Anticiper les Risques Juridiques des Systèmes d'IA

PHILIPPE BESSE

Université de Toulouse – INSA, IMT – UMR CNRS 5219, OBVIA – Université Laval
18 mars 2021

Résumé

Faisant suite au déploiement du RGPD, la Commission Européenne a publié, en février 2020 un livre blanc pour une approche de l'IA basée sur l'excellence et la confiance et dont les recommandations sont largement issues du guide pour une IA digne de confiance rédigé en 2019 par un groupe d'experts européens. Au delà des questions prioritaires de protection des données au cœur des missions de la CNIL, ce livre blanc soulève avec insistance d'autres questions relatives aux risques des impacts des algorithmes d'apprentissage automatique sur notre société : qualité, reproductibilité de décisions algorithmiques, opacité des algorithmes et explicabilité des décisions, biais et risques de discrimination. En nous basant sur l'exemple bien connu du score de crédit, nous décrivons quels outils, procédures, indicateurs (cf. tutoriel), pourraient participer à la construction d'un DIA ou Discrimination Impact Assessment souhaité par le rapport Villani (2018) et cohérent avec la liste d'évaluation du groupe des experts européens. L'exemple traité montre les difficultés voire l'impossibilité d'un audit ex post d'algorithmes d'apprentissage. Pour réponde au légitime besoin d'auditer les algorithmes, nous concluons sur l'incontournable mise en place d'une documentations précise et exhaustive ex ante d'un système d'IA et donc sur la nécessité de former à ces questions les futurs responsables de ces systèmes comme le préconise la Commission Européenne.

1 Introduction

1.1 Battage médiatique

L'intelligence artificielle (IA) dite *faible*, opposée à une IA *forte* supposée disposer d'une conscience de soi et que nous laisserons à la science fiction, recouvre une grande variété d'objets, méthodes et algorithmes susceptibles d'imiter des comportements humains "intelligents" : robots, véhicules autonomes, systèmes experts, algorithmes d'apprentissage automatique... Depuis 2012 nous sommes soumis à une déferlante médiatique sans précédent sur les applications des algorithmes d'IA associés à des succès retentissants : reconnaissance d'images et diagnostic automatique, véhicules autonomes, victoire au go, traduction automatique... Ce battage médiatique fait suite à celui sur l'avènement du stockage tous azimuts de données massives ou *big data* et leur utilisation pour alimenter les nouveaux algorithmes d'IA exécutés dans des environnements technologiques en constante progression. Cette convergence entre données massives, algorithmes performants et puissance de calcul est à l'origine de l'expansion exceptionnelle des usages de l'IA dans tous les domaines de nos quotidiens. Les principaux acteurs technologiques comme *Google, Facebook, Amazon* ou *Microsoft*, ont tout intérêt à sur-médiatiser ces succès puisque des revenus considérables proviennent de la vente de l'application de ces technologies à notre profilage publicitaire. Ils se doivent donc d'en promouvoir l'efficacité, même si ses succès diffèrent largement en fonction du domaine d'application et si elle peut s'avérer anxiogène dans ses conséquences sociétales, tant sur la destruction d'emplois même qualifiés, que sur la déresponsabilisa des acteurs humains ou encore l'exposition des données de la vie privée.

1.2 Confiance et acceptabilité

Une composante importante de la publicité excessive autour de l'IA concerne son acceptabilité, comme celle de toute nouvelle technologie pénétrant ou plutôt envahissant nos quotidiens. Le principal enjeu est de cultiver ou conquérir la confiance des utilisateurs, qu'ils soient consommateurs, clients, patients, contribuables, justiciables ou citoyens, pour une IA acceptable. En première ligne, les entreprises privées spécialistes des réseaux sociaux et technologies numériques, rejoints ensuite par plus de 90 partenaires, se sont empressées, dès 2015, de signer une charte de partenariat pour une IA au bénéfice du peuple

et de la société. Dès lors, tous les acteurs publics institutionnels ont rejoint le mouvement ; citons parmi les plus récents la partie 5 du rapport Villani pour donner un sens à l'IA (Villani et al. 2018), les lignes directrices pour une IA digne de confiance des hauts experts désignés par la Commission Européenne (High Level Expert Group 2019), ou encore la déclaration de Montréal pour le développement d'une IA responsable (2018). C'est plus largement une avalanche de recommandations pour une IA éthique au service de l'humanité dont Jobin et al. (2019) explore le paysage. Les enjeux sont considérables car, en l'absence de confiance, les utilisateurs n'accepteront pas l'IA. Sans acceptation sociale, les entreprises technologiques ne pourront plus collecter toutes les données nécessaires et ne pourront pas développer une IA pertinente, source de profits. Les conséquences de l'affaire *Cambridge Analytica* sur l'encours boursier de Facebook, en mars 2018, en furent une démonstration éclatante.

1.3 Éthique et protection juridique

Cette affaire qui peut être citée parmi d'autres : condamnations successives de Google pour entrave à la concurrence, fuites massives et répétées de données personnelles, utilisations abusives de celles-ci... nous rappelle que le but premier des entreprises commerciales ou de leurs dirigeants n'est pas l'altruisme ou la philanthropie mais le montant des encours boursier ou celui des dividendes distribués à leurs actionnaires. Ces profits nécessitent des pratiques éthiques pour être acceptables mais la confiance des usagers sera nettement plus franche et massive si elle repose sur une protection juridique, plutôt que sur de bonnes intentions éthiques (*ethical washing*), aussi louables soient-elles. En France, la première version de la loi Informatique et Liberté date de 1978. Ce texte précurseur marqua une réelle anticipation des problèmes à venir. En revanche, à l'heure actuelle, la loi peine à suivre les évolutions ou disruptions technologiques.

L'entrée en vigueur du RGPD (Commission Européenne 2018), puis son intégration dans les textes nationaux des États membres signe une avancée majeure pour la protection des données personnelles en Europe. Le principe de sécurité et confidentialité, au cœur de l'action de la Commission Nationale de l'Informatique et des Libertés (CNIL) en France, est en effet une priorité mais d'autres aspects, tant juridiques qu'éthiques, sont à considérer pour instaurer ou restaurer la confiance des usagers envers ces nouvelles technologies. Ainsi, l'article 22.1 du RGPD (Commission Européenne 2018) accorde aux personnes concernées le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, produisant des effets juridiques la concernant ou l'affectant de manière significative. Repris dans les lois nationales des États membres, cet article a pu servir de fondement en droit français pour reconnaître un droit à l'explicabilité des décisions algorithmiques, dans le souci de lutter contre les risques de discrimination. Ces préoccupations rejoignent les exigences publiques exprimées dans un sondage réalisé au Royaume-Uni (Vayena et al. 2018) au sujet des applications de l'IA en médecine.

Un large consensus est donc établi sur la nécessité de pratiques en IA respectueuses de l'éthique. Néanmoins, compte tenu des pressions financières, un cadre juridique s'avère indispensable. Il est un préalable à des pratiques vertueuses génératrices de confiance. Tel est bien l'objectif de la Commission Européenne (CE) qui propose les éléments clefs d'un futur cadre réglementaire dans le livre blanc (Commission Européenne 2020) pour une *IA basée sur l'excellence et la confiance fondée sur les droits fondamentaux de la dignité humaine et la protection de la vie privée*. La rédaction de ce livre blanc s'appuie sur les lignes directrices pour une IA de confiance (High Level Expert Group 2019) rédigées par un groupe d'experts et dont il est important d'*anticiper l'impact à venir* notamment en l'intégrant dans nos formations. En résumé, les technologies de l'IA se développent à grande vitesse dans un contexte juridique très complexe mais insuffisant à encadrer les risques sociaux susceptibles de se produire. Ce cadre légal est appelé à évoluer, au moins en Europe, afin de minimiser les risques et créer les conditions d'une acceptabilité sociale de l'IA.

La section deux, qui peut être sautée par un lecteur déjà expert, précise de quelle IA il est question, la troisième tente de résumer le cadre juridique actuel et celui réglementaire européen à venir. Puis, chaque section aborde un risque particulier : qualité des décisions algorithmiques, explicabilité ou opacité, avec un focus particulier sur ceux de biais et discriminations avant de conclure sur la nécessité d'une approche documentaire exhaustive *ex ante* d'un algorithme en vue d'un audit. Cette approche, future réglementation de la Commission Européenne (2020) annoncée dans le livre blanc, doit être anticipée par les responsables des systèmes d'IA et donc ses outils inclus dans leur formation comme demandée par la CE.

2 De quelle IA est-il question ?

2.1 IA du quotidien et apprentissage statistique

Nous n'aborderons pas les questions d'IA forte ou celles de science fiction : transhumanisme, singularité technologie, lois d'Asimov. Nous n'évoquerons pas non plus les questions sociologiques anxiogènes : destruction d'emplois qualifiés, surveillance généralisée de la population. Ce chapitre s'intéresse aux usages quotidiens des systèmes d'IA. Le choix d'un traitement médical, d'une action commerciale, d'une action de maintenance préventive, d'accorder ou non un crédit, de surveiller plus particulièrement un individu, de bloquer un paiement... Toutes les décisions qui en découlent sont la conséquence d'une prévision. La prévision du risque ou de la probabilité de diagnostic d'une maladie, le risque de rupture d'un contrat par un client qui est le score d'attrition (*churn*), le risque de défaillance d'un système mécanique, de défaut de paiement d'un client, de radicalisation ou passage à l'acte d'un individu, de fraude... Les exemples sont très nombreux et envahissent notre quotidien. Ces prévisions de risques ou scores, par exemple de crédit, sont produits par des algorithmes d'*apprentissage statistique*, après entraînement sur une base de données.

