Statistique & Règlement Européen des Systèmes d'IA

Philippe Besse

Statistique & Règlement Européen des Systèmes d'IA (*AI Act*)

PHILIPPE BESSE

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

8 juin 2021

Résumé

Suite à la publication du livre blanc pour une approche de l'IA basée sur l'excellence et la confiance, la Commission Européenne (CE) a publié de nombreuses propositions de textes réglementaires dont un Artificial Intelligence Act (AI Act) (2021) établissant des règles harmonisées sur l'intelligence artificielle *(IA). Quels seront les conséquences et impacts de l'adoption à venir de ce texte du point de vue d'un mathématicien ou plutôt statisticien impliqué dans la conception de système d'intelligence artificielle (IA) à haut risque au sens de la CE ? Quels outils et méthodes vont permettre de répondre à l'obligation d'une analyse rigoureuse et documentée des données traitées, des performances, robustesse, résilience de l'algorithme, de son explicabilité, des risques, pour les droits fondamentaux, de biais discriminatoires ? Ces questions sont illustrées par un exemple numérique analogue à un score de crédit (cf. tutoriel) à la recherche d'un moins mauvais compromis entre toutes les contraintes. Nous concluons sur les avancées et limites de ce projet de règlement pour les systèmes d'IA à haut risque.*

1 Introduction

L'adoption en 2018 du Règlement Général de la Protection des Données (RGPD) a profondément modifié les comportements et pratiques des entreprises dans leurs gestions des données, messageries et sites internet. L'Europe poursuit sa démarche visant à harmoniser réglementations et innovations technologiques pour le respect des droits humains fondamentaux mais aussi la défense des intérêts commerciaux de l'Union. Cela a conduit à la publication d'un livre blanc sur l'Intelligence Artificielle : une approche européenne axée sur l'excellence et la confiance (2020) basé sur le guide pour une IA digne de confiance rédigé en 2019 par un groupe d'experts européens. Citons également la publication de propositions de règlements :

- *Digital Market Act* (2020) : recherche d'équité dans les relations commerciales et risques d'entraves à la concurrence à l'encontre des entreprises européennes ;
- *Digital Services Act* (2020) : sites de service intermédiaire, d'hébergement, de plateforme en ligne et autres réseaux sociaux ; comment contrôler les contenus illicites et risques des outils automatiques de modération ;
- *Data Governance Act* (2020) contractualisation des utilisations, réutilisations, des bases de données tant publiques que privées (fiducie des données) ;
- *Artificial Intelligence Act* (2021) : proposition de règlement établissant des règles harmonisées sur l'intelligence artificielle.

En plus du RGPD pour la protection des données à caractère personnel, l'adoption européenne à venir de ce dernier texte (*AI Act*) va profondément impacter les conditions de développements et d'exploitations des systèmes d'Intelligence Artificielle (systèmes d'IA). En conséquence, le présent document propose une réflexion sur la prise en compte méthodologique de ce projet de réglementation concernant plus spécifiquement les compétences en Statistique, Mathématiques, des équipes de développement d'un système d'IA, notamment ceux jugés à haut risque selon les critères européens.

La section 2 suivante extrait de ce texte les éléments clefs impactant les choix les plus méthodologiques puis la section 3 en commente les conséquences tout en proposant, décrivant, quels outils statistiques semblent les plus adaptés pour satisfaire aux futures obligations réglementaires. Enfin la section

4 déroule un cas d'usage numérique analogue à la recherche d'un score de crédit sur un jeu de données concret. Cet exemple permet d'illustrer une démarche de recherche d'un moins mauvais compromis à élaborer entre confidentialité, performance, explicabilité et sources de discrimination tout en soulevant les difficultés posées par la rédaction de la documentation qui devra accompagner tout système d'IA à haut risque. La conclusion synthétise les principales avancées de ce projet d'*AI Act* et en soulève, dans sa version actuelle, les quelques principales limites.

2 Proposition d'un cadre réglementaire européen de l'IA

2.1 Considérations éthiques & protection juridique

Les risques provoqués par les impacts dus aux erreurs des décisions, à l'opacité, aux biais algorithmiques mentionnés dans le considérant (71) du RGPD n'ont finalement pas ou peu été pris en compte dans cette 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 : rapport Villani (2018), déclaration de Montréal (2018)... Le livre blanc sur l'Intelligence Artificielle : une approche européenne axée sur l'excellence et la confiance (2020) annonce la volonté de remédier à ces insuffisances. Il 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.

Les lois actuelles en vigueur, dont en France la nouvelle rédaction de la Loi Informatique et Libertés (2019) qui intègre les dispositions du RGPD, ne sont pas contraignantes ou finalement inapplicables à des décisions complexes issues d'un algorithme d'apprentissage (Besse et al. 2019). 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 l'adoption d'une version même amendée de ce futur règlement européen sur l'IA.

En conséquence, nous proposons dans les sections suivantes, non pas une analyse exhaustive du projet de règlement, mais une sélection des questions concernant plus spécifiquement les compétences en Mathématiques et Statistique des équipes de développement d'un système d'IA. 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 (Health Center for Devices and Radiological, 2019) ou de l'autorisation de remboursement (Haute Autorité de Santé, 2020, annexe 5) en France des DSC (dispositifs de santé connectés) embarquant un algorithme d'apprentissage.

2.2 Proposition de règlement établissant des règles harmonisées sur l'IA

Structure du règlement

Le texte de la proposition (*Artificial Intelligence Act*) comporte 108 pages complété par 17 pages de 9 annexes. Seuls sont pris en compte ci-dessous une sélection des aspects les plus méthodologiques d'un système d'IA ayant des impacts sur l'usager final, personne physique, dans le cas de systèmes d'IA évalués à haut risque.

Les objectifs annoncés de ce règlement sont le développement et la diffusion, commercialisation, dans le marché européen de systèmes d'IA *sûrs, légaux et respectueux des droits fondamentaux*. Il s'agit aussi de *garantir la sécurité juridique pour faciliter les investissements et l'innovation*, par le *développement d'un marché unique pour les applications d'IA licites, sûres et dignes de confiance et empêcher la fragmentation du marché*. La proposition prévoit des règles qui se veulent *proportionnées et souples* pour faire face aux *risques spécifiques* liés aux systèmes d'IA en mettant l'accent sur les *risques inacceptables et les risques élevés*. L'objectif de ce texte vise, en principe, la recherche d'un meilleur équilibre entre bénéfices attendus et risques encourus, notamment en matière de droits fondamentaux. Plus concrètement il s'inscrit dans une logique de sécurité des produits et s'appuie, essentiellement (Meneceur 2021-b), sur la législation relative au marché intérieur : article 114 du Traité sur le Fonctionnement de l'Union Européenne.

