contrôler. Sans envisager d'inverser la charge de la preuve, cette dernière peut être facilitée par l'acceptation de procédés probatoires en principe interdits.

On pense en particulier aux dispositifs de type « *testing* », *test de situation* ou *test de discrimination*, qui est un moyen d'investigation et une forme d'expérimentation sociale en situation réelle destiné à déceler une situation de discrimination. Cette dernière peut notamment porter sur des données sensibles comme l'origine ethnique, le handicap, le sexe, l'orientation sexuelle, la religion, l'adhésion syndicale. Ce type de test ne respecte pas le principe de la loyauté de la preuve mais est le moyen le plus efficace, voire souvent le seul, de

prouver la discrimination. Dans le cas le plus simple, on compare le comportement d'un tiers envers deux personnes ayant exactement le même profil pour toutes les caractéristiques pertinentes, à l'exception de celle que l'on soupçonne de donner lieu à discrimination. Naturellement, lorsque la discrimination ne repose pas sur une seule ou même plusieurs données sensibles, mais est le résultat de croisement de données permettant indirectement la discrimination, il faut pouvoir interroger les résultats.

Cette méthode, utilisée par les associations comme SOS racisme, est reconnue par les juridictions françaises, dans la mesure où, bien que considérée comme pratique déloyale, elle ne peut être écartée comme moyen de recherche de la preuve depuis un arrêt de la Cour de cassation rendu en juin 2002 dans l'affaire du Pym's de Tours. La solution a été par la suite consacrée à l'article 225-3-1 du code pénal selon lequel : « *Les délits prévus par la présente section sont constitués même s'ils sont commis à l'encontre d'une ou plusieurs personnes ayant sollicité l'un des biens, actes, services ou contrats mentionnés à l'article 225-2 dans le but de démontrer l'existence du comportement discriminatoire, dès lors que la preuve de ce comportement est établie* ».

La difficulté toutefois est qu'il s'agit d'apporter la preuve de *l'intention discriminatoire*. Or, s'agissant d'une discrimination par un traitement algorithmique, la discrimination n'est pas forcément le fruit d'une intentionnalité.

Conclusion

L'afflux massif de données en Science, comme dans notre quotidien, a pour conséquence directe un recours de plus en plus systématique aux algorithmes d'apprentissage statistique, guidés par les données, pour l'aide à la décision voire même des prises de décision automatisées ou algorithmiques.

La recherche scientifique est largement impactée et se doit d'intégrer ces nouveaux outils dans sa démarche pour surfer au mieux les vagues de données. Cette adaptation méthodologique n'est pour autant pas une rupture épistémologique quant à la démarche et la rigueur nécessaire. Les confrontations des analyses et de leurs résultats, indispensables au débat scientifique sur la réfutation des hypothèses, nécessitent une ouverture pour une accessibilité systématique aux données, comme aux codes de calcul. C'est une condition indispensable de vérifiabilité donc de loyauté de la démarche, un remède préalable contre le manque de reproductibilité des résultats publiés.

En aval de la recherche ces stratégies de prises de décision se répandent dans tous les secteurs d'activité publiques, commerciaux, administratifs, économiques, industriels… alors que l'analyse fine des textes juridiques (section 3) montrent que même récents ceux-ci sont peu ou pas adaptés à la complexité des décisions algorithmiques.

- L'obligation de transparence ou d'explicabilité n'est effective que pour les décisions administratives explicables par construction : procédurales comme APB ou issues d'un modèle statistique (linéaire, logistique, arbre) élémentaire pour répondre aux attentes du décret relatif à l'article R. 311-3-1-2 du code des relations entre le public et l'administration (CRPA). A moins qu'une jurisprudence très restrictive n'en interdise l'usage, les algorithmes d'apprentissage machine récents et complexes (SVM, *boosting, random forest, deep learning.*…) ne peuvent être concernés.
- Aucun texte n'oblige à publier ou renseigner la qualité de prévision ou le taux d'erreur associé à l'utilisation d'un algorithme d'apprentissage.
- Une pratique discriminatoire est punie par la loi mais il revient à la victime d'en apporter la preuve.

