# Anticiper les Risques Juridiques des Algorithmes d'IA

## Exemple de la Santé : l'IA au risque de la (loi de) Bioéthique

Philippe Besse

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

# Introduction

## Le buzz de l'IA

## Battage médiatique sur l'Intelligence Artificielle (IA)

- Convergence entre puissance de calcul, stockage, accès à des masses de données & algorithmes d'apprentissage automatique
- Après le buzz grosses data celui sur l'IA
- Succès médiatisés d'algorithmes : *e.g.* apprentissage profond ou *deep learning*
  Reconnaissance d'images, véhicules autonomes, jeu de go...
- l'IA consomme beaucoup de données
- Données confidentielles à fort impact personnel notamment en Santé

## Historique : de la Statistique à l'*IA hybride* par la *Science des Données*

| | | Statistique | Informatique | Algos–Technos |
|---|---|---|---|---|
| 1930-60s | HO | Statistique Inférentielle | Début de l'IA (1955) | Régression / Perceptron |
| 1970s | KO | *Exploratory Data Analysis* | Systèmes experts | Composantes Principales |
| 1980s | MO | Statistique fonctionnelle | Réseaux de neurones | *CARTrees* |
| 1990s | GO | *Data mining* données pré-acquises | | *Boosting, SVM* |
| 2000s | TO | $p >> n$ | *Machine Learning* | *Lasso, random forest* |
| 2008 | | *Data Scientist* | | |
| 2010s | PO | $p$ et $n$ très grands | *Big Data* | *Hadoop* |
| 2012 | | | *Deep Learning* | *ConvNet, TensorFlow* |
| 2016 | | Intelligence Artificielle AlphaGo, Zero… | | *XGBoost, GAN* |
| 2020 | | IA hybride, *federated learning*... | | |

# Introduction

## IA éthique ou *soft law*

*Amazon, Apple, Facebook, Google, IBM, Microsoft... (2015)*




# Introduction

## Éthique, Confiance & Acceptabilité

## Éthique, Confiance, Acceptabilité, Loi

- Enjeux sociétaux & financiers considérables
- Acceptabilité des nouvelles technologies
- Pas de confiance ⇒ pas de données ⇒ pas d'IA
- Entreprises philanthropiques et altruistes ?


## Faire confiance à la Loi plutôt qu'à l'Éthique

- Une loi simple applicable est préférable à des dizaines de chartes éthiques
- Attention à l'*éthical washing*
- Applicabilité des textes de loi *vs.* disruptions technologiques
- Auditabilité des algorithmes (Villani 2018)
- Capacité de détection des transgressions de la loi

## Plan de l'exposé

1. De quelle IA est-il question ?
2. Cadre juridique
3. Réglementation à venir (CE)
4. Les algorithmes d'IA en Santé
   - Domaines de Santé concernés
   - Risques des impacts de l'IA

# 1. Quelle IA ?

## IA du quotidien & apprentissage statistique

## Intelligence Artificielle au quotidien

- Pas de Science Fiction : transhumanisme, singularité technologique, lois d'Asimov
- Pas de Sociologie : destruction des emplois qualifiés, *big data big brother*
- Décisions algorithmiques ou aides automatiques à la décision (IA faible)
- Apprentissage statistique (*statistical learning*) entraînés sur des bases de données $\subset$ apprentissage automatique (*machine learning*) $\subset$ IA
  - Risque de défaut de paiement (**score de crédit**), comportement à risque (assurance)
  - Risque de rupture de contrat (marketing), récidive (justice), passage à l'acte (police)
  - Profilage automatique publicitaire, **professionnel (CV, vidéos, carrière)**
  - Risque de fraude (assurance, banque), défaillance d'un système industriel
  - Diagnostic en imagerie médicale (*deep learning*)
  - Autres applications en **Santé**
  - ... 95% des applications de l'IA (Yan Le Cun)
- NMF, MLG, Arbres binaires, SVM, *random forest, boosting*, *deep learning*...

## Statistique Inférentielle vs. Apprentissage Statistique

- Objectif explicatif de la statistique inférentielle : tests
  - Montrer l'influence d'un facteur en contrôlant le risque d'erreur
  - Essais cliniques phase III : effet d'une molécule *vs.* placebo
  - Tests statistiques : outils de "preuve" scientifique
- Objectif prédictif en Statistique et apprentissage automatique
  - Prévision & explication
  - Prévision brute
  - Sélectionner le modèle ou algorithme minimisant le risque ou erreur de prévision
- Deux types de risque ou d'erreur :
  - Décision (oui / non) sur l'impact d'un facteur sur un phénomène
  - Prévoir l'occurrence ou la valeur prise par ce phénomène

## Principe de l'apprentissage

$p$ Variables ou caractéristiques $\{X^j\}_{j=1,\dots,p}$ observées sur $i = 1, \dots, n$ individus

$Y$ : Variable cible à modéliser ou prédire et observée sur le même échantillon

$$Y = \boldsymbol{f}\left( X^1 \quad X^2 \quad \cdots \quad X^j \quad \cdots \quad X^p \right)$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} = \widehat{\boldsymbol{f}}\left( \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^j & \cdots & x_1^p \\ & & & \vdots & & \\ x_i^1 & x_i^2 & \cdots & x_i^j & \cdots & x_i^p \\ & & & \vdots & & \\ x_n^1 & x_n^2 & \cdots & x_n^j & \cdots & x_n^p \end{bmatrix} \right) + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\widehat{y_0} = \widehat{\boldsymbol{f}}\left( x_0^1 \quad x_0^2 \quad \cdots \quad x_0^j \quad \cdots \quad x_0^p \right)$$