De façon générale, un modèle est estimé ou un algorithme entraîné pour rendre visibles des relations entre une variable Y cible (le risque, le diagnostic...) et un ensemble de variables ou caractéristiques (*features*) dites explicatives $X_{j=1,\ldots,p}$: socio-économiques biologiques... Toutes ces variables (Y, X_j) sont mesurées, observées, sur un ensemble $i = 1, \ldots, n$ d'individus ou instances appelé échantillon d'apprentissage ou d'entraînement. Une fois un modèle estimé ou un algorithme entraîné sur ces données, la connaissance d'un vecteur x_0, contenant les observations des variables X_j pour un nouvel individu, permet d'en déduire une prévision de la valeur ou de la classe y_0 le concernant. Le modèle ou l'algorithme calcule automatiquement cette valeur y_0 en combinant, en fonction de l'algorithme utilisé, celles y_i observées sur les individus présents dans la base d'apprentissage et proches de x_0, en un certain sens, au regard des valeurs x_{ij}. Autrement dit, la prévision d'une nouvelle situation et donc la décision qui en découle, est construite automatiquement à partir des situations lui ressemblant le plus dans la base d'apprentissage et dont les décisions sont déjà connues. Le principe repose sur la stationnarité des données : la loi apprise sur l'échantillon d'apprentissage est la même que celle des données que l'on veut tester. En conséquence, l'apprentissage statistique n'invente rien, il reproduit un modèle connu et le généralise aux nouvelles données, au mieux selon un critère spécifique d'ordre statistique à optimiser. Plus on possède de données, meilleure sera la connaissance fournie par ce modèle. Ceci souligne le rôle fondamental joué par les données et donc le succès des grands acteurs d'internet et des réseaux sociaux qui bénéficient d'une situation de monopole sur des masses considérables de données comportementales des internautes pour les traduire en profilage et donc en recettes publicitaires. Transposé à d'autres domaines dont celui de la santé, où l'objectif est une prise en compte toujours plus fine de la complexité du vivant, le premier enjeu est l'accès à de grandes masses de données personnelles excessivement sensibles, objet de toutes les convoitises.

2.2 Statistique inférentielle *vs.* apprentissage statistique

Deux objectifs doivent être clairement distingués dans les applications, tant de la statistique que de l'IA pour lever une ambiguïté trop répandue. Le premier objectif est celui explicatif de la statistique inférentielle, poursuivi par la mise en œuvre de tests, afin de montrer l'influence d'un facteur en contrôlant le risque d'erreur, soit le risque de rejeter à tort une hypothèse dite H_0 et donc de considérer que le facteur à un impact, alors qu'il n'en a pas. C'est le cas typique des essais cliniques de phase III, durant lesquels une molécule est prescrite en double aveugle à un groupe témoin, tandis que le groupe contrôle reçoit un placebo. Pour beaucoup de disciplines académiques, le test statistique constitue un outil de preuve scientifique même si son usage, parfois abusif, est mis en cause voire controversé à cause du manque de reproductibilité de trop nombreuses publications scientifiques (Ioannidis 2016).

Le deuxième objectif est prédictif, en utilisant des modèles statistiques classiques ou les algorithmes d'apprentissage automatique plus récents et sophistiqués. Deux sous-objectifs sont à considérer ; le premier est une prévision avec explication des résultats, de la façon dont les variables X_j influent sur la cible ou variable réponse Y. Le deuxième est une prévision brute sans recherche ou possibilité d'explication. Mais dans les deux cas, le *data scientist* sélectionne le modèle ou algorithme minimisant une estimation ou mesure d'une erreur de prévision qui contrôle le risque d'erreur de la décision qui en découle. In fine l'erreur de prévision de l'algorithme sélectionné est estimée

sur un échantillon test indépendant, différent de l'échantillon d'apprentissage sur lequel il a été entraîné ; c'est aussi à la base de toute procédure de certification précédant sa mise en exploitation.

Il y a donc, selon les objectifs, deux types de risque ou d'erreur. Celui de se tromper en affirmant qu'un facteur est influent et celui de se tromper de décision à cause d'une erreur de prévision. Laissons la question, largement débattue par ailleurs (Ioannidis 2016), de la pertinence des tests statistiques pour nous focaliser sur celle de la qualité de prévision plus spécifique à l'IA. Il existe de très nombreux critères ou métriques pour évaluer une erreur de prévision. Ce peut être un simple taux d'erreur pour la prévision d'une variable binaire : tissus pathologique ou sain, une erreur quadratique moyenne pour une variable Y quantitative. Dans beaucoup de publications du domaine de la santé, il est fait référence à l'aire sous la courbe ROC (*Area Under the Curve, AUC*) pour évaluer la qualité d'un algorithme pour une prévision binaire.

2.3 Facteurs de qualité d'une prévision

Plus précisément, quels sont les composants d'un modèle statistique ou algorithme d'apprentissage qui sont déterminants pour la qualité de prévision et donc pour les risques d'erreur de la décision qui en découle ?

Le point fondamental pour la qualité ou robustesse, voire la certification d'un algorithme d'apprentissage statistique, est, en tout premier lieu, la qualité des données disponibles, ainsi que leur représentativité du domaine d'étude ou d'application concerné. Les données d'entraînement de l'algorithme sont-elles bien représentatives de l'ensemble des situations ou cas de figure susceptibles d'être, par la suite, rencontrés lors de l'exploitation de l'algorithme ? Il s'agit d'anticiper une capacité de généralisation de son usage. En effet si des groupes ou des situations sont absents ou simplement sous-représentés c'est-à-dire si les données sont, d'une façon ou d'une autre, biaisées, le modèle ou l'algorithme qui en découle ne fait que reproduire les biais ou s'avère incapable de produire des prévisions correctes de situations qu'il n'a pas suffisamment apprises lors de son entraînement. Ce problème est très bien référencé dans la littérature et souligné dans les rapports et guides éthiques. C'est même un vieux problème déjà formalisé en statistique pour la constitution d'un échantillon relativement à une population de référence en planification d'expérience ou en théorie des sondages. Ce n'est pas parce que les données sont volumineuses, déjà acquises, qu'il faut pour autant tout prendre en compte ou ne pas se préoccuper d'en acquérir d'autres. Considérons l'exemple typique de la prévision d'événements rares mais catastrophiques. Un algorithme naïf, pour ne pas dire trivial, conduit à un très faible taux d'erreur, s'il ne prévoit aucune occurrence de l'événement rare mais est inutile voire dangereux. L'expérience du *data scientist* le conduit alors à sur-représenter (sur-échantillonnage) les événements rares, ou sous-échantillonner ceux très fréquents ou encore à introduire des pondérations dans le choix de la fonction objectif à optimiser. Ces pondérations dépendent de l'asymétrie des coûts, à évaluer par des experts métier, d'un faux positif ou prévision à tort d'un événement exceptionnel, relativement au coût induit par un faux négatif qui n'anticipe pas la catastrophe.

La précédente question concerne la représentativité des individus ou situations présentes dans la base d'entraînement relativement à une population théorique de référence. La deuxième soulève celle du choix ou de la disponibilité des caractéristiques (*features*) ou variables observées sur ces individus. Elle peut se formuler de la façon suivante : les causes effectives de la cible ou variable Y à modéliser, ou des variables qui lui seraient très corrélées, sont-elles bien prises en compte dans les observations ? Dans le même ordre d'idée et avec les mêmes conséquences, des mesures peuvent être erronées, soumises à du bruit. Ces questions ne sont pas plus faciles à résoudre que celles de représentativité précédentes. Sauf par imputation de données manquantes (prévision) il n'est pas possible de palier une absence d'information. Il n'est pas possible non plus de rectifier des erreurs de mesures, à moins qu'elles ne soient détectées comme atypiques, ou de labellisations, mais il est plus simple d'en circonscrire les conséquences en estimant précisément les erreurs d'ajustement du modèle ou d'entraînement de l'algorithme puis celles de prévision ; elles resteront plus ou moins importantes mais évaluables, quel que soit le nombre de variables pris en compte ou le volume des données accumulées.

Plus précisément, la taille de l'échantillon ou le nombre d'instances de la base d'apprentissage intervient à deux niveaux sur la qualité de prévision. La taille nécessaire dépend, d'une part, de la complexité de l'algorithme, du nombre de paramètres ou de poids qui en définit la structure et, d'autre part, de la variance du bruit résiduel ou erreur de mesure. Un algorithme est entraîné, en moyenne, et la taille de l'échantillon doit être d'autant plus grande que la variance de l'erreur de mesure est importante. Les réseaux de neurones profonds appliqués à des images de plusieurs millions de pixels sont composés de

dizaines de couches pouvant comporter des millions de paramètres ou poids à estimer ; ils nécessitent des bases de données considérables.

