IA : attention aux biais et problèmes éthiques

IA: comment valider que les modèles d'intelligence artificielle sont corrects d'un point de vue éthique et ne génèrent aucune forme de discrimination ?

9 min. de lecture · Voir l'original

On se retrouve aujourd'hui pour parler IA spécifiquement sur un sujet qui ne cesse de prendre de l'importance depuis plusieurs mois : comment **valider que les modèles d'intelligence artificielle sont corrects d'un point de vue éthique** et ne génèrent aucune forme de **discrimination** ?

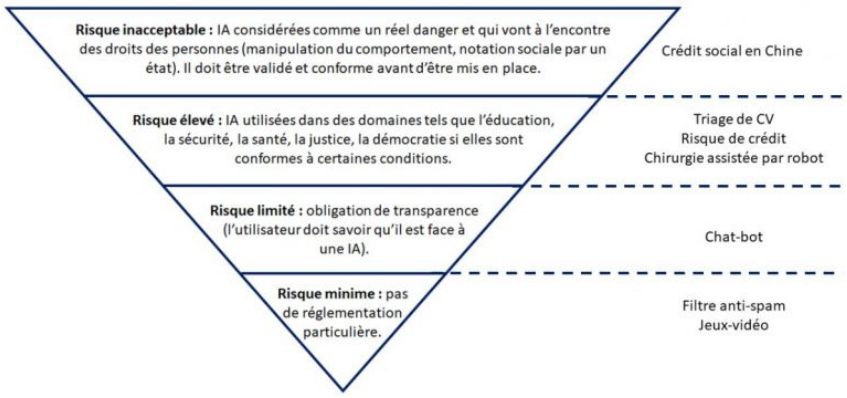
Ce sujet mérite en effet toute notre attention, car nous sommes amenés à proposer des solutions de plus en plus autonomes avec plus ou moins d'impact sur les utilisateurs finaux. Afin de contrôler le développement de ces modèles, la Commission européenne a d'ailleurs proposé des nouvelles normes de risques afin d'encadrer l'usage de ces technologies. En effet, à la suite de ces modèles avancés, les **prises de décisions automatiques peuvent avoir plus ou moins de risques sur la société**. Les solutions exploitant les intelligences artificielles devront désormais être classées selon leur risque ou dangerosité.

C'est pourquoi à AVISIA nous prenons les devants et travaillons activement sur ce point afin de vous garantir que nos modèles respectent un maximum de critères. Afin de bien comprendre les enjeux et les validations à mettre en place, il est important de maîtriser les risques possibles, à éviter lors de la construction de tels algorithmes. C'est ce que nous appelons généralement les biais.

Ces biais, qu'ils soient connus ou inconnus par les personnes construisant ces solutions, peuvent en effet être à l'origine de nombreuses dérives. Bien souvent ces biais ne sont relevés qu'après la mise à disposition au grand public et malheureusement certaines Intelligences peuvent conduire à de véritables problèmes éthiques comme l'exclusion de minorité dans l'usage de ces solutions.

CLASSIFICATION DES RISQUES

L'Union Européenne a décidé depuis le 21 avril 2021 de répertorier les systèmes d'IA de façon à leur imposer plus ou moins de règles en fonction de leur dangerosité. Les objectifs de ces mesures sont d'anticiper les possibles dérives de l'IA et veiller à ne pas porter atteinte à la vie privée de ses utilisateurs. Ils sont classés en **quatre catégories : de risque minime à risque inacceptable** en passant par risque limité et risque élevé. Voici ci-dessous les caractéristiques de chaque risque ainsi que quelques exemples associés :



LES PRINCIPAUX BIAIS À ÉVITER

Lors de la construction d'un modèle d'IA, il est important de bien prendre en compte les nombreux biais pouvant provenir des données, de l'algorithme ou même l'humain.