$\widehat{y_0}$ : prévision de $Y$ après observation de $[x_0^1, x_0^2, \cdots, x_0^p]$

# 1. Quelle IA ?

## Qualité et limites de l'apprentissage

## Facteurs de qualité d'une prévision

- Les données
  - Représentatives, sans biais de l'échantillon
  - Classes équilibrées
  - Observation des variables pertinentes, "causales"
  - Variance réduite du bruit
  - Données manquantes et imputation
  - Erreurs de mesure et détection d'anomalies
- Taille $n$ de l'échantillon dépend de
  - Nombre $p$ de variables ou plutôt de
  - paramètres de l'algorithme (*deep learning*)
  - Variance du bruit

**Limites de l'apprentissage statistique**

- Essentiel : qualité des données : représentativité, "quantité"
  apprentissage & reproduction des biais
- Ne pas confondre qualité d'ajustement, qualité de prévision d'une moyenne
  & qualité de prévision individuelle
- L'efficacité du *deep learning* *e.g.* en reconnaissance d'image
  n'est pas transposable à tout problème
- Algorithmes non linéaires *boîte noire* : pas interprétable
- Applications en Santé & Complexité du vivant

# 1. Quelle IA ?

## Risques des impacts sociétaux de l'IA : résumé

**Risques des impacts sociétaux des décisions algorithmiques**
(Besse et al. 2017 Besse et al. 2019-a)

**Cinq questions Juridiques et / ou Éthiques**

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
5. Entraves à la concurrence : comparateurs, *pricing* automatique

Situation complexe : les risques sont interdépendants

## IA, santé et éthique : risques spécifiques

Deux points majeurs : Racine et al. (2019), Wiens et al. (2019) plus un (Besse et al. 2019-b)

1. Consentement éclairé, responsabilité *vs.* Opacité de l'IA
2. Quels sont les risques de discrimination ?
3. Quel équilibre bénéfice / risque ? Intérêt public *vs.* Confidentialité des données


*Source : Le Monde du 8 décembre 2015*

# 2. Textes juridiques

## Mille-feuille de textes

## Mille-feuille de textes juridiques

- Loi n° 78-17 du 6/01/1978 relative à l'informatique aux fichiers et aux libertés
- Loi n° 2015-912 du 24/07/2015 relative au renseignement
- Loi n° 2016-1321 du 7/10/2016 pour une République Numérique (Lemaire)
- 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 (LIL 3)
- Conseil Constitutionnel Décision n° 2018-765 DC du 12 juin 2018
- Code pénal
- Code des relations entre le public et les administrations
- Code de la Santé publique
- ...

# 2. Textes juridiques

## RGPD

## Règlement Général sur la Protection des Données

- Considérant 71 : 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**, et 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 **qui prévienne, entre autres, les effets discriminatoires** à l'égard des personnes physiques fondées sur la l'origine raciale ou ethnique, les opinions politiques, la religion ou les convictions, l'appartenance syndicale, le statut génétique ou l'état de santé, ou l'orientation sexuelle, ou qui se traduisent par des mesures produisant un tel effet. La prise de décision et le profilage automatisés fondés sur des catégories particulières de données à caractère personnel ne devraient être autorisés que dans des conditions spécifiques

## Règlement Général sur la Protection des Données

- Article 12 : Le **responsable du traitement** prend des mesures appropriées pour fournir toute information [...] ainsi que pour procéder à toute communication [...] en ce qui concerne le traitement à la personne concernée d'une **façon concise, transparente, compréhensible et aisément accessible, en des termes clairs et simples**, [...]
- Articles 14 et 15 : [...] le responsable du traitement fournit à la personne concernée les informations suivantes nécessaires pour garantir un **traitement équitable et transparent** à l'égard de la personne concernée : [...] l'existence d'une prise de **décision automatisée**, y compris un profilage, visée à l'article 22, paragraphes 1 et 4, et, au moins en pareils cas, des **informations utiles concernant la logique sous-jacente**, ainsi que l'importance et les conséquences prévues de ce traitement pour la personne concernée.

## Article 22 : Décision individuelle automatisée, y compris le profilage

1. La personne concernée a le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un **traitement automatisé**, y compris le **profilage**, produisant des effets juridiques la concernant ou l'**affectant de manière significative** de façon similaire.

2. Le paragraphe 1 ne s'applique pas 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 ; ou
   - *c* est fondée sur le **consentement** explicite de la personne concernée.

3. Dans les cas visés au paragraphe 2, points a) et c), le responsable du traitement met en œuvre des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée, au moins du droit de la personne concernée d'**obtenir une intervention humaine** de la part du responsable du traitement, d'exprimer son point de vue et de contester la décision.

4. Les décisions visées au paragraphe 2 **ne peuvent être fondées** sur les catégories particulières de **données à caractère personnel** (cf. article 9 : biométriques, génétiques, de santé, ethniques ; orientation politique, syndicale, sexuelle, religieuse, philosophique) **sous réserve** d'un intérêt public substantiel et que des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée ne soient en place.