Le règlement dont Castets Renard (2021) propose un court résumé et Meneceur (2021-a) une analyse débute par l'énoncé de 89 considérants et est composé de 85 articles structurés en 12 titres :

Titre I Champ d'application et définitions

Titre II Pratiques interdites de l'IA

Titre III Systèmes d'IA à haut risque

Titre IV Obligation de transparence pour les autres systèmes

Titre V Mesures en faveur de l'innovation

Titre VI Gouvernance

Titre VII Base de données européenne des systèmes à haut risque

Titre VIII Surveillance post-commercialisation

Titre IX Codes de conduite

Titre X Confidentialité et pénalités

Titre XI Délégation de pouvoir

Titre XII Provisions finales

Retenons les éléments concernant plus directement le statisticien ou scientifique des données impliqué dans la conception d'un système d'IA.

Extraits du règlement

Voici quelques uns des considérants qui sont, en principe, déclinés dans les articles :

> (13) Afin d'assurer un niveau cohérent et élevé de protection des intérêts publics en ce qui concerne la santé, la sécurité et les droits fondamentaux, des normes communes pour tous les systèmes d'IA à haut risque devraient être établies. Ces normes devraient être cohérentes avec la charte des droits fondamentaux de l'Union européenne (la charte) et devraient être non discriminatoires et conformes aux engagements commerciaux internationaux de l'Union.
>
> (44) Une haute qualité des données est essentielle ... afin de garantir que le système d'IA à haut risque fonctionne comme prévu et en toute sécurité et qu'il ne devienne pas une source de discrimination ... Des ensembles de données de formation, de validation et de test ... pertinents, représentatifs, exempts d'erreurs et complets compte tenu de la finalité du système ... propriétés statistiques appropriées, y compris en ce qui concerne les personnes ou les groupes de personnes sur lesquels le système d'IA à haut risque est destiné à être utilisé. Afin de protéger le droit d'autrui contre la discrimination qui pourrait résulter du biais dans les systèmes d'IA, les fournisseurs devraient être en mesure de traiter également des catégories spéciales de données à caractère personnel...
>
> (47) Pour remédier à l'opacité qui peut rendre certains systèmes d'IA incompréhensibles ou trop complexes pour les personnes physiques, un certain degré de transparence devrait être requis pour les systèmes d'IA à haut risque. Les utilisateurs doivent être capables d'interpréter la sortie du système et de l'utiliser de manière appropriée. Les systèmes d'IA à haut risque devraient donc

être accompagnés d'une documentation et d'instructions d'utilisation pertinentes et inclure des informations concises et claires, y compris en ce qui concerne les risques potentiels pour les droits fondamentaux et la discrimination, le cas échéant.

(49) Les systèmes d'IA à haut risque doivent fonctionner de manière cohérente tout au long de leur cycle de vie et répondre à un niveau approprié de précision, de robustesse et de cybersécurité conformément à l'état de la technique généralement reconnu. Le niveau d'exactitude et les mesures d'exactitude doivent être communiqués aux utilisateurs.

Nous trouvons dans ces considérants la demande de normes internationales indispensables, la priorité au respect des droits fondamentaux dont la non-discrimination, la nécessaire représentativité statistique des données d'entraînement, la nécessité de documentations exhaustives notamment sur les performances d'un système d'IA, les possibilités d'interprétation de ses sorties ou décisions en découlant, l'obligation de journalisation ou archivage des décisions et données afférentes.

Ces principes sont plus ou moins gravés dans la séquence des 85 articles dont voici un extrait.

L'**Article 3** Titre I précise les principales définitions :

(1) Système d'IA logiciel développé avec une ou plus des techniques listées dans l'annexe I.

(2) Fournisseur qui rend accessible sur le marché un système d'IA à titre gratuit ou non.

(29) Données d'apprentissage pour ajuster les paramètres entraînables d'un algorithme.

(30) Données de validation pour optimiser les paramètres non entraînables et éviter le sur-ajustement.

(31) Données de test pour une évaluation indépendante des performances d'un système d'IA avant commercialisation ou mise en service.

La définition adoptée de l'IA est pragmatique et très flexible en se basant sur la liste exhaustive d'algorithmes de l'annexe III : algorithmes procéduraux à base de règles logiques, systèmes experts et évidemment l'ensemble de l'apprentissage automatique : supervisé (statistique) ou non, par renforcement. Elle peut être facilement adaptée en fonction des évolutions technologiques. Ces définitions reconnaissent la place prépondérante de l'apprentissage statistique donc des données accessibles.

L'**Article 5** Titre II liste les applications prohibées de l'IA. Cela concerne les objectifs de manipulation : techniques subliminales, atteintes aux personnes vulnérables, les scores sociaux, l'identification biométrique en temps réel, sauf exception dictée par des obligations de sécurité publique.

L'**Article 6** Titre III annonce la liste dans l'annexe III des applications de systèmes d'IA classées à haut risque et donc visées par les articles suivants du titre III. Cela concerne en priorité les systèmes d'IA impactant directement ou non des personnes physiques : l'identification, la gestion de trafic et de ressources énergétiques, l'éducation, l'emploi, l'accès aux services publics ou non, les forces de l'ordre, la justice, le droit d'asile et le contrôle aux frontières, l'administration, la justice, les processus démocratiques.

Les articles 5 et 6 adoptent également le principe de définitions pragmatiques en listant explicitement les applications prohibées et celles à haut risque de l'IA. La liste à haut risque (annexe III) est facilement adaptable en fonction des évolutions technologiques.

L'**Article 9** Titre III impose l'implémentation d'un système de gestion du risque pour toute la durée de vie d'un système d'IA à haut risque. Il s'agit d'identifier ces risques, d'adopter des mesures adaptée de gestion, élimination ou au moins atténuation. Un système d'IA doit être testé afin d'identifier les meilleures mesures de risque.

La production de normes et standards par les autorités compétentes (*e.g.* comment mesurer des biais possiblement discriminatoires) et comme cela est requis dans le considérant 13 conduirait une mise en œuvre plus explicite de cet article.

> L'**Article 10** Titre III concerne précisément la gouvernance des données notamment, 2. d'apprentissage, validation et test qui doivent être soumises à une évaluation a priori (disponibilité, quantité, pertinence), une préparation précise et documentée (annotation, étiquetage, nettoyage, enrichissement, agrégation), (f) un examen en vue de possibles biais. 3. Les ensembles de données d'entraînement, validation et test doivent être pertinents, représentatifs, exempts d'erreurs et complets. Ils doivent avoir les propriétés statistiques appropriées, y compris, le cas échéant, en ce qui concerne les personnes ou groupes de personnes sur lesquels le système d'IA à haut risque est destiné à être utilisé. 5. Dans la mesure où cela est strictement nécessaire aux fins d'assurer la surveillance, la détection et la correction des biais ... les fournisseurs de ces systèmes peuvent traiter des catégories spéciales de données à caractère personnel ... sous réserve des garanties appropriées pour les droits et libertés fondamentaux des personnes physiques ... de protection de la vie privée de pointe, telles que la pseudonymisation ou le cryptage, lorsque l'anonymisation peut affecter de manière significative le but poursuivi.