Biais relatifs à la collecte des données

Les biais relatifs aux données sont principalement représentés par des données de mauvaise qualité ou non pertinentes à une étude. Nous distinguons ici deux biais :

- Les **biais de sélection** consistent à choisir des données qui ne sont pas pertinentes à l'étude. Ils se décomposent en deux types :
    - Le **biais d'échantillonnage** consiste à ne pas sélectionner des données de façon aléatoire.

Exemple : nous souhaitons prédire les ventes d'un produit à travers une campagne d'emails. Nous choisissons les 1000 premières réponses. Nous sommes face à un biais d'échantillonnage : pour le palier il faudrait tirer 1000 emails aléatoirement parmi l'ensemble des réponses. En effet, il est possible que les premiers répondants soient plus enthousiastes à répondre que des acheteurs ne souhaitant pas à tout prix le produit.

- Le **biais de participation**, quant à lui, se produit lorsque les données ne sont pas assez représentatives à cause d'écarts de participation lors de la collecte.

Exemple : reprenons l'idée de prédire les ventes d'un produit. Des personnes ne souhaitant pas acheter le produit seront moins susceptibles de répondre au mail, nous serions donc face à un biais de participation.

- Les **biais de fréquence** se caractérisent par des effectifs sur ou sous-représentés par rapport à la vie réelle.

Exemple 1 : Facebook a utilisé un dataset contenant des métiers (comme variable cible) et des caractéristiques des personnes qui l'exerçaient pour entraîner son algorithme. Ce dernier conseillait des offres d'emploi différentes selon le sexe de l'utilisateur. La répartition démographique dans les données n'était pas équilibrée. Par exemple, pour un poste de comptable, la répartition (dans le jeu de données) était de 70 % pour les femmes et 30 % pour les hommes, il aurait eu tendance à conseiller les offres de comptable aux femmes. Cette recommandation irait à l'encontre de la vie réelle où, admettons, le poste de comptable est homogène.

- Les **biais survenant de données non mesurables** sont caractérisés par des événements ne

pouvant être évalués comme la relation entre des personnes, les sentiments, l'implication d'un individu dans une entreprise.

Exemple : L'état de Washington a décidé de noter les enseignants pour équilibrer le niveau des élèves dans les différentes écoles : l'objectif est d'attribuer les meilleurs enseignants dans les écoles les plus faibles. Une enseignante très appréciée de sa direction et des parents d'élèves a été licenciée. Elle enseignait dans une classe avec beaucoup d'élèves en difficulté, les résultats étant moins bons qu'ailleurs, le modèle lui a attribué une note médiocre. Ici, des dimensions non-mesurables comme la difficulté d'un élève, la relation enseignant/élèves ; enseignants/parents d'élèves ne sont pas prises en compte dans l'algorithme. 205 autres enseignants de l'Etat de Washington D.C ont également été licenciés.

Biais relatifs à la qualification des données

- Les **biais d'appartenance** consistent à préférer un groupe auquel on appartient.

Exemple : une entreprise souhaite mettre en place un modèle pour trier les CV. Les développeurs pensent que les candidats ayant fréquenté leur école sont meilleurs que les autres et les favorisent dans la construction de l'algorithme : il s'agit d'un exemple d'un biais d'appartenance.

- Le **biais implicite** consiste à généraliser notre expérience personnelle ou encore notre culture.

Exemple : Nous souhaitons construire un algorithme de reconnaissance de gestes : le modèle cherche à reconnaître si un individu dit oui avec sa tête en hochant celle-ci de haut en bas. Or, ce signe de la tête peut être interprété de manière différente selon les pays.

Biais relatifs à l'apprentissage de l'algorithme

Les biais relatifs aux algorithmes se dessinent de deux façons différentes :

- Le **sur-apprentissage** : phénomène statistique survenant lorsque le modèle s'adapte trop au jeu

de données d'apprentissage. Le modèle devient alors inexploitable lorsqu'il s'agit de prédire des nouvelles données : on dit que le modèle n'est pas généralisable.
- Le **sous-apprentissage** : à l'inverse, le sous-apprentissage peut survenir si le modèle n'a pas suffisamment appris lors de la phase d'entraînement ou si les variables d'entrées n'étaient pas suffisantes ou pertinentes pour expliquer et prédire la cible.