# 2. Textes juridiques

## Explicabilité

## Loi n° 2016-1321 du 7/10/2016 pour une République Numérique (Lemaire)

- Article 6 : Sous réserve des secrets protégés, les administrations ... publient en ligne les règles définissant les principaux traitements algorithmiques utilisés dans l'accomplissement de leurs missions lorsqu'ils fondent des décisions individuelles.
- Article 50 : Les opérateurs de plateformes en ligne dont l'activité dépasse un seuil de nombre de connexions défini par décret élaborent et diffusent aux consommateurs des bonnes pratiques visant à renforcer les obligations de clarté, de transparence et de loyauté.

## Décret du 16/03/2017 Art. R. 311-3-1-2. (APB)

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

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.

# 2. Textes juridiques

## Risque de discrimination

## 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
- 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, 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.

## Loi claire et explicite sur la non-discrimination

- Article 21 de la Charte des Droits Fondamentaux de l'Union Européenne
- Code pénal
  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

## Faire appliquer la loi :

- Comment détecter une discrimination algorithmique ?

# 2. Textes juridiques

## Textes spécifiques en Santé

## Redevabilité *vs.* Opacité

- **Article L1111-4 du code de la santé publique**
  *Aucun acte médical ni aucun traitement ne peut être pratiqué sans le* **consentement libre et éclairé** *de la personne et ce consentement peut être retiré à tout moment*
- Quelle explication Intelligible d'une décision algorithmique
- issue d'un algorithme d'IA opaque ?
- Comment caractériser les responsabilités en cas d'échec ou d'erreur ?

## Non-discrimination

- **Article L1110-3 du code de santé publique** Modifié par Loi 2012-954 du 6 août 2012
- Aucune personne ne peut faire l'objet de discriminations dans l'accès à la prévention ou aux soins

## Bénéfice *vs.* Risque

### Article L1461-3 du code de santé publique

- I.-Un **accès aux données** à caractère personnel du système national des données de santé ne peut être autorisé que pour permettre des traitements : 1. Soit contribuant à une finalité mentionnée au III de l'article L. 1461-1 et répondant à un motif d'intérêt public ;
- III.-Le système national des données de santé a pour **finalité** la mise à disposition des données, dans les conditions définies aux articles L. 1461-2 et L. 1461-3, pour contribuer : 1. À l'information sur la santé ainsi que sur l'offre de soins, la prise en charge médico-sociale et leur qualité ; 2. À la définition, à la mise en œuvre et à l'évaluation des politiques de santé et de protection sociale ; 3. À la connaissance des dépenses de santé, des dépenses d'assurance maladie et des dépenses médicosociales ; 4. À l'information des professionnels, des structures et des établissements de santé ou médico-sociaux sur leur activité ; 5. À la surveillance, à la veille et à la sécurité sanitaires ; 6. À la **recherche**, aux études, à l'évaluation et à l'**innovation dans les domaines de la santé** et de la prise en charge médico-sociale.

# 2. Textes juridiques

## Résumé

## Cadre juridique : résumé

- Cadre lourd et complexe
- Qualité des décisions : rien d'explicite
- Explicabilité des décisions : flou et inadapté
- Discrimination : stricte mais inapplicable
- Intérêt public substantiel de la recherche en Santé : Définition ?
  - Ouverture (*Health Data Hub*) des données pour la Recherche (publique, privée ?)
  - Sous réserve de Confidentialité
  - Objections de la CNIL et de la CNAM pour un dépôt "Microsoft Azure"

# 3. Réglementation à venir

## Annonces européennes


## Lignes directrices en matière d'éthique pour une IA de confiance

Groupe d'experts indépendants de hauts niveaux sur l'Intelligence artificielle (2018–2020)

- (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**


## IA – Une approche européenne axée sur l'excellence et la confiance

*Livre blanc* — 19/02/2020

- IA, qui combine données, algorithmes et puissance de calcul
- Risques potentiels, tels que l'opacité de la prise de décisions, la discrimination
- Enjeu majeur : acceptabilité et adoption de l'IA nécessite une IA digne de confiance
- Fondée sur les droits fondamentaux de la dignité humaine et la protection de la vie privée
- **Proposer les éléments clefs d'un futur cadre réglementaire**
- Déceler et prouver d'éventuelles **infractions à la législation**
- Notamment aux dispositions juridiques qui protègent les droits fondamentaux, à cause de l'opacité des algorithmes


## Chapitre III : Liste d'évaluation pour une IA digne de confiance

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 & bien commun ? Balance bénéfice / risque
7. Responsabilité (auditabilité, recours...)

# 3. Réglementation à venir

## Qualité, robustesse, résilience des décisions algorithmiques

**Qualité des décisions & vide juridique**

- Algorithme d'apprentissage : erreur de prévision, qualité de décision, confiance
- Taux d'erreur de 3% en image *vs.* 30 à 40% pour le risque de récidive
- Considérant (71) du RGPD mais loi française muette
- *Ethical washing* & intérêt commercial : cf. publication des sondages d'opinion
- Ne pas confondre estimation / prévision d'une moyenne (*loi des grands nombres*) et celle d'un comportement individuel
- Éthique : Obligation de moyen, pas de résultat mais obligation de transparence
- Industrie et Santé : objectif de certification


**Exemple de questions de la liste d'évaluation**

**2** *Robustesse technique et sécurité (résilience, précision...)*

- 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 supplémentaires sont nécessaires, par exemple pour améliorer la précision et éliminer les biais ?