La disruption technologique qui en découle, l'espace de non droit ou *far west* juridique, autorise toutes les possibilités de comportements ou de pratiques, éthiques ou pas. Les questions de discrimination sont celles le mieux encadrées par la loi mais aussi celles les plus complexes à appréhender. Les deux exemples présentés : justice prédictive et médecine personnalisée, montrent bien que les décisions qui en découlent ne peuvent être que largement statistiquement biaisés donc collectivement discriminatoires, sur certains critères, mais sans pour autant qu'il soit facile, pour une personne, de montrer qu'elle en a été lésée. Ces exemples montrent par ailleurs très bien que les données, bases de l'apprentissage des algorithmes, et leurs modes sélectifs de recueil, reflets de nos sociétés, sont la principale source de biais.

Cette situation motive en retour la recherche fondamentale pour définir des modèles ou construire des algorithmes répondant à ces critiques. Les investigations en cours consistent à rechercher des meilleurs compromis entre différentes contraintes : explicabilité et qualité de prévision, réduction du biais et confidentialité des données.

Quels contrôles ?

Vérifier l'interprétabilité ou l'explicabilité d'un algorithme ou du modèle sous-jacent, contrôler, par exemple sur un échantillon test, ses qualités prédictives et enfin détecter des biais potentiels, collectifs ou individuels sont des tâches complexes. À l'heure actuelle, aucun acteur ne peut à lui seul prétendre pouvoir contrôler la loyauté algorithmique. Une pluralité de contre-pouvoirs est donc nécessaire. Quels sont les acteurs susceptibles de prendre en charge ces contrôles ? Certains sont les régulateurs publics : CNIL, DGCCRF (répression des fraudes), Autorité de la Concurrence, juges (juridictions françaises et CJUE) mais en ont-ils les moyens. D'autres sont privés : plateformes collaboratives (Data transparency lab, TransAlgo INRIA, Conseil National du Numérique), Médias (ProPublica aux USA), ONG Data (Bayes Impact) mais ne sont que balbutiants.

Quelles normes ?

Faut-il aller plus loin que les principes énoncés par la loi pour une république numérique ? À l'heure actuelle, il convient d'abord de laisser le temps à cette loi de s'appliquer pour en mesurer la portée. Cela peut paraître prématuré, lors même que l'efficacité des dispositifs de contrôle est encore incertaine. En outre, comment formuler plus précisément les conditions d'encadrement dans l'utilisation des différentes méthodes algorithmiques ?

Dans ce contexte encore flou, il paraît peu pertinent de s'en remettre à une nouvelle fois et dès à présent au législateur. D'autres normes vont probablement apparaître, simples règles éthiques, bonnes pratiques (*soft law*), qui pourraient aider à mieux cerner les conditions d'une loyauté et transparence algorithmique. Les recherches académiques émergent seulement depuis 2-3 ans et il convient de prendre un peu de recul avant d'imposer une règle précise à respecter. Ces tâches de contrôle n'apportent pas de valeur ajoutée et ne sont donc pas auto-finançables ; elles doivent être imposées par des contraintes juridiques ou des clauses de loyauté vis-à-vis des utilisateurs, consommateurs, citoyens.

Dans le flou juridique et une disruption technologique toujours active, il appartient prioritairement aux fournisseurs et vendeurs de ces technologies de montrer que celles-ci peuvent surpasser l'expertise humaine et surtout qu'elles sont suffisamment loyales pour encourager leur acceptation dans les sphères publiques et commerciales et donc éviter un rejet massif, voire violent, de la part des citoyens et consommateurs. Le partenariat pour une Intelligence artificielle au service des gens et de la société, preuve d'un fond idéologique

transhumaniste, témoigne d'une intention ou d'une nécessité de stratégie commerciale en ce sens.

Références

Allègre C. (2010). L'imposture climatique ou la fausse écologie, Plon.

Angwin J., Larson J., Mattu S., Kirchner L. (2016). How we analyzed the compas recidivism algorithm. ProPublica, en ligne consulté le 28/04/2017.

Anserson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, *Wired*.

Benjamini Y., Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society*, Series B. 57 (1), 289–300.