Exemple : l'application de reconnaissance faciale qui ne reconnaissait pas les personnes d'origine asiatique car elle avait très peu appris sur ce type de visages.

Nous avons synthétisé ces biais selon les différentes phases de réalisation du projet :

Une fois notre modèle construit en limitant au maximum les biais listés précédemment, nous pouvons passer à une phase de validation. A cette étape du projet, nous pouvons proposer une **batterie de tests afin de contrôler que les probabilités restent similaires** si l'on fait varier des variables exogènes au modèle, variables potentiellement discriminantes... De même, il est important de mettre en place une validation continue du modèle afin d'éviter d'éventuelles dérives qui pourraient apparaître dans le temps. Il est recommandé d'effectuer des **contrôles continus sur un faible pourcentage des prédictions** qui ont été réalisées. Enfin, l'intelligence humaine restera toujours celle à prioriser en cas de doute sur les résultats.

Si vous souhaitez plus d'exemples ou d'explications sur les biais existants, cette section du Crash Course de Google est particulièrement intéressant.

NOTRE EXEMPLE SUR UN CAS CONCRET

Afin de conceptualiser ces aspects théoriques, nous avons créé un exemple de toute pièce. Dans cet exemple, nous partons de l'hypothèse que construire une intelligence artificielle pourrait nous aider à déterminer, voire à terme, choisir la prochaine Miss France du célèbre concours de beauté annuel. Pour arriver à notre objectif, nous avons nourri notre algorithme des images des précédentes candidates à l'élection puis labellisé celles qui avaient atteint le top 5 (label 1) ou non (label 0). Nous allons voir étape par étape les différents biais qui sont ressortis d'une démarche « naïve » et inattentive à l'éthique.

1. **Biais survenant de données non-mesurables :** Le premier problème que l'on doit immédiatement remarquer sur ce type de problématique repose sur la définition globale du projet : en se basant seulement sur les images, nous tentons de modéliser une cible basée uniquement sur des critères (features) complètement subjectifs et propres à chacun, qu'est la beauté. Pour qu'une modélisation soit valide, celle-ci doit s'appuyer sur des critères objectifs, des faits universels. On peut facilement déduire qu'ici, un modèle de ce type ne prédit pas la beauté d'une femme, mais des caractéristiques qu'il a appris (couleur de la peau, forme de la bouche, des yeux, etc). Une nouvelle fois, la beauté est quelque chose de personnel, chacun d'entre nous jugera la beauté d'une personne différemment. Le concours Miss France est d'autant plus subjectif car il ne s'appuie pas seulement sur la beauté, mais également sur sa capacité à communiquer par exemple, variables que nous n'avons pas intégré au modèle.
2. **Biais de fréquence :** Le second souci de ce projet repose sur l'échantillon d'images utilisé pour l'apprentissage du modèle. En effet, en ne considérant que l'historique des candidates Miss France des dernières années, nous aurions fait face à un déséquilibre des populations représentées. En effet, nous avons remarqué que le nombre de candidates d'origine asiatique, africaine ou d'autres régions du monde était inférieur au nombre de candidates de type européen notamment. Le modèle serait donc

moins entraîné sur ces personnes et arriverait plus difficilement à séparer nos 2 classes à prédire. Dans ce cas, une femme d'origine asiatique verrait sa probabilité de victoire mécaniquement plus faible toutes choses égales par ailleurs, ce qui d'un point de vue éthique est injustifiable. Nous sommes ici face à des biais de fréquence.

3. **Sous-apprentissage du modèle :** Enfin, il est probable que notre algorithme entraîné trop rapidement souffre de sous-apprentissage. Celui-ci serait alors inadapté pour résoudre ce problème de classification. De plus, en apprenant sur des images sélectionnées rapidement sur le net, l'algorithme n'a pas forcément utilisé sur le visage des Miss mais peut être exploité l'arrière-plan de l'image, élément complètement décorrélé de notre objectif. Pour lutter contre ces phénomènes en reconnaissance d'images, il existe des méthodes permettant d'interprétation permettant de localiser les parties de l'image ayant le plus contribué à l'attribution de la cible.