Attention, lorsque n est très grand (*big data*) le modèle peut être bien estimé car c'est une estimation en moyenne dont la précision s'améliore proportionnellement avec la racine de n. En revanche, une prévision individuelle est toujours impactée par le bruit résiduel du modèle, sa variance, quelle que soit la taille de l'échantillon. Aussi, même avec de très grands échantillons, la prudence est de mise quant à la précision de la prévision d'un comportement individuel surtout s'il est mal ou peu représenté dans la base : acte d'achat, acte violent, défaut de paiement, occurrence d'une pathologie.

En résumé, les applications quotidiennes de l'IA sont l'exploitation d'algorithmes d'apprentissage statistique, particulièrement sensibles à la *qualité des données d'entraînement*. Leur quantité est importante mais ne suffit pas à garantir la précision de prévisions individuelles qui doit être évaluée avec soin, afin de garantir, certifier, l'usage d'un algorithme. Malgré les abus de communication, l'IA ne se résume pas à l'utilisation de l'apprentissage profond. Le succès très médiatisé de certaines de ses applications ne doit pas laisser croire que ces relativement bons résultats en reconnaissance d'images ou traduction automatique sont transposables à tout type de problème.

Enfin, à l'exception des modèles statistiques élémentaires car linéaires ou à celle des arbres binaires de décision, les algorithmes d'apprentissage statistique sont opaques à une interprétation fine et directe de l'influence des caractéristiques d'entrée ou variables explicatives sur la prévision de la variable cible Y. Ce point soulève des problèmes délicats lorsqu'il s'agit de fournir l'explication intelligible d'une décision automatique.

Ce tout d'horizon met en évidence que l'usage de l'IA au quotidien soumet la société à des impacts dont les risques sont maintenant bien identifiés (Besse et al. 2017, Besse et al. 2019a) mais interdépendants et donc dans une situation très complexe qui nécessite la recherche permanente d'un meilleur compromis.

1. Protection : propriété, confidentialité des données personnelles (RGPD, CNIL) ;
2. Qualité, robustesse, résilience des prévisions donc des décisions ;
3. Explicabilité *vs.* opacité des algorithmes ;
4. Biais & Discrimination des décisions algorithmiques.

Plus un cinquième risque d'*entrave à la concurrence* de la part des navigateurs ou des comparateurs de prix. Ces dernières pratiques font appel à d'autres types d'algorithmes (*ranking*) qui ne seront pas pris en compte dans ce chapitre mais dont les dérapages délictueux sont régulièrement condamnés par la justice.

3 Cadre juridique

Le cadre juridique français est composé d'un mille feuille de textes :

- Loi n ° 78-17 du 6/01/1978 relative à l'informatique aux fichiers et aux libertés (LIL1) ;
- Loi n ° 2015-912 du 24/07/2015 relative au renseignement ;
- Loi n ° 2016-1321 du 7/10/2016 pour une République Numérique (Lemaire ou LIL2) ;
- Décrets d'applications (2017) ;
- RGPD Règlement Général pour la Protection des Données 05-2018 ;
- Loi n ° 2018-493 du 20 juin 2018 informatique et libertés (LIL3) ;
- Code pénal ;
- Code des relations entre le public et les administrations ;
- Code de la Santé publique ;
- ...
- Conseil Constitutionnel Décision n ° 2018-765 DC du 12 juin 2018.

dont il est fort complexe de tirer une synthèse globale. L'analyse ci-dessous reprend celle de Besse et al. (2019a).

3.1 Protection des données

La publication du RGPD (Règlement Général européen sur la Protection des Données n°2016/679/UE) et son intégration dans les lois nationales ont considérablement impacté la gestion des données dont celles impliquant des personnes physiques avec l'introduction de la notion de *Data Privacy Impact Assessment* (DPIA). Cette évaluation du bon usage des données est produite par un outil développé par la CNIL sous la forme d'un logiciel d'Analyse d'Impact relative à la Protection des données. Du point de vue juridique, il s'agit d'un *renversement de la charge de la preuve*. Ce n'est pas à la CNIL ou un usager d'apporter la preuve d'une fuite mais au DPO (*data protection officer*) d'une entreprise de montrer, en cas de contrôle, qu'il maîtrise la sécurité des données personnelles dans toute la chaîne de traitement de l'acquisition

à la décision. La constatation de défaillances est l'occasion de très lourdes sanctions financières : jusqu'à 20M€ et majorée pour une entreprise à 4 % du chiffre d'affaire annuel mondial

3.2 Qualité d'une décision

La question délicate de la qualité d'une décision algorithmique associée à une estimation d'erreur de prévision n'est pas explicitement présente dans RGPD ni les lois nationales qui en découlent. Notons néanmoins le considérant 71 du RGPD qui recommande :

> [...] Afin d'assurer un traitement équitable et transparent à l'égard de la personne concernée, [...] le responsable du traitement *devrait* utiliser des *procédures mathématiques ou statistiques* adéquates aux fins du profilage, appliquer les mesures techniques et organisationnelles appropriées pour faire en sorte, en particulier, que les facteurs qui entraînent des erreurs dans les données à caractère personnel soient corrigés et que le *risque d'erreur soit réduit au minimum*, sécuriser les données à caractère personnel d'une manière qui *tienne compte des risques* susceptibles de peser sur les intérêts et les droits de la personne concernée, et prévenir, entre autres, les *effets discriminatoires* [...]

Actuellement, la loi n'oblige pas de façon générale à communiquer les risques d'erreur, comme c'est le cas pour un sondage d'opinion. En revanche cet aspect est pris en charge dans chaque domaine d'application lorsqu'il est question de certification ou, pour un dispositif de santé connecté (DSC), de sa demande de remboursement. Il est traité dans la section 4 suivante.

3.3 Explication d'une décision

Le rapport Villani appelle à *ouvrir les boîtes noires* de l'IA car une grande partie des questions éthiques soulevées tiennent de l'opacité de ces technologies. Compte tenu de leur place grandissante, pour ne pas dire envahissante, le rapport considère qu'il s'agit d'un enjeu démocratique.

L'article 10 de la loi n ° 78-17 relative à l'informatique, aux fichiers et aux libertés du 6 janvier 1978 prévoyait à l'origine 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. 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 n'a pour l'instant pas donné lieu à sanction.

Parallèlement, le RGPD, et avant lui la directive 95/46/CE, consacre un certain nombre de droits en cas de décision individuelle prise sur le fondement d'un traitement automatisé de données :

1. Le droit d'accès et d'être informé de l'existence d'une prise de décision automatisée (RGDP, art. 13-15h) ;
2. 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) ;
3. Le droit d'obtenir une intervention humaine de la part du responsable du traitement (RGDP, art. 22.3) ;
4. Le droit d'exprimer son point de vue et de contester la décision (RGDP, art. 22.3) ; Les données sensibles doivent en principe être exclues des traitements exclusivement automatisés (art. 22.4), sauf en cas de consentement explicite ou pour des raisons d'intérêt public. Cependant, des exceptions sont aussi prévues (art. 22.2), lorsque la décision :
   - a) est nécessaire à la conclusion ou à l'exécution d'un contrat entre la personne concernée et un responsable du traitement ;
   - b) est autorisée par le droit de l'Union ou le droit de L'État 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 ;
   - c) 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. Par ailleurs, en dehors des activités du numériques, les hypothèses précédemment citées d'accès à un crédit, un logement, à des biens ou services, reposent le plus souvent sur la conclusion d'un contrat.

En outre, le point c) du paragraphe précédent 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, spécialement lorsque ce consentement est recueilli en ligne.

Ces dispositions ont été intégrées au droit français avec l'adoption récente de la loi n ° 2018-493 du 20 juin 2018 qui vient modifier la loi n ° 78-17 dite informatique et libertés du 6 janvier 1978. L'article 21 modifie l'article 10 de la loi du 6 janvier 1978 afin d'étendre les cas dans lesquels, par exception, une décision produisant des effets juridiques à l'égard d'une personne ou l'affectant de manière significative peut être prise sur le seul fondement d'un traitement automatisé de données à caractère personnel. L'article 10 alinéa 1er de la loi n ° 78-17 dispose désormais que *Aucune décision de justice impliquant une appréciation sur le comportement d'une personne ne peut avoir pour fondement un traitement automatisé de données à caractère personnel destiné à évaluer certains aspects de la personnalité de cette personne*.

L'alinéa 2 ajoute que *Aucune décision produisant des effets juridiques à l'égard d'une personne ou l'affectant de manière significative ne peut être prise sur le seul fondement d'un traitement automatisé de données à caractère personnel, y compris le profilage*. À ce principe, deux exceptions sont prévues.