L'article 10 est fondamental, il insiste sur l'importance d'une exploration statistique préalable fouillée des données avant de lancer les procédures devenues automatiques d'apprentissage et optimisation. Il autorise, sous réserve de précautions avancées pour la confidentialité, la constitution de bases de données personnelles sensibles permettant par exemple des statistiques ethniques, afin de pouvoir traquer directement des biais potentiels.

> L'**Article 11** Titre III impose la rédaction d'une documentation technique avant la mise en exploitation d'un système d'IA à haut risque puis sa mise à jour. Celle-ci doit démontrer que le système d'IA est conforme aux exigences ; elle fournit aux autorités nationales compétentes toutes les informations nécessaires pour évaluer sa conformité et contient au minimum les éléments indiqués dans l'annexe IV qui peut être complétée à la lumière des progrès techniques.

Cet article est essentiel pour ouvrir la possibilité d'audit *ex-ante* d'un système d'IA à haut risque. C'est au concepteur de montrer qu'il a mis en œuvre ce qu'il était techniquement possible en matière de sécurité, qualité, explicabilité, non discrimination, pour atteindre les objectifs attendus.

> L'**Article 12** Titre III impose un archivage du journal garantissant la traçabilité du fonctionnement d'un système d'IA. Le contenu du journal doit être approprié aux objectifs et fonctionnel tout au long de son cycle de vie ; une liste *a minima* des capacités est décrite.

Cette obligation est nouvelle par rapport aux textes européens précédents. Elle est indispensable pour assurer le suivi des mesures de performances, de risques et donc pour être capable de détecter des failles nécessitant des mises à jour voire un ré-entraînement du système ou même son arrêt. Les conditions d'archivage sont précisées dans l'article 61 Titre VIII (*post-market monitoring*).

> L'**Article 13** Titre III concerne la transparence et l'information des utilisateurs. 1. Les systèmes d'IA à haut risque doivent être conçus et développés de manière à garantir que leur fonctionnement est suffisamment transparent pour permettre aux utilisateurs d'interpréter les résultats du système et de l'utiliser de manière appropriée. 2. Ils doivent être accompagnés d'instructions d'utilisation comprenant des informations concises, complètes, correctes et claires, pertinentes, accessibles et compréhensibles pour les utilisateurs : 3. (b), (ii) le niveau de précision, de robustesse et de cybersécurité (iii) les conditions d'utilisation abusive raisonnablement prévisible, pouvant entraîner des risques pour la santé et la sécurité ou les droits fondamentaux ; (iv) les performances concernant les personnes ou groupes de personnes sur lesquels le système est destiné à être utilisé.

En résumé, un utilisateur devrait pouvoir interpréter les sorties, et doit être clairement informé des performances, éventuellement en fonction des groupes concernés, ainsi que des risques notamment de biais et donc de discrimination. Il s'agit ici d'un point sensible directement dépendant de la complexité des systèmes d'IA à base d'algorithmes sophistiqués donc opaques d'apprentissage statistique. L'absence de normes pour mesurer un biais discriminatoire et le manque de recul sur les recherches en cours en matière d'explicabilité d'une décision algorithmique sont tout à fait préjudiciables à la bonne application de cet article.

Ceci est complété par :

l'**Article 14** Titre III qui impose une surveillance par des personnes physiques pendant la période d'utilisation. La surveillance humaine vise à prévenir ou à minimiser les risques pour la santé, la sécurité ou les droits fondamentaux. Elle doit permettre d'interpréter correctement les résultats du système d'IA à haut risque, en tenant compte notamment des caractéristiques du système et des outils et méthodes d'interprétation disponibles.

L'**Article 15** Titre III Précision, robustesse et cybersécurité. Les niveaux de précision et les mesures de précision pertinentes des systèmes d'IA à haut risque doivent être déclarés dans les instructions d'utilisation jointes. Les systèmes d'IA à haut risque doivent être résilients en ce qui concerne les erreurs, les défauts ou les incohérences qui peuvent survenir dans le système ou l'environnement dans lequel le système fonctionne, en particulier en raison de leur interaction avec des personnes physiques ou d'autres systèmes. La robustesse des systèmes d'IA à haut risque peut être obtenue grâce à des solutions de redondance technique. Les systèmes d'IA à haut risque qui continuent d'apprendre après avoir été mis sur le marché ou mis en service doivent être développés de manière à garantir que les sorties éventuellement biaisées en raison des sorties utilisées comme intrants pour les opérations futures ("boucles de rétroaction") sont dûment traités avec des mesures d'atténuation appropriées. Ils doivent se montrer résilients aux attaques de leur vulnérabilité : falsification des données d'entraînement, exemple contradictoire, faille du modèle.

Cet article comble une lacune importante par l'obligation de déclaration des performances (précisions, robustesse, résilience) d'un système d'IA à haut risque. Il fait également allusion aux algorithmes d'apprentissage par renforcement soumis à des risques spécifiques : dérives potentielles (biais) et attaques malveillantes (cybersécurité).

Les articles suivants du Titre III notifient des obligations sans apporter de précision techniques ou méthodologiques : obligations faites au fournisseur (art. 16), obligation de mise en place d'un système de gestion de la qualité (art. 17), notamment de toute la procédure de gestion des données de la collecte initiale à leurs mises à jour en exploitation, ainsi que de la maintenance post-commercialisation ; obligation de documentation technique (art. 18), d'évaluation de la conformité (art. 19), obligation des utilisateurs (art. 29)...

Meneceur (2021-a) résume le processus de certification des principaux articles suivants :

Les États membres sont, par ailleurs, invités à désigner une autorité notifiante comme responsable du suivi des procédures relatives aux systèmes à haut risque et un organisme notifié (art. 30 à 39), tout à fait classique des mécanismes de certification déjà en œuvre. Un label "CE" sera délivré aux systèmes conformes (art. 49).

Le fournisseur devra suivre soit la procédure d'évaluation de la conformité sur la base du contrôle interne visée à l'annexe VI, soit la procédure d'évaluation de la conformité fondée sur l'évaluation du système de gestion de la qualité et l'évaluation de la documentation technique, avec l'intervention d'un organisme notifié, visée à l'annexe VII (art. 43). L'initiative de la mise en conformité repose, en toute hypothèse, sur le fournisseur. Une base de données enregistrera les systèmes autonomes d'IA à haut risque (art. 60). Des sanctions pourront être prononcées en cas de manquement, entre 2, 4 ou 6% du chiffre d'affaires annuel selon les situations (art. 71).

Ce processus de certification CE est essentiel pour les systèmes d'IA à haut risque, il repose sur un audit *ex-ante* requièrant, dans le cas d'une évaluation externe, des compétences très élaborées de la part des autorités qui en portent la responsabilité afin d'être à même de pouvoir déceler des manquements intentionnels ou non.