### Précision & choix d'une métrique

- 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 $F_\beta$, entropie...
- Multiclasse
  Taux d'erreur moyen, $F_\beta$ moyen...

### Robustesse

- Valeurs atypiques et choix de la fonction perte
- Détection des anomalies (*outliers*) de la base d'apprentissage, en exploitation

### Résilience

- Données manquantes de la base d'apprentissage, en exploitation

# 3. Réglementation à venir

## Explicabilité d'une décision


**Exemple de questions de la liste d'évaluation**

**4** *Transparence (explicabilité, communication...)*

- 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 ?

## Quelle niveau d'explication ? Pour qui ? (Barredo Arrieta et al. 2020)

426 références !

1. Fonctionnement général de l'algorithme, domaines de défaillances
   - Modèles linéaires, arbres *vs.* algorithme opaque : neurones, agrégation, SVM...
     - Approximation : linéaire, arbre, règles,...
     - Importance des variables, stress de l'algorithme et impact (Bachoc et al. 2020)
2. Décision spécifique
   - Concepteur : Expliquer une erreur, y remédier : ré-apprentissage
   - Personne concernée : client, patient, justiciable...
     - Interprétable : modèle linéaire, arbre de décision
     - Approximation locale : LIME, contre-exemple, règles,...
     - *a minima* : risque d'erreur

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

# 3. Réglementation à venir

## Risques de discrimination


## Exemple de questions de la liste d'évaluation

**5** *Diversité, non-discrimination et équité*

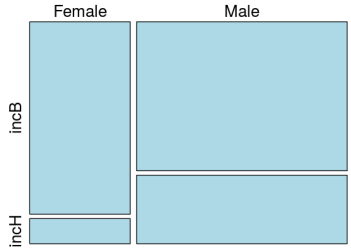
- Avez-vous prévu une définition appropriée de l'équité que vous appliquez dans la conception des SIA ?
- Avez-vous mis en place des processus pour tester et contrôler les biais éventuels au cours de la phase de mise au point, de déploiement et d'utilisation du système ?
- Avez-vous prévu une analyse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

**Détection d'une discrimination de groupe ou indirecte :** *critères statistiques*

- Pas de définition juridique de l'équité : absence de discrimination
- Indicateurs de discrimination : Zliobaité (2017), 70 sur `aif360.mybluemix.net`
- Critères, redondants, corrélés : Friedler et al. (2019), Verma et Rubin (2018)
- En pratique : Trois niveaux de biais estimés par IC (Besse et al. 2020) : Dépôt Github des fonctions en R et Python
  1. Effet disproportionné ou *Disparate Impact* (*demographic equality*) : Code du travail aux USA : $DI = \frac{\mathbb{P}(\hat{Y}=1|S=0)}{\mathbb{P}(\hat{Y}=1|S=1)}$ (Barocas et Selbst, 2016)
  2. Taux d'erreur conditionnels (*overall error equality*) : $\frac{\mathbb{P}(\hat{Y}\neq Y|S=0)}{\mathbb{P}(\hat{Y}\neq Y|S=1)}$ Reconnaissance faciale, santé (Besse et al. 2019), emploi (De Arteaga et al. 2019)
  3. Égalité des cotes (*equali odds*) : $\frac{\mathbb{P}(\hat{Y}=1|Y=0,S=0)}{\mathbb{P}(\hat{Y}=1|Y=0,S=1)}$ et $\frac{\mathbb{P}(\hat{Y}=1|Y=1,S=0)}{\mathbb{P}(\hat{Y}=1|Y=1,S=1)}$ Justice "prédictive" : Propublica *vs.* equivant (Compas)

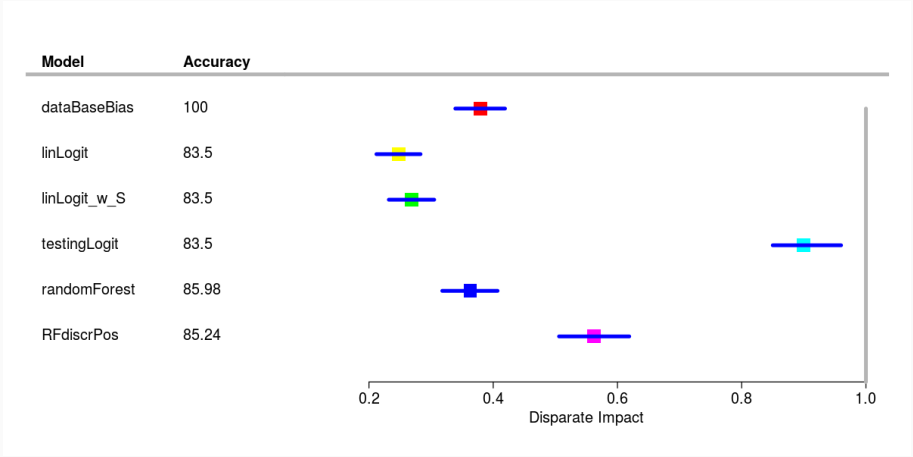
**Cas d'Usage illustratif :** *Adult Census Dataset*