La première se réfère aux exceptions du RGPD, c'est-à-dire *les cas mentionnés aux a et c du 2 de l'article 22 précité, sous les réserves mentionnées au 3 de ce même article et à condition que les règles définissant le traitement ainsi que les principales caractéristiques de sa mise en œuvre soient communiquées, à l'exception des secrets protégés par la loi, par le responsable de traitement à l'intéressé s'il en fait la demande*. Outre les garanties prévues par le texte européen à l'article 22.3 (droit d'obtenir une intervention humaine de la part du responsable du traitement, droit d'exprimer son point de vue et de contester la décision), le législateur français a ajouté l'obligation de communiquer les règles définissant le traitement, ainsi que les principales caractéristiques de sa mise en œuvre à la demande de la personne concernée. Cette garantie n'a plus cours si ces règles font l'objet de secrets protégés par la loi. Cette réserve vient ici aussi substantiellement affaiblir le principe, alors même qu'une communication des règles préservant le respect des secrets pourrait aisément s'envisager.

Quant à la deuxième exception prévue à l'article 10 al. 2 de la loi n° 78-17 modifiée, elle s'appuie sur le point b) de l'article 22.2 du RGPD, selon lequel chaque État membre peut prévoir librement des exceptions, dès lors qu'elles sont légalement prévues et respectent certaines garanties. Le législateur français a posé une exception pour les décisions administratives individuelles, à condition que le traitement ne porte pas sur des données sensibles, que des recours administratifs sont possibles et qu'une information est délivrée sur l'usage de l'algorithme. Cette exception ici précisée était déjà consacrée à l'article 4 de la loi n ° 2016-1321 pour une république numérique du 7 octobre 2016, codifiée à l'article L. 311-3-1 du CRPA, selon lequel une décision administrative individuelle prise sur le fondement d'un traitement algorithmique doit comporter une mention explicite en informant l'intéressé. L'article 1er du décret n ° 2017-330 du 14 mars 2017, codifiée à l'article R. 311-3-1-1 CRPA, précise que la mention explicite doit indiquer la finalité poursuivie par le traitement algorithmique. Elle rappelle le droit d'obtenir la communication des règles définissant ce traitement et des principales caractéristiques de sa mise en œuvre, ainsi que les modalités d'exercice de ce droit à communication et de saisine, le cas échéant, de la commission d'accès aux documents administratifs. La loi n ° 2018-493 du 20 juin 2018 est venue préciser que la mention explicite précitée est exigée à peine de nullité. La sanction de la violation de cette obligation d'information est donc explicitement prévue.

Depuis l'adoption de la loi pour une république numérique le 7 octobre 2016, l'article L. 311-3-1 prévoit par ailleurs que *les règles définissant ce traitement ainsi que les principales caractéristiques de sa mise en œuvre sont com-*

muniquées par l'administration à l'intéressé s'il en fait la demande. Le décret n ° 2017-330, codifié à l'article R. 311-3-1-2, précise les informations à fournir sous une forme intelligible et sous réserve de ne pas porter atteinte à des secrets protégés par la loi :

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.

On constate qu'est maintenue la dérogation en cas de secrets protégés par la loi.

La loi n ° 2018-493 va plus loin quant à l'utilisation d'un système de traitement automatisé pour la *prise de décision administrative* et prévoit désormais une obligation d'explication. Elle dispose ainsi que *le responsable de traitement s'assure de la maîtrise du traitement algorithmique et de ses évolutions afin de pouvoir expliquer, en détail et sous une forme intelligible, à la personne concernée la manière dont le traitement a été mis en œuvre à son égard*. Un fameux *droit à explication* est explicitement consacré par la loi française, alors que le RGPD n'y fait clairement référence que dans le considérant 71. Les articles 13 à 15 se contentent de prévoir un droit d'information et d'accès sur l'utilisation d'un dispositif automatisé et la *logique sous-jacente*, ce qui constitue une approche très générale, déconnectée des situations individuelles des personnes concernées.

Ajoutons que la loi n ° 2018-493 a fait l'objet d'une décision du Conseil constitutionnel n ° 2018-765 DC le 12 juin 2018. Notons seulement le point 71 : *... Il en résulte que ne peuvent être utilisés, comme fondement exclusif d'une décision administrative individuelle, des algorithmes susceptibles de réviser eux-mêmes les règles qu'ils appliquent, sans le contrôle et la validation du responsable du traitement* remettant ainsi en cause l'utilisation administrative d'algorithme d'apprentissage par renforcement.

Différentes situations peuvent être schématiquement considérées pour l'application de ces règles. Dans le cas d'un algorithme procédural de type ParcoursSup, les règles de fonctionnement doivent être clairement explicitées ; le Ministère concerné s'y est préparé à la suite des difficultés rencontrées par le prédécesseur APB (admission post bac). En effet, le code de l'algorithme Parcoursup est certes rendu public mais, source d'un débat ou controverse car les règles de délibérations locales à un établissement peuvent rester confidentielles rendant finalement opaque et éventuellement discriminatoire le processus. Enfin, la loi n ° 2018-493 prévoit que, s'agissant plus particulièrement des décisions prises en matière éducative dans le cadre de ParcoursSup, *le comité éthique et scientifique mentionné à l'article L.612-3 du code de l'éducation remet chaque année, à l'issue de la procédure nationale de préinscription et avant le 1er décembre, un rapport au Parlement portant sur le déroulement de cette procédure et sur les modalités d'examen des candidatures par les établissements d'enseignement supérieur. Le comité peut formuler à cette occasion toute proposition afin d'améliorer la transparence de cette procédure*.

Actuellement, L'obligation d'explicabilité, impose au mieux une intervention humaine pour assumer une décision et n'est contraignante que pour les décisions administratives françaises. Plus ou moins adaptée à des algorithmes procéduraux de type Parcousup, elle semble (décret n ° 2017-330) inapplicable à des algorithmes complexes (opaques) d'apprentissage statistique et interdit l'usage d'algorithmes auto-apprenants sans contrôle ou validation humaine, comme ce peut être le cas de ventes de publicités en ligne avec un algorithme d'optimisation stochastique dit de bandit manchot.

3.4 Biais et discrimination d'une décision

Selon l'article 225-1 du code pénal : *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.*

Concernant les groupes, l'alinéa 2, art. 1er de la loi n ° 2008-496 du 27 mai 2008 portant diverses dispositions d'adaptation au droit communautaire dans le domaine de la lutte contre les discriminations prévoit : *Constitue une discrimination indirecte une disposition, un critère ou une pratique neutre en apparence, mais susceptible d'entraîner, pour l'un des motifs mentionnés au premier alinéa (critères de la discrimination que l'on retrouve aussi dans le code pénal), un désavantage particulier pour des personnes par rapport à d'autres personnes, à moins que cette disposition, ce critère ou cette pratique ne soit objectivement justifié par un but légitime et que les moyens pour réaliser ce but ne soient nécessaires et appropriés.*

Apparaît aussi la notion de discrimination systémique qui serait : un processus qui met en jeu un système d'acteurs dans lequel personne ne manifeste directement d'intention discriminatoire, mais dont le résultat sera de produire une situation de discrimination. Les discriminations systémiques ne sont pas intentionnelles, elles proviennent de la somme de plusieurs représentations qui, cumulées, forment un contexte discriminant. Ce concept découle de la reconnaissance de l'existence de déséquilibres socio-économiques ou d'inégalités sociales qui sont historiquement constitués : Les discriminations systémiques sont donc constituées par les processus qui produisent et reproduisent les places sociales inégalitaires en fonction de l'appartenance à une classe, une "race" ou un sexe, cette appartenance pouvant être réelle ou supposée.

L'article 225-2 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 à :*

1. *refuser la fourniture d'un bien ou d'un service ;*
2. *entraver l'exercice normal d'une activité économique quelconque ;*
3. *refuser d'embaucher, à sanctionner ou à licencier une personne.*

La loi française insiste plus particulièrement sur une approche individuelle de la notion du risque de discrimination même si la notion de discrimination envers un groupe ou discrimination indirecte y est citée. La définition ou caractérisation de cette dernière n'est pas explicitée dans la loi tandis que le rapport Villani insiste sur la nécessité de définir un outil d'évaluation afin d'en faciliter la preuve et en permettre la sanction. Il évoque le *Discrimination Impact Assessment* (DIA) en complément du *Data Protection Impact Assessment (DPIA)* prévu par le RGPD et qui protège les données personnelles des individus et non des groupes. Ce n'est pas du tout évoqué dans le rapport Villani mais il existe une littérature abondante sur ce sujet sous l'appellation de *disparate impact* (effet disproportionné) depuis les années soixante-dix aux USA.

De son côté, le règlement européen encadre strictement la collecte de données personnelles sensibles (orientation religieuse, politique, sexuelle, origine ethnique...) et interdit aux responsables de décisions algorithmiques de les prendre en compte dans les traitements automatisés (art. 22.4), sous réserve du consentement explicite de la personne ou d'un intérêt public substantiel. Par opposition à discriminatoire, une décision est dite loyale équitable (*fair*) 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.

Actuellement et contrairement aux risques évoqués précédemment (qualité, explicabilité), les lois européenne et nationales condamnent très explicitement les risques discriminatoires. Le problème, à peine évoqué dans le rapport Villani est le manque d'élément qui permettrait de qualifier une situation discriminatoire, individuelle ou de groupe et donc d'en apporter la preuve. Ce point est développé dans la section 4 suivante.