Commentaires

L'analyse de ces quelques articles amène des commentaires ou questions, notamment sous le prisme d'une approche mathématique ou statistique.

Projet Le projet d'*AI Act* entre dans un long processus (3 ou 4 ans comme le RGPD ?) de maturation avant une adoption européenne puis par chacun des parlements nationaux. Les amendements à venir devront être successivement pris en considération pour en analyser lers conséquences

en espérant que des réponses, précisions, corrections, seront apportées aux points ci-dessous. Néanmoins et compte tenu des temps et coûts de conception d'un système d'IA, il est important d'anticiper dès maintenant l'adoption de ce cadre réglementaire.

Exigences essentielles À la suite du guide des experts, le livre blanc appelle à satisfaire sept *exigences essentielles* dont celles de non discrimination et équité, bien être sociétal et environnemental.

Environnement la prise en compte de l'impact environnemental reste anecdotique, simplement évoquées dans les considérants (28) et (81), puis l'article 69 (*codes of conducts*) 2. sans aucune obligation formelle de calculer une balance bénéfices / risques (environnementaux ou autres) d'un système d'IA. Ainsi, l'obligation de l'archivage des données de fonctionnement d'un système d'IA génère un coût environnemental qui mériterait d'être pris en compte dans les risques afférents à son déploiement au regard de son utilité.

Équité La demande exprimée qu'un système d'IA satisfasse au respect des droits fondamentaux en référence à la Charte de l'UE, notamment celui de non-discrimination, est très présente dans le livre blanc (cité 16 fois), comme dans les considérants (15, 17, 28, 39) de la proposition de règlement. En revanche, ce principe n'apparaît plus explicitement dans les articles ; est-ce sa présence dans des textes de plus haut niveau comme la Charte des droits fondamentaux de l'UE qui n'a pas justifié ici une répétition mais sans ajouter de précision sur une façon de "mesurer" une discrimination ?

Normes Le considérant (13) appelle en effet à la définition de normes internationales notamment à propos des droits fondamentaux. En l'absence d'une définition juridique de l'équité d'un algorithme, celle-ci est définie en creux par l'*absence de discrimination* interdite explicitement. Le souci est que la littérature regorge de dizaines de définitions de biais statistiques pouvant être à l'origine de sources de discrimination ; lesquels considérer en priorité ? Espérons que les autorités compétentes (AFNOR, IEEE ?) se prononcent rapidement afin de préciser ce cadre indispensable. L'obligation de détailler les performances (précision) par groupe ou sous groupe d'un système d'IA (art. 13, 3., (b), iv) permet de prendre en compte ce type de biais, donc de discrimination spécifique. En revanche la détection d'un biais systémique ou de société, simplement évoquée dans l'analyse préalable des données (art. 10, 2. (f)) requiert d'urgence une définition normative pour être actionnable ou opposable. Des indicateurs statistiques de biais devenus relativement consensuels dans la communauté académique sont proposés dans la section suivante. Il est regrettable qu'aucune indication, recommandation, contrainte, ne vienne ensuite préciser ce qui pourrait ou devrait être fait pour atténuer ou éliminer un biais systémique détecté dans les données. Cet aspect est précisément détaillé dans l'exemple numérique de la section suivante.

Utilisateur & Usager Le règlement traite en priorité les considérations commerciales dont de la chaîne de responsabilité de l'acquisition des données à l'utilisateur du système d'IA. Tout système doit satisfaire aux exigences de performance annoncées, selon un principe de sécurité des produits ou responsabilité du fait des produits défectueux. En revanche, l'usager final, les dommages auxquels il peut être confronté, ne sont pas du tout pris en compte. L'obligation d'information (art. 13) est ainsi au profit de l'utilisateur et pas à celui de l'usager, personne physique impactée, qui ne semble donc protégé à ce jour que par les seules obligations de l'article 22 du RGPD. Il est informé de l'usage d'un système d'IA le concernant, il peut en contester la décision auprès de l'utilisateur humain mais l'explication de la décision, des risques encourus, est soumise aux compétences et à la déontologie professionnelle voire le cade juridique spécifique (*e.g.* code de santé public) de l'utilisateur : médecin pour un patient, conseiller financier pour un client, juriste pour un justiciable, responsable des ressources humaines pour un candidat...

Données le règlement reconnaît le rôle prépondérant des algorithmes d'apprentissage automatique et donc de la nécessité absolue (considérant 44) de qualité et pertinence des données conduisant à leur entraînement. L'article 10 impose en conséquence des compétences en Statistique pour conduire les études préalables à l'entraînement d'un algorithme. Nous assistons à un renversement de tendance, un retour de balancier, du tout automatique à une approche raisonnée sous responsabilité humaine de cette phase d'analyse des données longue et coûteuse mais classique du métier de statisticien.

Responsabilité De façon générale, l'objectif essentiel n'est plus la per-

formance absolue comme dans les concours de type *Kaggle* et conduisant à des empilements inextricables d'algorithmes opaques mais de satisfaire à un ensemble de contraintes dont celle de transparence en identifiant chaque responsabilité aux différents niveaux d'élaboration, diffusion et exploitation d'un système d'IA.

Documentation Tous les choix opérés lors de la conception d'un système d'IA : ensembles de données, algorithmes, procédures d'apprentissage et de tests, optimisations des paramètres, compromis entre confidentialité, performances, interprétabilité, biais... doivent (art. 11 et annexe IV) être explicitement documentés en vue d'un audit *ex-ante*. C'est un renversement de la charge de preuve sous la responsabilité première du concepteur qui doit montrer avoir mis ce qui était techniquement possible en œuvre pour satisfaire aux obligations légales de sécurité, transparence, performances et non discrimination.

Autorité de notification (Chapitre 4 Titre III) Chaque pays va se doter ou désigner (art. 30) un organisme de contrôle chargé entre autres de l'audit *ex-ante* d'un système d'IA à haut risque avant son déploiement qu'il soit commercialisé ou non.

Archivage & confidentialité Le règlement cible donc, en première lecture, les obligations commerciales du fournisseur plutôt que celles étiques ou déontologiques envers l'usager. Néanmoins le règlement apporte la possibilité de prendre en compte des données sensibles (art. 10, 5.), les obligations d'archivage des décisions (art. 12), de suivi des performances selon les groupes (art. 13), une surveillance humaine (art. 14) pendant toute la période d'utilisation et de correction rétroactive des biais (art. 15). Cette obligation d'archivage et surveillance du fonctionnement notamment à destination des groupes sensibles oblige implicitement à l'acquisition, en toute sécurité (cryptage, anonymisation, pseudonymisation...), de données confidentielles (*e.g.* origine ethnique). Cela ne rend-il pas indispensable, selon le domaine d'application, la mise en place d'un protocole explicite de consentement libre et éclairé, d'un engagement éthique, entre l'utilisateur et l'usager, protégé par le RGPD. Comment sont évalués les risques encourus d'un usager ou groupe d'usager par le recueil et l'exploitation de leurs données face aux bénéfices attendus pour eux mêmes ou l'intérêt public ?