Besse P., Laurent B.(2015). De Statisticien à Data Scientist; développements pédagogiques à l'INSA de Toulouse, *Statistique et Enseignement*, 7(1), 75-93.

Breiman L. (2001). Random forest, *Machine Learning*, 4, 5-32.

Cadiet L., L'accès à la justice, D. 2017. 522.

Cailliez F., Pages J.P., (1976). *Introduction à l'Analyse des Données*, Smash.

Calders T., VerwerThree S. (2010). Naive Bayes Approaches for Discrimination-Free Classication in Data Mining and Knowledge Discovery 21(2), 277–292.

Chang D., Gao F., Slavney A.,Ma L., Waldman Y., Sams A., Billing-Ross P., Madar A., Spritz R., KeinanA. (2014). Accounting for eXentricities: Analysis of the X Chromosome in GWAS Reveals X-Linked Genes Implicated in Autoimmune Diseases, *PLoS One*, 9(12).

Chen T., Guestrin C. (2016). XGBoost: A Scalable Tree Boosting System. In 22nd SIGKDD Conference on Knowledge Discovery and Data Mining.

Chouldechova A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, arXiv pre-print.

Cluzel-Métayer Lucie, La loi pour une République numérique : l'écosystème de la donnée saisi par le droit, AJDA 2017. 340.

Conseil d'Etat (2014). *Le numérique et les droits fondamentaux*, étude annuelle du Conseil d'Etat, La Documentation Française, en ligne, consulté le 29/042017.

Datta A., Sen S., Zick Y. (2016). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems, in IEEE Symposium on Security and Privacy.

Demidenko, E. (2016). The p-Value You Can't Buy, *The American Statistician*, 70(1), 33-38.

Dondero B., Justice predictive : la fin de l'aléa judiciaire ?, D. 2017. 532.

Ezrachi A., Stucke M. (2016). *Virtual Competition The promise and perils of algorithmic-driven economy*, Harward University Press.

Feldman M., Friedler S., Moeller J., Scheidegger C., Venkatasubramanian S. (2015). Certifying and removing disparate impact, arXiv-preprint.

Goodman B. (2016). A Step Towards Accountable Algorithms?:Algorithmic Discrimination and the European Union General Data Protection, in 29th Conference on Neural Information Processing Systems (NIPS 2016).

Goodman B., Flaxman S. (2016). EU regulations on algorithmic decision-making and a "right to explanation", ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York.

Guillevic-Alpérovitch A. (2016). Ethique et Big data, conference présentée à l'IPSED Bordeaux, en ligne, consultée le 28/04/2017.