3.5 Futur cadre réglementaire européen de l'IA

Les risques provoqués par les impacts dus aux erreurs des décisions, à l'opacité, aux biais algorithmiques (considérant 71 du RGPD) n'ont finalement pas ou peu été pris en compte dans une réglementation européenne visant en priorité la protection des données. Ils ont été en revanche largement commentés dans de très nombreuses déclarations, chartes pour une IA éthique au service de l'humanité. Pour remédier à ces lacunes, la Commission Européenne (CE) a réuni un groupe d'experts indépendants de haut niveau sur l'IA qui ont rédigé un guide sous la forme de lignes directrices en matière d'éthique pour une IA digne de confiance (2019) qui s'achève sur la proposition d'une liste d'évaluation sur le principe de celle sur la protection des données. Ces recommandations ont ensuite donné lui à la rédaction et la publication d'un livre blanc sur l'Intelligence Artificielle: une approche européenne axée sur l'excellence et la confiance (2020).

Ce livre souligne l'importance prise par l'IA, qui *combine données, algorithmes et puissance de calcul*, dans tous les aspects de la vie des citoyens,

en liste les bénéfices attendus, mais met également en exergue les *risques potentiels, tels que l'opacité de la prise de décisions, la discrimination*, qui accompagnent son développement et sa mise en œuvre. C'est un enjeu majeur car l'acceptabilité de l'IA et donc son adoption par les citoyens ne seront possibles que si celle-ci est *digne de confiance*. La CE, qui ambitionne de faire de l'Europe un *acteur mondial de premier plan en matière d'innovation dans l'économie fondée sur les données et dans ses applications*, insiste sur la nécessité de cette confiance fondée sur les *droits fondamentaux* de la *dignité humaine et la protection de la vie privée.*

Il s'agit donc pour la CE de proposer les *éléments clefs d'un futur cadre réglementaire* basé sur un *écosystème de confiance* en prenant en compte les lignes directrices en matière d'éthique élaborées par le groupe d'experts et dont une *liste d'évaluation* servirait de base pour un *programme indicatif destiné aux développeurs de l'IA* et une *ressource mise à la disposition des établissements de formation*. La CE insiste sur la liste des exigences énumérées par le groupe d'experts en remarquant que si certaines sont prises en compte par les régimes législatifs ou réglementaires existants, d'autres (*e.g.* transparence, contrôle humain) ne sont pas couvertes ou qu'il est de toute façon *difficile de déceler et de prouver d'éventuelles infractions à la législation, notamment aux dispositions juridiques qui protègent les droits fondamentaux*, à cause de l'opacité des algorithmes d'IA.

Par ailleurs, suivant en cela le groupe d'experts, la CE insiste tout particulièrement sur la classe de systèmes d'intelligence artificielle basés sur des *algorithmes d'apprentissage automatique* et donc sur le rôle fondamental des *données* utilisées pour leur entraînement.

Remarque : algorithmes déterministes ou procéduraux ou encore d'IA symbolique. Ce chapitre laisse apparemment de côté cette classe d'algorithmes décisionnels (*e.g. calcul de taxes, impôts, allocations ou prestations sociales,...* basés sur un ensemble de règles de décision déterministes qui peuvent tout autant présenter des impacts de désavantage ou risques de discrimination indirecte malgré une apparente neutralité. La détection de ces risques relève de l'analyse experte des règles de décisions codées dans l'algorithme. Néanmoins, la complexité de l'algorithme peut être telle (cf. Parcoursup) qu'une l'analyse experte *ex post* ne sera pas en mesure d'évaluer l'étendue des risques. En conséquence, l'algorithme déterministe peut être traité avec le même niveau d'opacité et les mêmes outils qu'un algorithme d'apprentissage statistique.

En résumé Les lois actuelles listées dans cette section ne sont pas contraignantes ou finalement inapplicables à des décisions complexes issues d'un algorithme d'apprentissage. Néanmoins et compte tenu du temps nécessaire au déploiement d'un système d'IA, de l'acquisition des données à sa mise en exploitation, il est urgent pour les responsables d'un système d'IA d'anticiper sur le cadre réglementaire européen à venir qui se présentera sous la forme d'une procédure d'évaluation (version pilote) des points fondamentaux identifiées par les experts européens :

1. Action humaine et contrôle humain ;
2. Robustesse technique et sécurité (résilience, précision...) ;
3. Respect de la vie privée et gouvernance des données (qualité...) ;
4. Transparence (explicabilité, communication...) ;
5. Diversité, non-discrimination et équité ;
6. Bien-être sociétal et environnemental (durabilité, interactions...), Utilité ;
7. Responsabilité (auditabilité, recours...).

Voici à titre illustratif quelques unes des lignes directrices rédigées par les experts de la CE, première ébauche d'une base probable à l'évaluation de la confiance d'un système d'IA :

- (52) Si les *biais injustes* peuvent être évités, les systèmes d'IA pourraient même *améliorer le caractère équitable de la société*.
- (53) L'*explicabilité* est essentielle... les décisions – dans la mesure du possible – doivent pouvoir être expliquées.
- (69) Il est important que le système puisse indiquer le *niveau de probabilité de ces erreurs*.
- (80) *Absence de biais injustes*, La persistance de ces biais pourrait être *source de discrimination et de préjudice (in)directs*. Dans la mesure du possible, les *biais détectables et discriminatoires devraient être supprimés* lors de la phase de collecte.
- (106) (107) besoin de *normalisation* (IEEE, ANSI, AFNOR...).

Comme pour la sécurité des données, ce ne sera plus à un individu d'apporter la preuve d'un manquement à la loi mais bien au responsable d'un système d'IA de montrer qu'il a pris toutes les mesures nécessaires pour que celle-ci soit respectée.

En conséquence, nous proposons dans les sections suivantes, non pas une liste exhaustive des questions auxquelles, il sera important de chercher des réponses mais une sélection illustrative de celles-ci en attendant une version plus aboutie de normes souhaitées. Notons que cette anticipation est déjà une réalité dans le domaine de la santé à la demande des organismes responsables de la certification (FDA aux USA) ou de l'autorisation de remboursement (HAS 2020) en France des DSC (dispositifs de santé connectés) embarquant un algorithme d'apprentissage.

4 Qualité, précision et robustesse

4.1 Évaluation réglementaire

Comme expliqué section 2, l'évaluation des erreurs de prévision qui conditionnent directement la qualité de décision et donc son niveau de confiance, est essentielles lors de la mise au point d'un système d'IA. Elle est même partie intégrante de la procédure d'entraînement. Elle doit être menée avec une grande rigueur notamment dans la constitution de l'échantillon test indépendant et représentatif du domaine d'exploitation de l'algorithme. Actuellement un grand flou est entretenu en faisant valoir et même en communiquant exagérément sur les très bonnes performances de certains systèmes d'IA en reconnaissance d'image afin de masquer les piètres performances d'autres systèmes dédiés à la prévision de comportements individuels humains. Rappelons que les taux d'erreur de prévision de la récidive d'un détenu varient entre 30 et 40%, pas beaucoup mieux qu'un tire à pile ou face, de même que les prévisions de passage à un acte d'achat lors de publicités ciblées en ligne.

L'intérêt commercial des principaux acteurs de ce dernier secteur induit une stratégie bien identifiée de lobbying (*ethical washing*) qui consiste à afficher des principes éthiques (*soft law*) afin de freiner toute tentative de réglementation certes contraignante pour l'innovation technologique mais éclairante sur leurs pratiques et les performances effectives des algorithmes de recommandation en ligne, principale source de leurs revenus.

D'un point de vue éthique il n'y a pas d'obligation de moyen, sauf dans le cadre explicite d'une norme industrielle de certification. Il y a en revanche une obligation de transparence qu'il importe de rendre obligatoire.

Ceci est pris explicitement en compte dans les questions de la liste d'évaluation des experts de la CE :

- Avez-vous évalué le *niveau de précision* et la *définition* de la précision nécessaires dans le contexte du système d'IA et du cas d'utilisation concerné ?
- Avez-vous réfléchi à la manière dont la *précision* est mesurée et assurée ?
- Avez-vous mis en place des mesures pour veiller à ce que les données utilisées soient *exhaustives* et à jour ?
- Avez-vous mis en place des mesures pour évaluer si des données supplémentair sont nécessaires, par exemple pour améliorer la précision et *éliminer les biais* ?

4.2 Éléments de réponse

Les mesures de *précision* de la prévision d'un système d'IA sont bien connues et maîtrisées, même si l'éventail des possibles. Le choix, précisément justifié, doit être adapté au domaine, au type de problème traité aux risques spécifiques encourus.

- *Régression* : variable cible Y quantitative
  Fonction perte L_2 (quadratique) ou L_1 (valeur absolue)
- *Classification* binaire
  Taux d'erreur, AUC (*area under the ROC Curve*), score F_β, entropie...
- *Multiclasse*
  Taux d'erreur moyen, F_β moyen...

L'évaluation de la *robutesse* est lié aux procédures de contrôle mises en place pour détecter des valeurs atypiques (*outliers*)ou anomalies dans la base d'apprentissage et au choix de la fonction perte de la procédure d'entraînement de l'algorithme. Impérativement, surtout dans les d'applications sensibles pouvant entraîner des risques élevés en cas d'erreur, la détection d'anomalie doit également être intégrée en exploitation afin de ne pas chercher à proposer des décisions correspondant à des situations inconnues de l'apprentissage.