3 Éléments de réponse méthodologique

Dans l'attente d'éléments de réponses politiques ou techniques (normes) aux questions ci-dessus, il est néanmoins indispensable d'anticiper pour répondre techniquement à certaines contraintes ou obligations faites aux système d'IA désignés à haut risque. Cet article laisse volontairement de côté certaines classes d'algorithmes mentionnées dans l'annexe I et donc intégrées dans la définition de l'IA de l'article 3.

Algorithmes déterministes ou procéduraux Il s'agit 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, désavantages ou risques de discrimination indirecte, malgré une apparente neutralité. Le Défenseur des Droits (2020) est très attentif en France à l'analyse et détection de ces risques. Celle-ci relèvent 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 analyse experte *ex post* ne sera pas en mesure d'évaluer l'étendue des risques indirects. Aussi, un algorithme déterministe complexe peut être analysé avec les mêmes outils statistiques que ceux adaptés à un algorithme d'apprentissage automatique.

Systèmes experts Un système expert est l'association d'une base de règles logiques ou base de connaissances construites par des experts du domaine concerné, d'un moteur d'inférence et d'une base de faits observés pour une exécution en cours. Le moteur d'inférence recherche la séquence de règles logiquement applicables à partir des faits observés de la base qui s'incrémente comme conséquences du déclenchements des règles. Le processus itère jusqu'à l'obtention ou non d'une décision recherchée et expliquée par la séquence de règles y conduisant. Très développée dans les années 70, la recherche a marqué le pas face à un problème dit NP-complet c'est-à-dire de complexité algorithmique exponentielle en la taille de la base de connaissance (nombre de règles). Supplantée par la ré-émergence des réseaux de neurones (années 80) puis plus largement par l'apprentissage automatique, la recherche dans ce domaine dit d'IA symbolique est restée active. Elle connaît un renouveau motivé par les capacités d'explicabilité des systèmes experts.

Nous insistons donc tout particulièrement sur les systèmes d'IA basés sur des algorithmes d'apprentissage supervisé ou statistique ou IA empirique. Ce sont très majoritairement les plus répandus au sein de ceux désignés à haut risque (art. 6) car susceptibles d'impacter directement ou non des personnes physiques.

3.1 Les données

Tout système d'IA basé sur un algorithme d'apprentissage statistique nécessite la mise en place d'une base de données d'entraînement fiable et représentative du domaine d'application visé. Ce travail d'exploration statistique, généralement long et fastidieux d'acquisition, vérification, analyse, préparation, nettoyage, enrichissement, archivage sécurisé des données, est essentiel à l'élaboration d'un système d'IA performant, robuste, résilient et dont les biais potentiels sont sous contrôle. Construire de nouvelles caractéristiques (*features*) adaptées à l'objectif, traquer et gérer éventuellement par imputation des données manquantes, identifier les anomalies ou valeurs atypiques (*outliers*) sources de défaillances, les sources de biais : classes ou groupes sous représentés, biais systémiques, nécessitent compétences et expériences avancées en Statistique. Elles sont indispensables pour répondre aux attentes de l'article 10 ainsi qu'aux besoins de la documentation (annexe IV) imposée par l'article 11.

3.2 Qualité, précision et robustesse

Les mesures de *précision* de la prévision d'un système d'IA sont bien connues et maîtrisées, parties intégrante du processus d'apprentissage. Néanmoins parmi un large éventail des possibles, le choix, précisément justifié, doit être adapté au domaine, au type de problème traité, aux risques spécifiques encourus. Citons par exemple les cas de la :

Régression ou modélisation et prévision d'une variable cible Y quantitative. Elle est généralement basée sur l'optimisation d'une mesure quadratique (norme L_2) pouvant intégrer, à l'étape d'entraînement, différents types de pénalisation dont celle de parcimonie (Ridge, Lasso) afin de contrôler la complexité de l'algorithme et éviter les phénomènes de sur-apprentissage. D'autres types de fonction objectif basée sur une perte en norme L_1 ou valeur absolue, moins sensible à la présence de valeurs atypiques (*outliers*) que la norme quadratique, permet des solutions plus robustes car tolérantes à des observations atypiques.

Classification ou modélisation, prévision d'une variable Y qualitative. Le choix d'une mesure d'erreur doit être opéré parmi de très nombreuses possibilités : taux d'erreur, AUC (*area under the ROC Curve* pour une variable Y binaire), score F_β, risque bayésien, entropie... avec la difficile prise en compte des situations de classes déséquilibrées qui oriente le choix du type de mesure et nécessite des précautions spécifiques dans l'équilibrage de la base d'apprentissage ou les pondérations de la fonction objectif en prenant en compte une matrice de coûts de mauvais classement éventuellement asymétrique.

Une démarche très rigoureuse doit conduire à l'évaluation de la précision et donc des performances d'un système d'IA basé sur un algorithme d'apprentissage. Comme énoncé dans l'article 3, 31. ce sont des *données de test indépendantes* de celles d'apprentissage qui sont utilisées à cet effet. *Attention* néanmoins d'évaluer les performances sur des données telles qu'elles se présenteront réellement en exploitation, avec leurs défauts, et pas un simple sous-ensemble aléatoire de la base d'apprentissage comme c'est le cas en situation académique. En effet cet ensemble de données peut bénéficier d'une homogénéité d'acquisition (*e.g.* même technologie) et de prétraitements qui peut faire défaut à de réelles données d'entrée à venir en exploitation. Cela demande donc une extrême rigueur dans la constitution d'un échantillon test pour éviter ces pièges et également une surveillance toute la durée de vie du système d'IA afin d'en détecter de possibles dérives ou dysfonctionnements (art. 12 et 15).

L'évaluation de la *robutesse* est liée 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 atypiques, étrangères à la base d'apprentissage.

La *résilience* d'un système d'IA est essentielle pour les dispositifs critiques (dispositifs de santé connecté, aide au pilotage). Cela 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, alerte et arrêt du système.

Notons que le Laboratoire Nationale de Métrologie et d'Essai a pris les devants et propose des procédures, formations, pour l'évaluation des systèmes d'IA, plus précisément ceux basés sur le traitement de la parole ou intégrant de la robotique.

3.3 Explicabilité

Il est bien trop tôt pour tenter un résumé opérationnel de ce thème et fournir des indications claires sur la démarche à adopter pour satisfaire aux exigences réglementaires. Il faut pour cela attendre que la recherche ait progressé et qu'une sélection "naturelle" en extrait les procédures les plus pertinentes parmi une grande quantité de solutions proposées ; un article de revue sur ce sujet (Barredo Arrieta et al. 2020) liste plus de 400 références. 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.