- Code disponible sur github/wikistat
- Données publiques de l'UCI
- 48 842 individus décrits par 14 variables issues d'un sondage aux USA (1994)
  - Genre, origine ethnique, niveau d'éducation, occupation, statut familial, nombre d'heures travaillées par semaine...
  - $Y$ : Seuil de Revenu inférieur ou supérieur à 50k$
  - Prévision de la classe ou "solvabilité"
  - Données largement biaisées selon le genre, biaisées selon l'origine




$$DI = \frac{\mathbb{P}(Y=1|S=0)}{\mathbb{P}(Y=1|S=1)} = 0.37$$

$$\mathbb{P}\left(DI \in [0.35, 0.38]\right) = 0.95$$




*Détection de la discrimination indirecte ($DI = \frac{\mathbb{P}(\widehat{Y}=1|S=0)}{\mathbb{P}(\widehat{Y}=1|S=1)}$) de différents algorithmes*

Attention : impact de la correction de l'effet disproportionné sur les deux autres biais

# 3. Réglementation à venir

## Conclusion

## Réglementation à venir : conclusion

- Questionnaire lourd d'évaluation *ex ante* d'un projet d'IA *vs.* audit *ex post*
- Analogue au *PIA (privacy impact assessment)* du RGPD (analyse d'impact de la protection des données)
- Renversement de la charge de preuve
- Documenter dès le lancement d'un projet : objectifs, données...
- En Santé : remboursement des systèmes connectés (CNEDiMTS de la HAS) (annexe)
- Travail en devenir : textes juridiques à venir et recherches en cours
- Anticipation néanmoins indispensable de cette documentation *ex ante*

# 4. L'IA en Santé

## Domaines concernés

## Domaines de Santé concernés par L'IA

- Révision de la Loi Bioéthique de 2011
- États Généraux de la Bioéthique (2018) abordent neuf points :
  - Procréation, Embryon, Dons d'organes, Fin de vie, Neurosciences, Environnement
  - Trois points concernent l'IA :
    1. Bases de données de santé
    2. Médecine génomique
    3. IA & robotisation de la médecine


*Source : Le Monde du 6 janvier 2018*

## Bases de Données

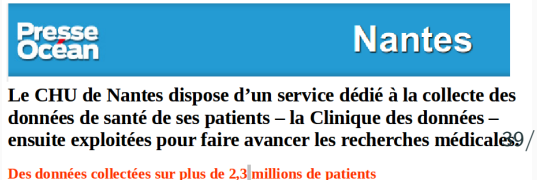
- SNDS (Système National des Données de Santé) $\subset$ *Health Data Hub* $\subset$ *Microsoft Azure*
  - Assurance maladie (base SNIIRAM)
  - Hôpitaux (base PMSI)
  - Causes médicales de décès
  - Données relatives au handicap
  - Assurance maladie complémentaire
- Plan Médecine France Génomique 2025 : SeqOIA (Paris) AURAGEN (Lyon)


Accès : INDS (Institut National des Données de Santé) après avis de la CNIL


Accueil > #HealthTech > Emmanuel Macron souhaite l'ouverture d'un « hub des données de santé » (...)

**Emmanuel Macron souhaite l'ouverture d'un « hub des données de santé » respectant l'anonymat le 29 mars 2018**

Le chef de l'Etat a mis en avant les progrès de la médecine prédictive déjà réalisés grâce à l'IA


Presse Océan — Nantes

**Le CHU de Nantes dispose d'un service dédié à la collecte des données de santé de ses patients – la Clinique des données – ensuite exploitées pour faire avancer les recherches médicales**

Des données collectées sur plus de 2,3 millions de patients

## Médecine génomique

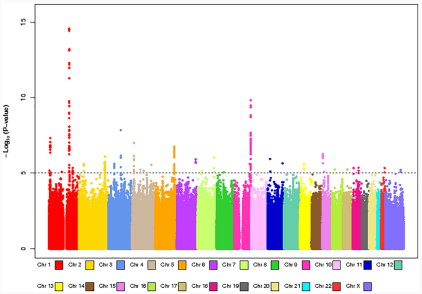
- Médecine 4p :
  - Prédictive d'un risque pathologique
  - Préventive de ce risque
  - Participative car participation nécessaire à la prévention
  - Personnalisée ou de précision car thérapie ciblée pour une personne
- Donc ... génomique
- Médecine translationnelle
  - Accélérer les applications de la recherche
  - Favoriser les échanges, pluridisciplinarité, données ouvertes...


*Source : Réseau Régional de Cancérologie Île-de-France*

*2017*

## Données de médecine Génomique

- Génome complet (3400Mpb), 26517 gènes protéiques (1.5%)
- *Genomic Wide Association Studies (GWAS)* ou études pangénomiques
- *Single Nucleotide Polymorphism (SNP)*




*Manhattan Plot (Lindström et al. 2017)*

## Études pangénomiques

- Bases de données jusqu'à 167 millions de *SNP* par individus
- Associer mutations avec phénotypes (pathologies) : tests statistiques classiques
- Maladies rares monogénique *vs.* polygéniques ou polyfactorielles, chroniques
- Problèmes : Reproductibilité (Ioannidis 2016), environnement, épigénétique

## IA et robotisation de la médecine

- Robots de micro-chirurgie
- Aide au diagnostic
  - Imagerie médicale, ECG, EEG : apprentissage profond ou *deep learning*
  - Identification de biomarqueurs préventifs : études "omiques"
- Aide aux choix thérapeutiques : *e.g.* IBM Watson
- Surveillance effets secondaires & base SNIIRAM (Morel et al. 2019)
- Suivi épidémiologique de grandes cohortes : `constances` (Zins et al. 2010)