Enfin, la *résilience* d'un système d'IA est essentielle pour les dispositifs critiques (dispositifs de santé connecté, aide au pilotage). Il concerne par exemple la prise en compte de données manquantes lors de l'apprentissage comme en exploitation. Il s'agit d'évaluer la capacité d'un système d'IA à

assurer des fonctions pouvant s'avérer vitales en cas, par exemple, de panne ou de fonctionnement erratique d'un capteur : choix d'un algorithme tolérant aux données manquantes, imputation de celles-ci, fonctionnement en mode dégradé.

5 Explicabilité

5.1 Évaluation réglementaire

Ce point est le plus complexe à traiter. Il est un domaine de recherche extrêmement actif notamment pour les applications industrielles de systèmes d'IA embarqués dans un véhicule autonome ou un avion à un seul pilote et qui nécessiteront des procédures de certification particulièrement exigeantes. Barredo Arrieta et al. (2020) proposent une revue de cette recherche en cours tentant une synthèse de plus de 400 références bibliographiques !

Exemples de questions posés par les experts dans la liste d'évaluation :

- Avez-vous évalué la mesure dans laquelle les *décisions prises*, et donc les résultats obtenus, par le système d'IA peuvent être *compris* ?
- Avez-vous veillé à ce qu'une *explication de la raison* pour laquelle un système a procédé à un certain choix entraînant un certain résultat puisse être rendue *compréhensible* pour l'*ensemble des utilisateurs* qui pourraient souhaiter obtenir une explication ?

5.2 Éléments de réponse

Il est encore beaucoup trop tôt pour tenter un résumé opérationnel de ce thème. Il faut pour cela attendre que la recherche progresse et qu'une "sélection naturelle" en extrait les procédures les plus pertinentes. Tentons de décrire les premiers embranchements d'un arbre de décision en répondant à quelques questions rudimentaires qu'il faudrait en plus adapter au domaine d'application car le type de réponse à apporter n'est évidemment pas le même s'il s'agit d'expliquer le refus d'un prêt ou les conséquences d'une aide automatisée au diagnostic d'un cancer.

L'explication peut concerner :

1. Le fonctionnement général de l'algorithme
   - dans le cas d'un modèle "transparent" : modèles linéaires, arbres de décision, l'explication est possible à condition que le nombre de variables, d'interactions reste raisonnable,
   - dans le cas d'un algorithme complexe opaque :
     — chercher une approximation : linéaire, arbre, règles de décision déterministes ;
     — chercher les variables importantes par randomisation des valeurs ou stress de l'algorithme (Bachoc et al. 2020).
2. Une décision spécifique pour :
   - le concepteur : expliquer une erreur, y remédier (ré-apprentissage) ;
   - la personne concernée : client, patient, justiciable :
     — modèle interprétable : linéaire, arbre de décision,
     — approximation locale : LIME, contre-exemple, règles,...
     — explication *a minima* du risque d'erreur.

Quelques démonstrations de procédures explicatives sont proposées sur des sites collaboratifs. Citons :

- https://www.gems-ai.com/
- https://aix360.mybluemix.net/
- https://github.com/MAIF/shapash

Ne pas perdre de vue que l'impossibilité ou simplement la difficulté à formuler une explication provient de l'utilisation d'algorithmes opaques mais dont la nécessité est inhérente à la complexité même du réel. Un réel complexe (*e.g.* les fonctions du vivant) impliquant de nombreuses variables, des interactions, voire des boucles de contre-réaction, est nécessairement modélisé par un algorithme complexe afin d'éviter des simplifications abusives pouvant gravement nuire à ses performances. C'est tout d'abord le réel qui peut s'avérer complexe à expliquer.

6 Biais et risques de discrimination

6.1 Évaluation réglementaire

La liste d'évaluation du groupe d'experts, base de réflexion de la CE, réserve la section 5 *Diversité, non-discrimination et équité* aux questions de discrimination. Relevons seulement trois questions de cette longue liste adressées aux concepteurs d'un système d'IA :

- Avez-vous prévu une définition appropriée de l'équité que vous appli-

quez dans la conception des systèmes d'IA ?
- Avez-vous mis en place des processus pour tester et contrôler les biais éventuels au cours de la phase de mise au point, de déploiement et d'utilisation du système ?
- Avez-vous prévu une analyse quantitative ou des indicateurs pour mesurer et tester la définition appliquée de l'équité ?

Il s'agit d'exemples typiques de questions auxquelles il est difficile de répondre sans définition claire, en termes juridiques, des concepts employés. Ainsi, le cadre juridique ne fournit aucune définition de l'équité mais condamne explicitement la discrimination. Corrélativement, l'équité d'une décision algorithmique devient l'absence de risque discriminatoire donc de biais.

6.2 Détecter une discrimination

Avant de s'intéresser à la détection d'une discrimination algorithmique, il est opportun d'évaluer les capacités de détecter une discrimination humaine. Prenons l'exemple critique de l'embauche identifié à haut risque par la CE.

Testing

La détection et même la preuve d'une discrimination directe envers une personne peut être obtenue par *testing*. Cette pratique consiste à adresser à des dates distinctes deux dossiers, par exemple de candidature à un emploi. A l'exception de la caractéristique discriminatoire à tester : genre, origine ethnique, tranche d'âge, quartier d'habitation... les dossiers sont strictement similaires tout en introduisant des différences mineures afin d'éviter d'éventer le procédé. Son usage à été élargi (cf. Riach et Rich 2002) avec le déploiement d'enquêtes systématiques afin de viser l'objectif d'une mesure statistique de la discrimination indirecte envers un groupe. Les communautés académiques en Sociologie et Économie ont produit une vaste bibliographie à ce sujet (Rich 2014). En France, c'est la doctrine officielle diffusée par le Comité National de l'Information Statistique et déployée périodiquement par la DARES (Direction de l'Animation, des Études, de la Recherche et des Statistiques) lorsqu'il s'agit d'étudier les risques de discrimination à l'embauche. D'autres enquêtes par *testing* se ont également ciblé l'accès à l'assurance, au crédit ou encore au logement (cf. les rapports de recherche du TEPP).

Effet disproportionné

Aux USA, une approche très différente est développée avec la notion d'*adverse* ou *disparate impact* (effet disproportionné). L'évaluation de l'effet disproportionné consiste à estimer le rapport de deux probabilités : probabilité d'une décision favorable pour une personne du groupe sensible au sens de la loi sur la même probabilité pour une personne de l'autre groupe. Elle est appliquée depuis 1971 (Barocas et Selbst 2016) pour mesurer des discriminations indirectes dans l'accès à l'emploi, le logement, et a donné lieu à une réglementation officielle de son usage notamment pour l'accès à l'emploi :

> *Civil Rights act & Code of Federal Regulations*
> TITLE 29 - LABOR: PART 1607—UNIFORM GUIDELINES ON EMPLOYEE SELECTION PROCEDURES (1978)
> - D. *Adverse impact and the "four-fifths rule.*" A *selection rate* for any race, sex, or ethnic group which is *less than four-fifths (4/5) (or eighty percent)* of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact. *Smaller differences* in selection rate may nevertheless constitute adverse impact, where they are *significant in both statistical and practical* terms or where a user's actions have discouraged applicants disproportionately on grounds of race, sex, or ethnic group. Greater differences in selection rate may not constitute adverse impact where the differences are based on small numbers and are not statistically significant, or where special recruiting or other programs cause the pool of minority or female candidates to be atypical of the normal pool of applicants from that group.

L'estimation de ce rapport de probabilités (*odds ratio*) est donc comparée à une valeur arbitraire $0,8$, jugée suffisamment faible pour signifier un effet important malgré un aléa statistique de son estimation. Une valeur inférieure n'induit pas nécessairement des poursuites juridiques mais oblige une entreprise à justifier, pour des raisons économiques, les raisons de ce déséquilibre.

Remarques

Les éléments de ces approches statistiques sont également présents dans un guide publié par le Défenseur des Droits et la CNIL (2012). Il décrit une approche méthodologique à l'intention des acteurs de l'emploi pour mesurer et progresser dans l'égalité des chances sans volonté coercitive ni obligation juridique. En préalable, ce guide pose la question de l'opportunité de construire des statistiques ethniques alors que, contrairement aux USA, l'origine des personnes ne peut être enregistrée dans une base de données. Cette apparente protection des droits des personnes soulève un problème lorsqu'il est question d'évaluer une possible discrimination. La difficulté peut être contournée en adoptant une identification de l'origine par le patronyme au prix d'une perte sans doute mineure mais à évaluer de précision. Ce guide évoque la pratique du *testing* mais incite également les services de ressources humaines d'une entreprise à produire des tableaux statistiques (tables de contingence) desquels il serait facile d'extraire une évaluation quantitative de l'effet disproportionné.

