Impacts & Risques des Algorithmes d'IA

PHILIPPE BESSE, CÉLINE CASTETS RENARD & JEAN-MICHEL LOUBES

Intelligence Artificielle (IA) 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
- Apprentissage statistique (*statistical learning*) entraînés sur des bases de données
  $\subset$ apprentissage automatique (*machine learning*) $\subset$ IA
  - Profilage automatique publicitaire, **professionnel (CV, vidéos, carrière)**
  - Risque de défaut de paiement (crédit), comportement à risque (assurance)
  - Risque de rupture de contrat (marketing), récidive (justice), passage à l'acte (police)
  - Risque de fraude (assurance, banque), défaillance d'un système industriel
  - Diagnostic en imagerie médicale (*deep learning*)
  - ... 95% des applications de l'IA (Yan Le Cun)
- NMF, MLG, Arbres binaires, SVM, *random forest, boosting*, *deep learning*...

Principe de l'apprentissage statistique

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 $[x_0^1, x_0^2, \cdots, x_0^p]$

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

Pourquoi se préoccuper d'Éthique en IA ?

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

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

- Applicabilité des textes de loi *vs.* disruptions technologiques
- **Auditabilité** des algorithmes (Villani 2018)
- Capacité de détection des transgressions de la loi

Quels risques des impacts sociétaux des décisions algorithmiques ?

Cinq questions Juridiques et / ou Éthiques

1. Protection : propriété, confidentialité des données personnelles (RGPD, CNIL)
2. Entraves à la concurrence : comparateurs, *pricing* automatique
3. Biais & Discrimination des décisions algorithmiques
4. Explicabilité *vs.* opacité des algorithmes
5. Qualité, robustesse des prévisions donc des décisions

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 (RGPD) : 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.

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.

Article 225-2 du 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

BUSINESS NEWS OCTOBER 10, 2018 / 5:12 AM / 8 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Source : Agence Reuters

On Artificial Intelligence - A European approach to excellence and trust

European Commission : White paper — 19/02/2020

- 16 occurrences du mot ”discrimination”
- there is a need to examine whether current legislation is able to address the **risks of AI** [...] **new legislation is needed**
- Risks for fundamental rights, including personal data and privacy protection and **non-discrimination**
- AI applications for **recruitment processes** as well as in situations impacting workers’rights would always be considered **“high-risk”**
- **Training data**
  Requirements to take reasonable measures aimed at ensuring that such subsequent use of AI systems does not lead to outcomes entailing **prohibited discrimination**. These requirements could entail in particular obligations to use **data sets** that are sufficiently **representative**, especially to ensure that all relevant dimensions of gender, ethnicity and other possible grounds of prohibited discrimination are appropriately reflected in those data sets ;
- The need to **verify the data used for training** and the relevant programming and **training methodologies**, processes and techniques used to **build, test and validate** AI systems.

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

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

- (52) Si les **biais injustes** peuvent être évités, les systèmes d'IA pourraient même **améliorer le caractère équitable de la société**.
- (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

Liste d'évaluation pour une IA digne de confiance

5 Diversité, non-discrimination et équité

- Avez-vous prévu une stratégie ou un ensemble de procédures pour **éviter de créer ou de renforcer des biais injustes** dans le système d'IA, en ce qui concerne tant l'utilisation des données d'entrée que la conception de l'algorithme ?
- Avez-vous réfléchi à la diversité et à la **représentativité** des utilisateurs dans les données ? Avez-vous procédé à des essais portant sur des populations spécifiques ou des cas d'utilisation problématiques ?
- Avez-vous recherché et utilisé les outils techniques disponibles pour améliorer votre **compréhension des données**, du **modèle** et de la **performance** ?
- 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 (in)directe : *Testing*

- Riach et Rich (2002) : économistes, sociologues
- Comité National de l'Information Statistique
- Observatoire des Discriminations (Paris 1)
- DARES (Direction de l'Animation, des Études, de la Recherche et des Statistiques)
- ISM Corum, Challe et al. (2020)

Une étude montre des discriminations à l'embauche « significatives » en fonction de l'origine

Sur les 103 entreprises testées, les chercheurs identifient « entre 5 et 15 entreprises discriminantes » à l'encontre du « candidat présumé maghrébin ».

Le Monde avec AFP • Publié le 08 janvier 2020 à 18h30 - Mis à jour le 09 janvier 2020 à 07h41

Source : Le Monde

USA : Civil Rights act & Code of Federal Regulations

Title 29 - Labor : PART 1607—UNIFORM GUIDELINES ON EMPLOYEE SELECTION PROCEDURES (1978)

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

Mesurer pour progresser dans l'égalité des chances

DdD & CNIL : Guide méthodologique à l'usage des acteurs de l'emploi (2012)

- 24 fiches 110 pages
- Mesurer des discriminations
- Question des statistiques ethniques : méthode patronymique

Tableau 3 : Probabilité d'être recruté suivant le type de fonction postulée, l'origine supposée et le niveau d'étude