Bien distinguer les niveaux d'explication : utilisateur ou usager même si ce dernier n'est pas directement concerné par le projet de règlement.

L'explication peut alors s'appliquer :

1. au 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 et d'interactions prises en compte reste raisonnable.
   - dans le cas d'un algorithme complexe opaque :
     — chercher une approximation explicable par un modèle linéaire, arbre, règles de décision déterministes ;
     — sinon, fournir des indications par l'identification des variables *importantes* par randomisation des valeurs de chaque variable (*mean decrease accuracy* Breiman, 2001) ou stress de l'algorithme (Bachoc et al. 2020).
2. à une décision spécifique pour :
   - le concepteur : identifier la cause d'une erreur, y remédier par exemple en complétant la base d'apprentissage d'un groupe sous-représenté avant de ré-entraîner l'algorithme ;
   - la personne concernée : client, patient, justiciable :
     — à l'aide d'un modèle interprétable : linéaire, arbre de décision,
     — sinon par une approximation locale : LIME, contre-exemple, règles,...
     — et sinon se limiter à l'explication *a minima* du risque d'erreur encouru.

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 certes 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, leurs interactions, des effets non linéaires voire des boucles de contre-réaction, est nécessairement modélisé par un algorithme complexe afin d'éviter des simplifications abusives pouvant gravement nuire aux performances. C'est tout d'abord le réel qui s'avère complexe à expliquer.

3.4 Biais & discrimination

Très prolixe, le monde académique a proposé quelques dizaines d'indicateurs (*e.g.* Zliobait 2017) afin d'évaluer des biais potentiels sources de discrimination bien que beaucoup de ces indicateurs s'avèrent très corrélés ou redondants (Friedler et al. 2019). Alors qu'il n'existe pas de définition juridique de l'équité et dans l'attente de normes à venir et éventuellement d'obligations légales plus contraignantes, il est nécessaire d'anticiper des choix. Empiriquement et après avoir consulter une vaste littérature sur l'IA éthique ou plutôt sur les risques identifiés de discrimination algorithmique, trois niveaux de biais statistique peuvent et doivent être pris en compte en priorité. De plus, formellement, la stricte équité s'exprime par des propriétés d'indépendance en probabilité mais cette approche théorique n'est pas concrètement praticable pour

détecter, mesurer, atténuer des risques de biais. En conséquence, sont considérés trois types de rapports de probabilités (égaux à 1 en cas d'indépendance stricte) dont Besse et al. (2021) proposent des estimations par intervalle de confiance afin d'en contrôler la précision.

3.4.1 Parité statistique et effet disproportionné

Le premier niveau de risque de discrimination algorithmique s'illustre simplement : si un algorithme est entraîné sur des données biaisées, il reproduit très fidèlement ces biais systémiques, de société ou de population par lequel un groupe est historiquement (*e.g.* revenu des femmes) désavantagé ; plus grave, l'algorithme risque même de renforcer le biais en conduisant à des décisions explicitement discriminatoires. Il importe donc de pouvoir détecter, mesurer, atténuer voire éliminer ce type de biais. L'équité ou parité statistique (ou *demographic equality*) serait l'indépendance entre la ou les variables sensibles S (*e.g.* genre, origine ethnique) et la variable de prévision $\widehat{Y}$ de la décision. Historiquement, l'écart à l'indépendance pour mesurer ce type de biais est évalué aux USA dans les procédures d'embauche depuis 1971 par la notion d'effet disproportionné ou *disparate impact* (Barocas et Selbst, 2016). L'évaluation de l'effet disproportionné consiste à estimer le rapport de deux probabilités : probabilité d'une décision favorable ($\widehat{Y} = 1$) pour une personne du groupe sensible ($S = 0$) au sens de la loi sur la même probabilité pour une personne de l'autre groupe ($S = 1$) :

$$DI = \frac{\mathbb{P}(\widehat{Y} = 1|S = 0)}{\mathbb{P}(\widehat{Y} = 1|S = 1)}.$$

Cet indicateur est intégré au *Civil Rights act & Code of Federal Regulations (Title 29, Labor : Part 1607 Uniform guidelines on employee selection procedures)* depuis 1978 avec la règle dite des 4/5 ème ; si DI est inférieur à 0.8, l'entreprise doit en apporter les justifications économiques. Les logiciels commercialisés aux USA et proposant des algorithmes de pré-recrutement automatique anticipent ce risque juridique (Raghavan et al. 2019) en intégrant une procédure automatique d'atténuation du biais (*fair learning*). Il n'y a aucune obligation ni mention en France de cet indicateur statistique, seulement une incitation de la part du Défenseur des Droits et de la CNIL (2012) envers les services de ressources humaines des entreprises. Il leur est suggéré de tenir des statistiques ethniques, autorisées dans ce cas sous réserve de confidentialité, sous la forme de tables de contingence dont il serait facile d'en déduire des estimations d'effet disproportionné.

La mise en évidence d'un biais systémique est implicitement citée lors de l'étape d'analyse préliminaire des données (art. 10, 2., (f)) mais sans plus de précision sur la façon dont il doit être pris en compte alors que renforcer algorithmiquement ce biais serait ouvertement discriminatoire. De plus serait-il politiquement opportun d'introduire une part de discrimination positive afin d'atténuer la discrimination sociale ? C'est évoqué dans le travail des experts (Commission Européenne, 2019, ligne directrice 52) pour *améliorer le caractère équitable de la société* et techniquement l'objet d'une vaste littérature académique nommée apprentissage équitable (*fair learning*). Cette opportunité n'est pas reprise explicitement dans l'*AI Act* mais nous verrons dans l'exemple numérique ci-dessous qu'elle ne peut être exclue et peut même être pleinement justifiée en prenant en considération les autres types de biais ci-après.

3.4.2 Erreurs conditionnelles

Les taux d'erreur de prévision et donc les risques d'erreur de décisions sont-ils les mêmes pour chaque groupe (*overall error equality*) ? Autrement dit, l'erreur est-elle indépendante de la variable sensible ? Ceci peut se mesurer par l'estimation (intervalle de confiance) du rapport de probabilités (probabilité de se tromper pour le groupe sensible sur la probabilité de se tromper pour l'autre groupe) :

$$\frac{\mathbb{P}(\widehat{Y} \neq Y|S = 0)}{\mathbb{P}(\widehat{Y} \neq Y|S = 1)}.$$

Ainsi, si un groupe est sous-représenté dans la base d'apprentissage, il est très probable que les décisions le concernant soient moins fiables. C'est une des première critiques formulées à l'encontre des algorithmes de reconnaissance faciale et ce risque est également présent dans les applications en santé (Besse et al. 2020) ou en ressources humaines (De-Arteaga et al. 2019). L'identification, la prise en compte et la surveillance de ce risque sont présents (art. 13, 3., (b), ii et art. 15, 1. & 2.) dans le projet de règlement et doivent donc être explicitement détaillés dans la documentation (art. 11).