Chacune des approches : *testing vs. disparate impact* présente des avantages mais également des défauts, biais ou difficultés de mise en œuvre. Le *testing* met bien en évidence une discrimination directe, intentionnelle, et peut conduire à une action en justice lorsqu'une personne est concernée. En revanche, utilisée lors d'une enquête systématique, il déploie des dossiers fictifs, fournis des résultats indicatifs, qui ne sont pas représentatifs de la politique d'embauche effective d'une entreprise sur l'ensemble des ses postes. Les enquêtes menées par la DARES ne conduisent pas à des actions en justice et la récente stratégie *name and shame* du gouvernement stigmatisant certaines entreprises a suscité de vives polémiques en janvier 2020.

Les enquêtes par *testing* ne nécessitent pas une participation des entreprises concernées mais sont d'un coût élevé et soulèvent de lourdes difficultés pour tenter d'approcher la réalité des embauches. En revanche, l'évaluation de l'effet disproportionné est de coût très faible mais implique une contribution loyale des services de ressources humaines ou une obligation réglementaire comme aux USA. Il est éventuellement biaisé puisque les dossiers ne sont pas identiques et nécessite donc une analyse ou la recherche d'autres explications possibles mais confondues des écarts observés.

6.3 Détecter une discrimination algorithmique

Une décision algorithmique ajoute une couche d'opacité sur une situation déjà complexe.

Indicateurs statistiques de discrimination

Le problème émergeant de la discrimination algorithmique s'exprime simplement : si un algorithme est entraîné sur des données biaisées, il reproduit très fidèlement ces biais systémiques ou de société ; plus grave, il risque même de les renforcer. Très prolixe, le monde académique a proposé quelques dizaines d'indicateurs (Zliobaitė 2017) afin d'évaluer des biais potentiels. Néanmoins, beaucoup de ces indicateurs s'avèrent très corrélés ou redondants (Friedler et al. 2019). Empiriquement, trois niveaux de biais discriminatoires doivent être pris en compte en priorité :

1. L'effet disproportionné reflet du biais social ou de population par lequel un groupe est historiquement (*e.g.* revenu des femmes) désavantagé. La mise en évidence de ce biais soulève des questions techniques, politiques évidentes. Renforcer algorithmiquement ce biais serait ouvertement discriminatoire, il importe de détecter, éliminer, un tel risque. Serait-il politiquement opportun d'introduire automatiquement une part de discrimination positive afin d'atténuer la discrimination sociale ? C'est techniquement l'objet d'une vaste littérature académique nommée apprentissage équitable (*fair learning*) et évoqué dans le travail des experts (ligne directrice 52) pour *améliorer le caractère équitable de la société*.
2. Les taux d'erreur de prévision et donc les risques d'erreur de décisions sont-ils les mêmes pour chaque groupe ? Ainsi, si un groupe est sous-représenté dans la base d'apprentissage, il est très probable que les décisions le concernant seront moins fiables. C'est typiquement le cas en reconnaissance faciale et ce risque est également présent dans les applications de l'IA en santé (Besse et al. 2019b).
3. Même si les deux critères précédents sont trouvés équitables et surtout si les taux d'erreur identiques sont relativement importants, les erreurs peuvent être dissymétriques (plus de faux positifs, moins de faux négatifs) au détriment d'un groupe. Cet indicateur (comparaison des rapports de cote ou *odds ratio*) est ainsi au cœur de la controverse

concernant l'évaluation COMPAS du risque de récidive aux USA (Larson et al. 2016).

Difficultés d'évaluation

Contrairement à des prises de position très naïves des entreprises proposant des algorithmes de prérecrutement, des décisions algorithmiques ne sont pas plus objectives que des décisions humaines. Il est même facile de montrer sur des exemples (numériques ci-après, De Arteaga et al. 2019) que les biais humains sont fidèlement reproduits voire amplifiés même si la variable sensible (genre, origine, âge...) est absente de la base de données car cette information est présente, d'une façon ou d'une autre, dans les autres variables jouant le rôle de variables de substitution ou *proxy*. Autre conséquence importante de cette situation, le *testing* est complètement inopérant (cf. sous section suivante) pour détecter une discrimination face à un algorithme.

En conséquence, les biais, risques de discrimination, doivent être soigneusement évalués très en amont lors de la constitution des bases de données et lors de la procédure d'apprentissage afin de les corriger ou les atténuer : *fairness by design*, au risque de ne plus être à même de pouvoir les détecter.

6.4 Exemple numérique de discrimination algorithmique

Nous proposons d'illustrer sur un exemple numérique élémentaire les difficultés rencontrées pour la détection et l'évaluation de ces risques fondamentaux.

Données

Les données publiques utilisées imitent le contexte du calcul d'un score de crédit. Elles sont extraites (échantillon de 45 000 personnes) d'un recensement de 1994 aux USA et décrivent l'âge, le type d'emploi, le niveau d'éducation, le statut marital, l'origine ethnique, le nombre d'heures travaillées par semaine, la présence ou non d'un enfant, les revenus ou pertes financières, le genre et le niveau de revenu bas ou élevé. Elles servent de référence ou *bac à sable* pour tous les développements d'algorithmes d'apprentissage automatique équitable. Il s'agit de prévoir si le revenu annuel d'une personne est supérieur ou inférieur à 50k$ et donc de prévoir, d'une certaine façon, sa solvabilité connaissant ses autres caractéristiques socio-économiques. L'étude complète et les codes de calcul sont disponibles mais l'illustration est limitée à un résumé succinct de l'analyse de la discrimination selon le sexe.

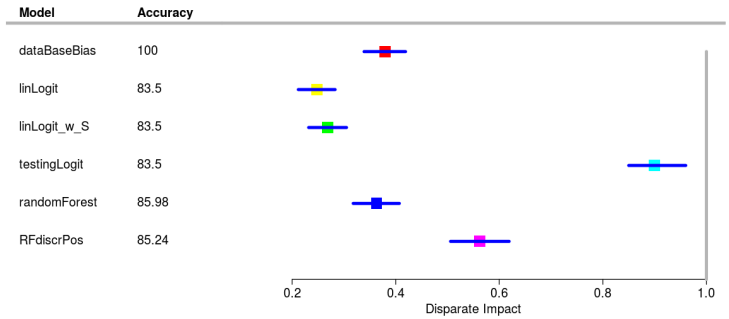

FIGURE 1 – *Précision de la prévision (accuracy) et effet disproportionné estimé par un intervalle de confiance sur un échantillon test (taille 9000) pour différents modèles ou algorithmes d'apprentissage.*

Résultats

Les données ont été aléatoirement réparties en deux échantillons d'apprentissage (36 000), destinés à l'estimation des modèles ou entraînement des algorithmes, et de test (9000) pour évaluer les différents indicateurs. Les résultats sont regroupés dans la figure 1.

Ils mettent en évidence un biais de société important : seulement $11,6\%$ des femmes ont un revenu élevé contre $31,5\%$ des hommes. Le rapport $DI = 0,38$ est donc très disproportionné. Il est comparé avec celui de la prévision de niveau de revenu par un modèle classique linéaire de régression logistique `linLogit` : $DI = 0,25$. Significativement moins élevé (intervalles de confiance disjoints), il montre que ce modèle renforce le biais et donc discrimine nettement les femmes dans sa prévision. La procédure naïve (`linLogit-w-s`) qui consiste à éliminer la variable dite sensible (genre) du modèle ne supprime en rien $(DI = 0,27)$ le biais discriminatoire car le genre est de toute façon

présent à travers les valeurs prises par les autres variables (effet *proxy*). Une autre conséquence de cette dépendance est que le *testing* (changement de genre toutes choses égales par ailleurs) ne détecte plus $(DI = 0.90)$ aucune discrimination !

Un algorithme non linéaire plus sophistiqué (*random forest*) est très fidèle au biais des données avec un indicateur $(DI = 0,36)$ pas significativement différent de celui du biais de société et fournit une meilleure précision : $0,86$ au lieu de $0,84$ pour la régression logistique. Cet algorithme ne discrimine pas plus mais c'est au prix de l'interprétabilité du modèle. Opaque comme un réseau de neurones, il ne permet pas d'expliquer une décision à partir de ses paramètres comme cela est facile avec le modèle de régression. Enfin, la dernière ligne propose une façon simple, parmi une littérature très volumineuse, de corriger le biais pour plus de *justice sociale*. Deux algorithmes sont entraînés, un par genre et le seuil de décision (revenu élevé ou pas, accord ou non de crédit...) est abaissé pour les femmes : $0,3$ au lieu de celui par défaut de $0,5$ pour les hommes. C'est une façon, parmi beaucoup d'autres, d'introduire une part de discrimination positive et d'atténuer le biais pour une *société plus équitable*.