Typiquement, après toutes ces explications, nous pourrions obtenir ce genre de résultats :

Probabilité d'être dans le top 5 :

Lot de Miss Monde 2012

Probabilité moyenne :
0.45

Lot d'une femme lambda

Probabilité moyenne :
0.70

D'après les probabilités, une personne d'origine asiatique (peu fréquente au concours Miss France) dont nous avons passé un set d'images à notre modèle, aurait moins de chances de gagner le concours qu'une personne de type européen a priori sans critères de beauté apparents. Ces résultats simulés prouveraient l'existence d'un ou plusieurs biais soulevant ensuite de nombreuses questions

éthiques si ce projet d'intelligence artificielle arrivait, à terme, à remplacer les choix humains. Même si les décisions humaines ne sont elles aussi pas toujours éthiquement correctes dans la vraie vie, **un algorithme ne doit pas les reproduire.**

INTELLIGENCES : ARTIFICIELLES, MAIS PAS QUE...

Comme nous l'avons, tout au long de cet article, il y a beaucoup d'étapes à valider afin de pouvoir exploiter concrètement et sereinement une intelligence artificielle qui est prête à **prendre des décisions ou à réaliser des actions plus ou moins impactantes pour l'individu ou la société.** On pourrait penser que ces algorithmes apprennent seuls et prennent leurs décisions seuls : en réalité c'est bien l'Homme qui dicte à la machine ce qu'elle doit apprendre **puis décide des actions à effectuer selon le résultat obtenu.** Dire qu'une IA est "éthique" résulte donc dans la capacité à prendre du recul sur chacune des étapes de son projet afin de ne pas reproduire d'éventuels phénomènes de société ou décisions humaines qui, elles, n'étaient peut-être pas éthiquement correctes.

AVISIA travaillant depuis plusieurs années autour de cette thématique, nous avons donc lancé différents travaux afin de pouvoir **proposer, sereinement, des solutions d'intelligences artificielles** à nos clients. Pour nous, cela passe notamment par :

- **la sensibilisation des consultants :** comment produire des outils irréprochables si ces sujets ne sont jamais abordés ? Avez-vous déjà été sensibilisé à ces biais lors de votre formation ? Non, tout simplement car les technologies d'apprentissage ont évolué plus rapidement que la société (éducation, législation, etc.). C'est donc pourquoi, nous devons préparer nos consultants à identifier et résoudre ces biais.
- **une expertise technique :** Naturellement nous pouvons nous appuyer sur notre expertise technique pour proposer des solutions permettant de valider qu'un modèle respecte bien les normes éthiques actuelles. Pour cela, nous travaillons sur la création d'un outil de validation permettant de tester et d'approuver que vos futurs modèles d'IA ne contiennent pas de biais pour ainsi limiter les dérives éthiques

que nous avons pu constater ces dernières années. L'objectif à terme est de développer un outil permettant de détecter automatiquement des biais discriminatoires pour un algorithme d'intelligence artificielle.

Nous espérons que cet article soulèvera d'éventuelles interrogations et vous incitera à vous poser des questions lors de la création de vos prochains scores. Quelques doutes subsistent encore notamment sur la frontière entre une variable dite « éthique » ou non : nous pouvons facilement comprendre que la couleur de peau ou le genre de la personne sont bien évidemment des champs sensibles sur lesquels il faut faire attention. Qu'en est-il des variables comme le département de résidence, lieu de naissance, l'âge, etc. ? Quoiqu'il en soit, chez AVISIA ces sujets méritent toute notre attention et alimenterons certainement nos prochains débats autour de l'intelligence artificielle. Nous serions aussi ravis d'en discuter avec vous !