3.4.3 Rapports de cote conditionnels

Même si les deux critères précédents sont trouvés équitables, les erreurs peuvent être dissymétriques (plus de faux positifs, moins de faux négatifs) au détriment d'un groupe avec un impact d'autant plus discriminatoire que le taux d'erreur est important. Cet indicateur (comparaison des rapports de cote ou *odds ratio* d'indépendance conditionnelle nommé aussi *equalli odds*) est au cœur de la controverse concernant l'évaluation COMPAS du risque de récidive aux USA (Larson et al. 2016). Il est également présent dans l'exemple numérique ci-après. Cet indicateur double est également mesuré par l'estimation (intervalle de confiance) de deux rapports de probabilités (rapports des taux de faux positifs du groupe sensible sur le taux de faux positifs de l'autre groupe et rapport des taux de vrais positifs pour ces mêmes groupes) :

$$\frac{\mathbb{P}(\hat{Y}=1|Y=0,S=0)}{\mathbb{P}(\hat{Y}=1|Y=0,S=1)} \quad \text{et} \quad \frac{\mathbb{P}(\widehat{Y}=1|Y=1,S=0)}{\mathbb{P}(\widehat{Y}=1|Y=1,S=1)}.$$

L'évaluation de ce type de biais n'est pas explicitement mentionné dans le projet de règlement. Néanmoins il fait partie de la procédure classique d'évaluation des erreurs en classification à l'aide d'une matrice de confusion ou de courbes ROC par groupes. Les instances de normalisation et / ou nationales d'évaluation auront à se positionner.

Notons qu'il est d'autant plus difficile de faire abstraction du dernier type de biais que les trois sont interdépendants et même en interaction avec les autres risques : précision et explicabilité. Ceci est clairement mis en évidence dans l'exemple numérique suivant. Il y a donc une forme d'obligation déontologique ou de cohérence statistique à devoir appréhender ces différents niveaux d'analyse.

4 Exemple numérique

L'exemple jouet ou bac à sable de cette section permet d'illustrer concrètement toute la complexité des principes précédemment évoqués en soulignant leur interdépendance.

4.1 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 dans un tutoriel (calepin *Jupyter*) mais l'illustration est limitée à un résumé succinct de l'analyse de la discrimination selon le genre.

4.2 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 élémentaire (arbre binaire de décision) augmente le biais mais pas de façon statistiquement significative car les intervalles de confiance ne sont pas disjoints. Sa précision est meilleure que celle du mo-

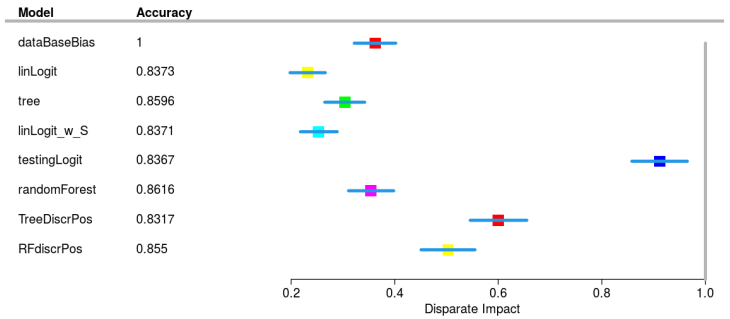

FIGURE 1 – *Précision de la prévision (accuracy) et effet disproportionné estimé par un intervalle de confiance sur un échantillon test (taille 9000) pour différents modèles ou algorithmes d'apprentissage.*

dèle de régression logistique mais, si l'objectif est une interprétation utile, il est nécessaire de réduire la complexité de l'arbre en pénalisant le nombre de feuilles. Dans ce cas la précision se dégrade pour atteinde celle de la régression logistique.

Un algorithme non linéaire plus sophistiqué (*random forest*) est très fidèle au biais des données avec un indicateur $(DI = 0,36)$ proche 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 ou un arbre binaire de décision de taille raisonnable. Enfin, les deux dernières lignes proposent 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 : $0,3$ pour les forêts aléatoires, $0,2$ pour un arbre binaire, 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*. L'arbre binaire utilisé est celui pénalisé (peu de feuilles) afin d'obtenir une interprétation facile au prix de la précision.

Les autres types de biais sont également à considérer. En 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 au désavantage des femmes. 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 $(0,38)$. La procédure élémentaire d'atténuation du biais en entraînant deux algorithmes, un pour chaque genre, et en adaptant le seuil de décision conduit à une augmentation attendue du taux d'erreur pour les femmes, qui se rapproche de celui des hommes, accompagnée corrélativement d'un taux de faux positifs plus élevés pour les femmes.

4.3 Discussion

Nous pouvons tirer quelques enseignements de cet exemple jouet 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.
- La suppression naïve de la variable sensible (genre) pour réduire le biais n'y change rien d'où l'importance (art. 10, 5.) d'autoriser la prise d'un risque contrôlé de confidentialité pour intégrer des données personnelles sensibles afin de pouvoir détecter des biais.
- Un algorithme 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 $(DI < 0,8)$. Seul un arbre binaire pénalisé permet de concilier accroissement peu important du biais et interprétabilité sans trop pénaliser la précision.
- Une procédure de *testing* (Rich 2014), qui consiste à envoyer des milliers de paires de CV comparables à l'exception de la modalité de la

variable sensible (*e.g.* origine ethnique) est largement utilisée pour mesurer des situations de discrimination à l'embauche. Elle est même la doctrine officielle promue par le Comité National de l'Information Statistique et commanditée par la DARES (Direction de l'Animation, des Etudes, de la Recherche et des Statistiques) du Ministère du travail. Elle est complètement inadaptée à la détection *ex-post* d'une discrimination algorithmique. Seule une analyse *ex-ante* rigoureuse d'une documentation loyale (art. 11) décrivant des données, la procédure d'apprentissage, les performances peut donc s'avérer convaincante sur les capacités non discriminatoires d'un algorithme.

- Sur cet exemple, l'introduction d'une dose de discrimination positive impacte les trois types de biais pour en réduire simultanément l'importance. C'est une façon de légitimer l'introduction d'une dose de discrimination positive qui réduit le désavantage fait aux femmes sans pour autant nuire aux hommes.
- Finalement dans cet exemple illustratif, un arbre pénalisé pour être suffisamment simple (nombre réduit de feuilles) et assorti d'une touche de discrimination positive autorise une explication des décisions avec un meilleur équilibre des biais au regard des risques de discrimination. Certes, dans le cas d'un score de crédit, cela aurait pour conséquence d'accroître le risque de la banque en réduisant la qualité de prévision et augmentant le taux de faux positifs pour les femmes mais lui fournirait des arguments tangibles de communication pour une image "éthique" : des décisions plus équitables et plus explicables sans trop nuire à la précision.