Hajian S., Domingo-Ferrer J. (2013). A Methodology for Direct and Indirect Discrimination Prevention in Data Mining, IEEE Transactions on Knowledge and Data Engineering, 25(7), 1445 – 1459.
Hajian S., Domingo-Ferrer J., Farràs O. (2014). Generalization-based Privacy Preservation and Discrimination Prevention in Data Publishing and Mining, Data Mining and Knowledge Discovery 28 (5-6), 1158-1188.
Hajian S., Domingo-Ferrer J., Monreale A., Pedreschi D., Giannotti F. (2014). Discrimination – and privacy – aware patterns in Data Mining and Knowledge Discovery 29(6).
Hek K. et al. (2013). A Genome-Wide Association Study of Depressive Symptoms, *Biological Psychiatry* , 73(7) , 667-678.
Ioannidis J. (2016). *Why Most Clinical Research Is Not Useful*, PLOS Medecine, 13(6).
Kamarinou D., Millard C., Singh J. (2016). Machine Learning with Personal Data: Profiling, Decisions and the EU General Data Protection Regulation, in 29th Conference on Neural Information Processing Systems (NIPS 2016).
Kamiran F., Calders T. (2011). Data Pre-Processing Techniques for Classification without Discrimination, Knowledge and Information Systems 33(1).
Kamiran F., Calders T, Pechenizkiy M. (2010). Discrimination Aware Decision Tree Learning in ICDM, 869-874.
Lee Y., Park S., Moon S., Lee J., Elston R., Lee W., Won S. (2014). On the Analysis of a Repeated Measure Design in Genome-Wide Association Analysis, *Int. J. Environ. Res. Public Health*, 11, 12283-12303.
Mittelstadt B., Allo P., Taddeo M., Wachter S., Floridi L. (2016). The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, 3(2).
Morozov E. (2012). The Net Delusion: The Dark Side of Internet Freedom, PublicAffairs.
Morozov E. (2013). To Save Everything, Click Here : Technology, Solutionism, and the Urge to Fix Problems that Don't Exist, Allen Lane.
Morozov E. (2014). The rise of data and the death of politics, *The Observer*.
O'Neil C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality, Crown Random House.
Pastor J.-M., Accès aux traitements algorithmiques utilisés par l'administration, AJDA 2017. 604.
Pedreschi D., Ruggieri S., Turini F. (2008). Discrimination-Aware Data Mining. In KDD, pp. 560-568.
Pedreschi D., Ruggieri S., Turini F. (2012). A Study of Top-K Measures for Discrimination Discovery. SAC. Proceedings of the 27th Annual ACM Symposium on Applied Computing, 126-131.
Popejoy A., Fullerton S. (2016). Genomics is failing on diversity, *Nature*, 538, 161-164.
Pulit S., Karaderi T., Lindgren C. (2017). Sexual dimorphisms in genetic loci linked to body fat distribution, *Bioscience Report*, 37(1).
Rochfeld J. – Zolynski C., La « loyauté » des « plateformes ». Quelles plateformes ? Quelle loyauté ?, Dalloz IP/IT 2016. 520.
Rouvroy A. (2011). Pour une défense de l'éprouvante inopérationnalité du droit face à l'opérationnalité sans épreuve du comportementalisme numérique, *Dissensus*, 4.
Rouvroy A., Berns T. (2013). Gouvernementalité algorithmique et perspectives d'émancipation, *Cairn*, 177 (1), 163-196.
Ruggieri S. (2014). Using t-closeness anonymity to control for non-discrimination, Transaction on Data Privacy, 7, 99-129.
Ruggieri S., Pedreschi D., Turini F. (2010). Data mining for discrimination discovery. In TKDD 4(2).
Stahl J.-H., Données publiques - *open data* et jurisprudence, Dr. adm. 2016, n° 11.

Stiegler B. Questions de pharmacologie générale. Il n'y a pas de simple pharmakon, *Psychotropes*, vol. vol. 13, no. 3, 2007, pp. 27-54.
Trafimow, D., Marks, M. (2015), Editorial, *Basic and Social Psychology*, 37, 1–2.
Tukey J. (1977). *Exploratory Data Analysis*, Addison-Wesley.
Verdier H. (2016). Science et Big data: la fin de la théorie ?, en ligne, consulté le 28/04/2017.
S. Wachter, B. Mittelstadt, L. Floridi, Why a Right to Explanation of Aumated Decision-Making Does Not Exist in the General Data Protection Regulation?, *International Data Privacy Law*, à paraître.
Wikistat (2017). Statistique et Déontologie Scientifique, en ligne, consulté le 27/04/2017.
Yanga S., Santillanab M., Koua S. (2015). Accurate estimation of influenza epidemics using Google search data via ARGO, *PNAS*, 112(47), 4473–14478.
Wachter S., Mittelstadt B., Floridi L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, *International Data Privacy Law*, à paraître.
Zafar M., Valera I., Rodriguez M., Gummadi K. (2017). Fairness Constraints: Mechanisms for Fair Classication in International Conference on Articial Intelligence and Statistics (AISTATS), vol. 5.
Zemel R., Wu Y., Swersky K., Pitassi T., Dwork C. (2013). Learning Fair Representations in JMLR W&CP 28(3), 325–333.
Zeng J., Ustuny B., Rudin C. (2016). Interpretable Classification Models for Recidivism Prediction, , arXiv pre-print.
Žliobaitė I. (2015). A survey on measuring indirect discrimination in machine learning. arXiv pre-print.
Žliobaitė I., Kamiran F., Calders T. (2011). Handling Conditional Discrimination, Proceedings of IEEE International Conference on Data Mining, 992-1001.
Rapport conjoint de l'Autorité de la concurrence et du Bundeskartellamt, « Droit de la concurrence et données », 10 mai 2016.