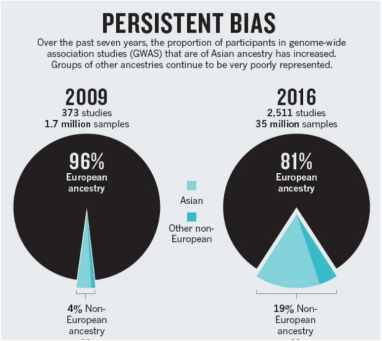
*Source : Life Science Daily News Desk on January 31, 2017*

# 4. L'IA en Santé

## Risques des impacts de l'IA en Santé

## Biais et discrimination en Santé

- Biais de société (Lee et al. 2019) exacerbés (Obermayer et Mullainathan 2019)
- Médecine 4p personnalisée : Biais des bases pangénomiques
  - Ethnique : population d'ascendance blanche européenne (Popejoy et Fullerton, 2016)
  - Âge et environnement : bases transversales et pas longitudinales
  - Genre : Chang et al. (2014), Pulit et al. (2017)




*Biais des études pangénomiques ; Popejoy et Fullerton (2016)*

## Réduire les biais en santé

- Constitution représentatives des cohortes
  - `constances` 200 000 personnes (Zins et Goldberg 2011)
  - Sous-ensemble de SNIIRAM (Schwarzinger et al. 2018)
- Réglementation
  - HAS (2019) Remboursement des DSC Dispositifs de Santé Connectés
  - FDA (2019) certification des AI/ML-SaMD *Artificial Intelligence and Machine Learning Software as a Medical Device*

*Source : INRIA Bordeaux*

## Consentement éclairé vs. Opacité des algorithmes

- Médecine de population : modèles épidémiologiques explicables
- *Deep learning* & diagnostic individualisé
  - Droit du RGPD et LIL3 d'une intervention humaine
  - Boîte noire : action ou sens de l'action inconnue des variables
  - Focaliser sur le risque (London 2019)
  - Estimation du risque (Liu et al. 2019)
  - Certification des dispositifs (HAS, FDA)
  - Protocole d'information du patient ?

## Risques de confidentialité *vs.* Intérêt public

- Risques
  - Pseudonymisation (Article L1461-4 code santé publique) du *HDH* : NIRPP codé, Nom, Adresse
  - Ré-identification : date de naissance, code postal, sexe, nombre d'enfants...
    Narayanan et Shmatikov (2008), Rubinstein et Hartzog (2015), Rocher et al. (2019)
  - Anonymisation par confidentialité différentielle Dwork et Roth (2014)
  - Génome : Clef d'identification (Robinson et Glusman, 2017) ou empreinte génétique
- Intérêt public
  - Recherche & nombre de publications
  - Résultats substantiels en santé publique (CEIP)


Healthcare IT News

Global Edition Privacy & Security

**Google, University of Chicago named in suit charging misuse of patient data**

The class action complaint alleges that, despite being deidentified, Google's expertise in data mining and AI makes it "uniquely able to determine the identity" of the medical records shared with it by the university.

By Nathan Eddy | July 01, 2019 | 04:03 PM

## Intérêt public : résultats substantiels

- Épidémiologie (*e.g.* Journées cohorte `constances`) : modèles statistiques linéaires
- Médecine génomique & maladies rares (Pujol 2019, SFMPP) : tests statistiques
  - Pénétrance : probabilité de développer la maladie
  - Actionnabilité : possibilités médicales ouvertes par un diagnostic de risque
  - Kim et al. (2019) Traitement de la maladie de Batten pour une fillette
- Imagerie & diagnostic (*deep learning*)
  - Esteva et al. (2017), De Fauw J. et al. (2018), Haenssle et al. (2018), Yala et al. (2019)
  - Reproductibilité : Liu et al. (2019)
  - Exception : Oakden-Rayner et al. (2019)
  - Certification de la FDA : Topol (2019)
- Biomarqueurs protéomiques
  - Williams et al. (2019) & maladies multifactorielles

**Intérêt public : résultats inconsistants**

Médecine génomique et maladies plurigéniques, chroniques

- Facteurs génétiques ne sont pas majeurs (Rappaport, 2016)
- Pénétrance généralement très faible au regard de l'environnement (Pujol, 2019)
- Capacités prédictives inexistantes (Patron et al. 2019)
- Inférieure à celle des variables cliniques (Udler et al. 2019)
- Déontologie scientifique & reproductibilité des résultats
  - "Nettoyage" et sélection des données (Ambroise et McLachlan, 2002)
  - Sur-apprentissage : Montañez et al. (2018)
  - Évaluation sur un échantillon test indépendant (Liu et al. 2019)

# Conclusion

**Tout est lié et affaire de compromis**

**Utilité** d'un système d'IA : équilibre bénéfice / risque

1. Confidentialité, protection ds données *vs.* connaissance de la variable sensible
2. Qualité, robustesse de la décision algorithmique
3. Explicabilité de la décision algorithmique
4. Types de bais donc risques de discrimination
    - Biais systémique, des erreurs, de leur asymétrie

**En chantier**

- Auditabilité et contrôle : liste d'évaluation & renversement de la charge de preuve
- Normes : ANSI, IEEE, ISO ?
- Certification en santé : *FDA*, HAS et dans l'industrie : projet DEEL, ANITI