5 Conclusion

Comme le rappelle Meneceur (2021-b) dans une comparaison exhaustive des démarches institutionnelles, les très nombreuses approches éthiques visant à encadrer le développement et l'application des systèmes d'IA ne sont pas des réponses suffisantes et convaincantes pour développer la confiance des usagers. Ceci motive la démarche de la CE aboutissant à la publication de ce projet de règlement alors que le *Conseil de l'Europe envisage également un mélange d'instruments juridiques contraignants et non contraignants pour prévenir les violations des droits de l'homme et des atteintes à la démocratie et à l'État de droit* ; la nécessité de conformité se substitue à l'éthique.

L'analyse du projet de règlement européen montre des avancées significatives pour plus de transparence des systèmes d'IA :

- importance fondamentale des données et donc de leur analyse préalable fouillée et documentée,
- évaluation et documentation explicite des performances et donc des risques d'erreur ou de manquement : robustesse, résilience,
- documentation explicite sur les capacités d'interprétation d'un système, d'une décision, à la mesure des technologies et méthodes disponibles,
- prise en compte de certains types de biais : performances différentes selon des groupes et suivi des risques possibles de discrimination associés,
- enregistrement de l'activité pour une traçabilité du fonctionnement,
- contrôle humain approprié pour réduire et anticiper les risques,
- autorité nationale d'évaluation et de contrôle.

Néanmoins ce projet de règlement principalement motivé par une harmonisation des relations commerciales au sein de l'Union selon le principe de sécurité des produits ou de la responsabilité du fait des produits défectueux ne prend pas en compte des dommages possibles afférents aux usagers. Les conséquences ou objectifs de la démarche adoptée par la CE rejoignent d'ailleurs les exigences de la FTC (*Federal Trade Commission*) (Jillson, 2021) de loyauté et transparence vis-à-vis des performances d'un système d'IA commercialisé. Aussi certains droits fondamentaux, bien que retenus comme *exigence essentielle* dans le livre blanc se trouvent pour le moins négligés.

- Plus largement que les seules applications de l'IA, une prise en compte d'une forme de frugalité numérique afin de réduire les impacts environnementaux ne semblent pas, dans ce projet d'*AI Act*, une préoccupation majeure de la CE. Cela concerne la consommation énergétique pour le stockage massif et l'entraînement des algorithmes mais surtout la sur-exploitation des ressources minières (cuivre, lithium, terres rares...) afférente à la fabrication des équipements numériques.
- Il est certes conseillé de rechercher des biais potentiels dans les données (art. 10, 2., (f)) avec même la possibilité de prendre en compte des données personnelles sensibles (art.10, 5.) pour traquer des biais systémiques sources potentielles de discrimination. Néanmoins, l'absence de normes précises sur la façon de mesurer ces biais, de les at-

ténuer ou les supprimer dans les procédures d'entraînement laisse un vide potentiellement préjudiciable à l'usager. Alors qu'il est déjà fort complexe pour un usager d'apporter la preuve d'une discrimination, par exemple par *testing*, lors d'une décision humaine, l'exemple numérique ci-dessus montre que c'est mission impossible face à une décision algorithmique. Seule une procédure rigoureuse d'audit *ex-ante* de la procédure d'apprentissage et des dispositions mises en place pour gérer, atténuer les biais peut garantir une protection *a minima* des usagers finaux contre ce type de discrimination. Il importe donc de doter l'autorité désignée pour cet audit de compétences et moyens humains ainsi que d'outils techniques, normes et protocoles, appropriés.

L'exemple numérique jouet a également pour mérite de montrer clairement l'*interdépendance* de toutes les contraintes : confidentialité, qualité, explicabilité, équité (types de biais), que devrait satisfaire un système d'IA pour gagner la confiance des usagers. Il montre aussi que le problème ne se réduit pas à un simple objectif de minimisation d'un risque quantifiable pour l'obtention d'un meilleur compromis. C'est plutôt la recherche d'une moins mauvaise solution imbriquant des choix techniques, économiques, juridiques, politiques qu'il sera nécessaire de clairement expliciter dans la documentation rendue obligatoire par l'adoption à venir d'un *AI Act* qui serait, de toute façon et malgré les limites actuelles du projet de texte, une avancée notable pour plus de transparence.

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.
- Barredo Arrieta A., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R., Chatila R., Herrera F. (2020). Explainable Artificial Intelligence (XAI) : Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion*, Vol. 58, pp 82-115.
- Besse P., Besse Patin A., Castets Renard C. (2020). Implications juridiques et éthiques des algorithmes d'intelligence artificielle dans le domaine de la santé, *Statistique & Société*, 3, pp 21-53.
- Besse P., Castets-Renard C., Garivier A., Loubes J.-M. (2019). L'IA du Quotidien peut elle être Éthique ? Loyauté des Algorithmes d'Apprentissage Automatique, *Statistique et Société*, **6**-3.
- Besse P., del Barrio E., Gordaliza P., Loubes J-M., Risser L. (2021) A survey of bias in Machine Learning through the prism of Statistical Parity for the Adult Data Set, *The American Statistician*, to appear.
- Breiman L. (2001). Random forests, *Machine Learning* 45, 532.
- Castets-Renard C. (2021). Nouvelles règles et actions pour l'excellence et la confiance en l'IA, blog, consulté le 29/05/2021.
- Commission Européenne (2019). Lignes 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, *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp 120128.
- Défenseur des Droits, CNIL (2012). Mesurer pour progresser vers l'égalité des chances. Guide méthodologique à l'usage des acteurs de l'emploi.
- Défenseur des Droits (2020). Algorithmes : Prévenir l'automatisation des discriminations.
- Friedler S., Scheidegger C., Venkatasubramanian S., Choudhary S., Hamilton E., Roth D. (2019). Comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, p. 329-338.
- Haute Autorité de Santé (2020) Guide : LPPR Dépôt d'un dossier auprès de la Commission nationale d'évaluation des dispositifs médicaux et des technologies de santé.
- Health Center for Devices and Radiological (2019). Artificial Intelligence and Machine Learning in Software as a Medical Device, *FDA*.
- Jillson E. (2021). Aiming for truth, fairness, and equity in your company's use of AI, blog, consulté le 29/05/2021.
- Larson J., Mattu S., Kirchner L., Angwin J. (2016). How we analyzed the compas recidivism algorithm. ProPublica, en ligne consulté le

28/04/2020.

- Meneceur Y. (2021-a) Proposition de règlement de l'IA de la Commission européenne : entre le trop et le trop peu ?, blog consulté le 28/05/2021
- Meneceur Y. (2021-b). Analyse des principaux cadres supranationaux de régulation de l'intelligence artificielle : de l'éthique à la conformité, projet d'étude, Institut des Hautes Études sur la Justice (IHEJ), version d'étude du 27/05/2021.
- 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.
- Zliobaité I. (2017). Measuring discrimination in algorithmic decision making, *Data Min. Knowl. Disc.*, 31, p 106089.