Les autres types de biais sont également à considérer. Par principe, la précision de la prévision pour un groupe dépend de sa représentativité. Si ce dernier est sous-représenté, l'erreur est plus importante ; c'est typiquement le cas en reconnaissance faciale mais pas dans l'exemple traité. Alors qu'elles sont deux fois moins nombreuses dans l'échantillon, le taux d'erreur de prévision est de l'ordre de $7,9\%$ pour les femmes et de 17% pour les hommes. Il faut donc considérer le troisième type de biais pour se rendre compte que c'est finalement à leur désavantage. Le taux de faux positifs est plus important pour les hommes $(0,08)$ que pour les femmes $(0,02)$. Ceci avantage les hommes qui bénéficient plus largement d'une décision favorable même à tort. En revanche, le taux de faux négatifs est plus important pour les femmes $(0,41)$, à leur désavantage, que pour les hommes $(0,38)$. Noter que la procédure élémentaire d'atténuation du biais en entraînant deux algorithmes, un pour chaque genre, conduit à une légère augmentation du taux d'erreur pour les femmes, qui se rapproche un peu de celui des hommes, et surtout produit un taux de faux positifs plus élevés pour les femmes. Aussi, sur cet exemple, l'introduction d'une dose de discrimination positive intervient sur les trois types de biais pour en réduire l'importance.

Discussion

Nous pouvons tirer quelques enseignements de cet exemple rudimentaire imitant le calcul d'un score d'attribution de crédit bancaire.

- Sans précaution, si un biais est présent dans les données, il est reproduit et même renforcé par un modèle linéaire élémentaire.
- Un algorithme plus sophistiqué, non linéaire et impliquant les interactions entre les variables, ne fait que reproduire le biais mais, opaque, ne permet plus de justification des décisions si l'effet disproportionné est juridiquement attaquable $(DI < 0,8)$.
- La procédure de *testing*, déjà peut convaincante pour évaluer une discrimination indirecte *ex post*, est complètement inadaptée face à une procédure algorithmique.
- Actuellement en Europe, une ou un *data scientist* est libre de produire ce qu'il peut ou veut, en fonction de ses compétences et de sa déontologie personnelle : de l'algorithme élémentaire interprétable mais discriminatoire à celui incluant une part arbitraire de discrimination positive. Aucune procédure de contrôle que ce soit *ex ante* ou *ex post*, n'est en vigueur à ce jour pour le remettre en cause.
- La recherche d'une moins mauvaise solution sera l'affaire d'un compromis entre les trois exigences de base pour une IA de confiance : contrôle de la discrimination, qualité (robustesse réplicabilité) d'une décision et explicabilité de cette décision. En effet, le meilleur algorithme en termes de précision est opaque, ininterprétable, et donc inadapté pour éviter aux USA, une procédure judiciaire si l'effet disproportionné est trop important. De plus, la correction ou l'atténuation de l'effet disproportionné entraîne une dégradation de la qualité de la prévision. Les récents travaux de recherche en apprentissage équitable (Fairness, Accountability, and Transparency conferences) visent cette recherche de meilleur compromis.

En résumé, la détection d'un risque algorithmique de discrimination indirecte vis à vis d'un groupe est une question complexe basée sur l'estimation d'un choix d'indicateurs statistiques impliquant également les autres exigences de qualité et explicabilité. Cette estimation est de plus soumise à l'accès à l'information sensible dont l'enregistrement (*e.g.* origine ethnique) peut être interdite par le RGPD ; interdiction contournable par des procédés (*e.g.* analyse du patronyme) pouvant nuire à la précision.

En conséquence et malgré un bagage de recherches anciennes et bien documentées, apporter des réponses aux simples questions de la liste d'évaluation n'est pas immédiat !

Conclusion

Une chose est à retenir de cette rapide présentation et de l'exemple numérique proposé. Sans documentation précise et exhaustive sur le processus qui a conduit à la mise en exploitation d'un système d'IA, du recueil des données à sa mise en exploitation, un audit *ex post* est impossible. Pour évaluer par exemple une discrimination, les enquêtes classiques par *testing* sont hors-jeu et tester un système sur des données réalistes, nécessiterait une immersion complète dans la complexité du domaine d'application concerné. Le risque serait évidemment de ne tester que certains aspects du système, certaines situations ou types de données. La question principale reste donc la représentativité de ces tests par rapport à l'usage réel qui est fait du système pour en détecter les biais potentiels. Cette question est *de facto* un préalable indispensable à la création d'un système d'IA : quelles données pour quel objectif ? Si elle n'a pas été posée explicitement et documentée *ex ante*, une analyse *ex post* ne peut conduire qu'à une remise en cause de la fiabilité du système, en termes de qualité de décision ou de biais, devant l'impossibilité d'en définir précisément le domaine d'usage.

La mise en place de cette documentation *ex ante* sera la conséquence de l'exécution de la liste d'évaluation du groupe des experts européens reprise par le livre blanc de la CE. Elle suit le même principe et la même logique que l'analyse d'impact relatif à la confidentialité des données (*privacy impact assessment*) et devrait être formalisée dans une réglementation à venir.

Des capacités et des compétences, à la fois techniques (statistique, apprentissage automatique) et juridiques de la part des régulateurs ou de sociétés *ad hoc*, sont indispensables pour auditer *a minima* une telle documentation. Il s'agira en tout premier lieu, de s'assurer que la vérification *ex ante* de conception d'un système d'IA a été mise en place très en amont dans un souci de contrôle exigeant de la qualité à toutes les étapes : représentativité statistique des données d'entraînement en fonction de l'objectif, procédure d'apprentissage et évaluation des erreurs, des biais, validation, éventuelle qualification et mise en exploitation. Le cahier des charges doit également intégrer une surveillance du bon fonctionnement du système d'IA afin d'en contrôler tout risque de dérive et d'identifier les causes et responsabilités humaines en cas d'erreurs. Ce processus qualité peut imposer de devoir ré-entraîner périodiquement l'algorithme afin d'y intégrer des situations ou cas de figures initialement omis de la base de données.

Il serait irresponsable de ne pas anticiper la construction d'une telle documentation dès l'initialisation de la réalisation d'un systèmes d'IA au risque de devoir reproduire toute l'analyse lorsque l'obligation en sera réglementée. En conséquence il est important de former les futurs responsables de systèmes d'IA à ce type d'évaluation et à ses outils dont cet article fournit quelques éléments.

Références

- Bachoc F., Gamboa F., Halford M., Loubes J.-M., Risser L. (2020). Entropic Variable Projection for Model Explainability and Intepretability, arXiv preprint : 1810.07924.
- Barocas S., Selbst A. (2016). Big Data's Disparate Impact, 104 *California Law Review*, 104 671.
- Besse P., Castets-Renard C., Garivier A., Loubes J.-M. (2019a). L'IA du Quotidien peut elle être Éthique? Loyauté des Algorithmes d'Apprentissage Automatique, *Statistique et Société*, **6**-3.
- Besse P., Besse Patin A., Castets Renard C. (2019b). Implications juridiques et éthiques des algorithmes d'intelligence artificielle dans le domaine de la santé, à paraître.
- Besse P., Castets Renard C., Loubes J.-M., Risser L. (2020). Évaluation des Risques des Algorithmes d'Apprentissage Statistique de l'IA: ressources pédagogiques, tutoriels R et python en ligne consultés le 8/05/2020.
- De-Arteaga M., Romanov A. et al. (2019). Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting, *Proceedings of the Conference on Fairness, Accountability, and Transparency*.
- Commission Européenne (2019). Lifgnes directrices pour une IA de confiance.
- Commission Européenne (2020). Livre blanc sur l'intelligence artificielle: une approche européenne d'excellence et de confiance.
- De-Arteaga M., Romanov A. et al. (2019). Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting, in FAT'19,

pp 120–128.
- Défenseur des Droits, CNIL (2012). Mesurer pour progresser vers l'égalité des chances. Guide méthodologique à l'usage des acteurs de l'emploi.
- Friedler S., Scheidegger C., Venkatasubramanian S., Choudhary S., Hamilton E., Roth D. (). Comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, p. 329–38.
- HAS (2019). Guide sur les spécificités d'évaluation clinique d'un dispositif médical connecté (DMC) en vue de son accès au remboursement, *Évaluation des dispositifs médicaux par la CNEDiMTS*, Janvier 2019.
- Health Center for Devices and Radiological (2019). Artificial Intelligence and Machine Learning in Software as a Medical Device, *FDA*.
- Ioannidis J. (2016). Why Most Clinical Research Is Not Useful, *PLOS Medicine*, Volume 13, Issue 6.
- Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. *Nat Mach Intell* 1 :389–399.
- Larson J., Mattu S., Kirchner L., Angwin J. (2016). How we analyzed the compas recidivism algorithm. ProPublica, en ligne consulté le 28/04/2020.
- Raghavan M., Barocas S., Kleinberg J., Levy K. (2019) Mitigating bias in Algorithmic Hiring : Evaluating Claims and Practices, *Proceedings of the Conference on Fairness, Accountability, and Transparency*.
- Riach P.A., Rich J. (2002). Field Experiments of Discrimination in the Market Place, *The Economic Journal*, Vol. 112 (483), p F480-F518.
- Rich J. (2014). What Do Field Experiments of Discrimination in Markets Tell Us? A Meta Analysis of Studies Conducted since 2000, *IZA Discussion Paper*, No. 8584.
- Vayena E, Blasimme A, Cohen IG (2018) Machine learning in medicine: Addressing ethical challenges. *PLoS Med* 15 :e1002689.
- Zliobaité I. (2017). Measuring discrimination in algorithmic decision making, *Data Min. Knowl. Disc.*, 31, p 1060–89.