## Recommandations

1. Accès aux données de santé (CNIL, INDS, CEIP)
   - Pseudonymisation (HDH) & réidentification : Audit des accès
   - Épidémiologie, diagnostics, maladies rares
   - Médecine génomique antinomique avec l'IA
2. Déontologie de la recherche
   - Détecter biais et discriminations indirectes
   - Rigueur d'analyse et d'estimation des erreurs : reproductibilité
   - Publication des codes
3. Réglementation (HAS FDA) des DSC AI/ML-SaMD
   - Certification : biais & adaptation
   - Protocole d'information du patient

## Les grandes manœuvres

*Google & Ascension, Fitbit...*

*Sanofi & Aection*

TECH

**Google sister-company Verily is teaming with big pharma on clinical trials**

PUBLISHED TUE, MAY 21 2019 • 8:00 AM EDT | UPDATED TUE, MAY 21 2019 • 11:43 AM EDT

*Source : CNBC Tuesday May 21th 2019*

**NHS DeepMind deal broke data protection law, regulator rules**

The UK's data watchdog has ruled that the NHS didn't comply with data protection legislation when it shared patient details with Google-owned DeepMind

*Source : Wired Monday July 3rd 2017*

# 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, *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. (2020). Détecter, évaluer les risques des impacts discriminatoires des algorithmes d'IA , Contribution au séminaire Défenseur des Droits et CNIL, 28 mai 2020, soumis.
- Besse P., Besse-Patin A., Castets-Renard C. (2019-b). Implications juridiques et éthiques des algorithmes d'intelligence artificielle dans le domaine de la santé, soumis.
- Besse P., Castets-Renard C., Garivier A. (2017). Loyauté des Décisions Algorithmiques, Contribution au Débat "Éthique et Numérique" de la CNIL.
- Besse P., Castets-Renard C., Garivier A., Loubes J.-M. (2019-a). L'IA du Quotidien peut elle être Éthique ? Loyauté des Algorithmes d'Apprentissage Automatique, *Statistique et Société*, Vol6 (3), pp 9-31.
- Besse P. del Barrio E. Gordaliza P. Loubes J.-M., Risser L. (2020-b). A survey of bias in Machine Learning through the prism of Statistical Parity for the Adult Data Set, à paraître.
- Chang D., Gao F., Slavney A.,Ma L., Waldman Y., Sams A., Billing-Ross P., Madar A., Spritz R., KeinanA. (2014). Accounting for eXentricities : Analysis of the X Chromosome in GWAS Reveals X-Linked Genes Implicated in Autoimmune Diseases, *PLoS One*, 9(12).
- Commission Européenne (2016). Règlement Général sur la Protection des Données.
- Commission Européenne (2018). 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 Fauw J. et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease, *Nature Medicine*, 24, pp 1342-1350.

# Références suite


- Dwork C., Roth A. (2014). The Algorithmic Foundations of Differential Privacy, *Foundations and Trends in Theoretical Computer Science*, vol. 9, n 3-4, 211-407.
- Esteva A., Kuprel ., Novoa R., Ko J., Swetter S., Blau H., Thrun S. (2017). Dermatologist-level classification of skin cancer with deep neural networks, *Nature* volume 542, pages 115-118.
- FDA (2019). Artificial Intelligence and Machine Learning in Software as a Medical Device.
- France (2018). États Généraux de la Bioéthique : Rapport de Synthèse du Comité Consultatif National d'Éthique - Opinion du comité citoyen, *La Documentation Française*.
- Friedler S., Scheidegger C., Venkatasubramanian S., Choudhary S., Ha-milton E., Roth D. (2019). Comparative study of fairness-enhancing interventions in machine learning. in FAT'19, p. 32938.
- Haenssle H. et al. (2018). Man against machine : diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists, *Annals of Oncology*, Volume 29, Issue 8.
- 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.
- Ioannidis J. (2016). Why Most Clinical Research Is Not Useful, *PLOS Medicine*, Volume 13, Issue 6.
- Kim J. et al. (2019). Patient-Customized Oligonucleotide Therapy for a Rare Genetic Disease, *New England Journal of Medicine*.
- Lee P., Le Saux M., Siegel R., Goyal M., Chen C., Ma Y., Meltzer A. (2019). Racial and ethnic disparities in the management of acute pain in US emergency departments : Meta-analysis and systematic review, *American Journal of Emergency Medecine*, 37(9), 1770-1777.
- Lindström S., Loomis S., Turman C., Huang H., Huang J. (2017). A comprehensive survey of genetic variation in 20,691 subjects from four large cohorts, *PIOS ONE*, (12) 3.
- Liu X. et al. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging : a systematic review and meta-analysis, *The Lancet Digital Health*, (1) 6, pp 271-297.
- London A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions : Accuracy versus Explainability, *Hasting Center Report*,

# Références suite


- Montanez C., Fergus P., Montanez A., Hussain A., Al-Jumeily D., Chalmers C. (2018). Deep Learning Classification of Polygenic Obesity using Genome Wide Association Study SNPs, *2018 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp 1-8.
- Morel M., Bavry E, Gaïffas S., Guilloux A., Leroy F. (2019). ConvSCCS : convolutional self-controlled case series model for lagged adverse event detection, *Biostatistics*, kxz003.
- Narayanan A., Shmatikov V. (2008). Robust De-anonymization of Large Sparse Datasets, *2008 IEEE Symposium on Security and Privacy*.
- Oakden-Rayner L. et al. (2019). Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging, arXiv :1909.12475.
- Obermayer Z., Mullainathan S. (2019). Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People, , FAT 19, Proceedings of the Conference on Fairness, Accountability, and Transparency. item Popejoy A., Fullerton S. (2016). Genomics is failing on diversity, *Nature*, 538, 161-164.
- Patron P., Serra-Cayuela A., Han B., Li, C. Wishart D. (2019). Assessing the performance of genome-wide association studies for predicting disease risk.
- Pujol P. (2019). *Voulez-vous savoir ? Ce que nos gènes disent de notre santé*, Humensciences, 192 p.
- Pulit S., Karaderi T., Lindgren C. (2017). Sexual dimorphisms in genetic loci linked to body fat distribution, *Bioscience Report*, 37(1).
- Racine E, Boehlen W, Sample M (2019) Healthcare uses of artificial intelligence : Challenges and opportunities for growth. Healthc Manage Forum 32 :272275.
- Rappaport S. (2016). Genetic Factors Are Not the Major Causes of Chronic Diseases, *PLoS ONE*, 11(4) : e0154387.
- Robinson M., Glusman G. (2017). Genotype fingerprints enable fast and private comparison of genetic testing results for research and direct-to-consumer applications, *Bioinformatics*.
- Rocher L. , Hendrickx, de Montjoye Y.-A. (2019), Estimating the success of re-identifications in incomplete datasets using generative models, *Nature Communications* volume 10 , Numéro d'article : 3069.
- Rubinstein I., Hartzog W. (2015). Anonymization and Risk, *New York University of Law, Public Law & Legal Thery Research Paper Series*,

# Références fin


- Schwarzinger M., Pollock B., Hasan O., Dufouil C., Rehm J., Baillot S., Guibert Q., Planchet F., Luchini S. (2018). Contribution of alcohol use disorders to the burden of dementia in France 2008-13 : a nationwide retrospective cohort study, *The Lancet Public Health*.
- Topol E. (2019). High-performance medicine : the convergence of human and artificial intelligence, *Nature Medecine* (25) 1, pp 44-56.
- Udler M., McCarthy M., Florez J., Mahajan A. (2019). Genetic Risk Scores for Diabetes Diagnosis and Precision Medicine *Endocrine Reviews*, (40) 6, pp 1500-1520.
- Verma S., Rubin J. (2018). Fairness Definitions Explained, ACM/IEEE International Workshop on Software Fairness.
- Villani C., Schoenauer M., Bonnet Y., Berthet C., Cornut A.-C., Levin F., Rondepierre B.(2018). Donner un sens à l'Intelligence Artificielle pour une stratégie nationale et européenne, *La Documentation Française*, rapport public.
- Wiens J., Saria S., Sendak M., Ghassemi M., Liu V. (2019). Do no harm : a roadmap for responsible machine learning for health care, *Nature Medecine*, (25) 9, pp 1337-1340.
- Williams S. et al. (2019). Plasma protein patterns as comprehensive indicators of health, *Natude Medecine*, (25) 12, pp 1851-1857.
- Wright K, Rand K, Kermany A, Noto K, Curtis D, Garrigan D, Slinkov D, Dorfman I, Granka J, Byrnes J, Myres N, Ball C, Ruby G. (2019). A prospective analysis of genetic variants associated with human lifespan, *G3 Genes, Genomes, Genetics*, vol. 9, $n^o$9, 2863-2878.
- Xu D., Yuan S., Zhang L., Wu X. (2018). FairGAN : Fairness-aware Generative Adversarial Networks, IEEE International Conference on Big Data, pp. 570-575.ala
- Yala A., Lehman C., Schuster T., PortnoiT., Barzilay R. (2019). A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction, *Radiology*, Vol. 292, No. 1. *
- Zins M. et al. (2010). The CONSTANCES cohort : an open epidemiological laboratory, *BMC Public Health*, (10) 1.
- Zins M., Goldberg M., Constances Team. (2015). The French CONSTANCES population-based cohort : design, inclusion and follow-up, *European Journal of Epidemiology*, (30) 12, pp 1317-1328.
- Zliobaité I. (2017). Measuring discrimination in algorithmic decision making, *Data Min Knowl Disc* 31, 10601089.

# Annexes

## Liste évaluation HAS

HAUTE AUTORITÉ DE SANTÉ

**Liste d'évaluation de la Haute Autorité de Santé**

- Évaluer les technologies de Santé
- Dépôt d'un dossier auprès de la Commission nationale d'évaluation des dispositifs médicaux et des technologies de santé(CNEDiMTS)
- p43 : Informations descriptives spécifiques à fournir pour les fonctionnalités du dispositif médical s'appuyant sur des procédés d'apprentissage automatique (technologies relevant du champ de l'intelligence artificielle)
- Grille descriptive de 42 questions
- Accès au document

# Annexes

## Liste évaluation experts européens


**Liste évaluation provisoire de la Commission Européenne**

- Lignes directrices en matière d'éthique pour une IA digne de confiance
- Groupe d'experts de haut niveau sur l'intelligence artificielle
- Chapitre III p32 : Liste d'évaluation pour une IA digne de confiance (version pilote)
- Accès au document