Type de fonction postulée	Candidatures évoquant une origine...					
	« européenne »			« extra-européenne »		
	< Bac	Bac/Bac+2	≥ Bac+5	< Bac	Bac/Bac+2	≥ Bac+5
Fonctions en contact clientèle requérant un Bac/Bac+2	1%	**5%**	4%	0%	**2%**	5%
Fonctions d'encadrement requérant un Bac+5	0%	1%	**9%**	0%	0%	**6%**
Ensemble	1%	4%	8%	0%	2%	5%

Algorithmes d'apprentissage et discrimination

- Biais structurel ou social des bases de données
- Renforcement du biais et discrimination
- Variable sensible absente mais prévisible

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

- Rapport Villani (2018) : *Discrimination Impact Assessment (DIA)*
- Monde académique (Zliobaité, 2017)
  Schématiquement trois niveaux de biais ou discrimination :
  1. Effet disproportionné : *Disparate* ou *adverse impact* (*Title vii Civil Rights Act*)
     - $DI = \frac{P(Y=1|S=0)}{P(Y=1|S=1)}$ par intervalle de confiance (Besse et al. 2008)
  2. Taux d'erreur conditionnels : reconnaissance faciale, santé, (De Arteaga et al. 2019)
  3. Égalité des cotes (*equality of odds*) : justice "prédictive" (Propublica vs. Compas)

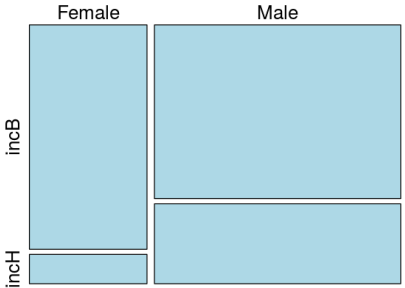
Entretien vidéo différé : modèle prédictif pour la pré-sélection de candidats sur la base du contenu verbal

Léo Hemamou[123], Grégory Wajntrob[1], Jean-Claude Martin[2], Chloé Clavel[3]
[1]EASYRECRUE, 3 bis rue de la Chaussée d'Antin, 75009 Paris
[2]LIMSI, CNRS, Université Paris-Saclay, Bât 508, rue John von Neumann, Campus Universitaire, F-91405 Orsay
[3]LTCI-CNRS, Telecom-Paristech, Université Paris-Saclay, 75013 Paris

- Données : transcription lexicale d'une vidéo & évaluation par recruteur humain
- $n = 305$ sélectionnés : favorable / réservé parmi 607
- *"Le score d'aire sous la courbe ROC (0.69) est bien supérieur à celui qui serait obtenu avec l'aléatoire (0.5)"*
- Aucune prise en compte du risque de discrimination contrairement aux USA (Raghavan et al. 2019)

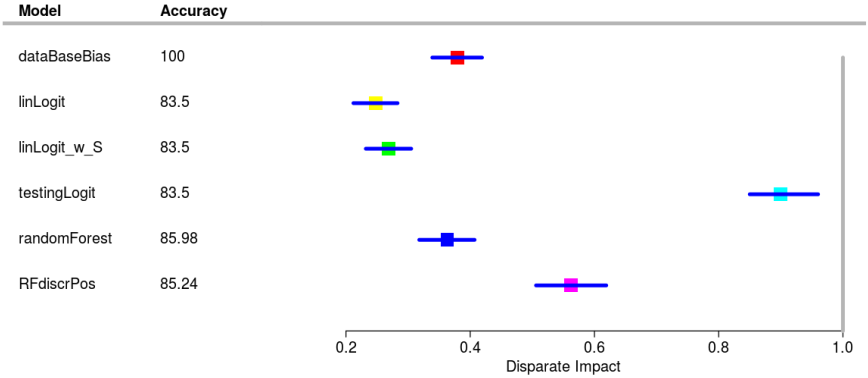
Cas d'Usage : *Adult Census Dataset*

- 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



$$DI = 0.37$$

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



Détection de la discrimination indirecte : DI vs. testing

Discrimination directe par *testing* (DARES) : $DI = 0.74 \quad P(DI \in [0.66, 0.83]) = 0.95$

Qualité des décisions & vide juridique

- Confiance envers une décision
- Algorithme d'apprentissage : qualité de décision et erreur de prévision
- Taux d'erreur de 3% en image *vs.* 30 à 40% pour le risque de récidive
- Considérant (71) du RGPD mais loi française muette
  Loi sur la 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

Santé : du buzz à la certification (Liu et al. 2019, *FDA*, HAS(?))

ARTIFICIAL INTELLIGENCE, DIAGNOSTICS, HEALTH TECH

Google's AI beats humans at detecting breast cancer — sometimes

A retrospective study published in Nature shows Google's DeepMind AI outperformed radiologists in detecting breast cancer. But it won't be replacing them anytime soon.

By ELISE REUTER

nature

Article | Published: 01 January 2020

International evaluation of an AI system for breast cancer screening

Scott Mayer McKinney, Marcin Sieniek, [...] Shravya Shetty

nature Subscribe

INNOVATIONS IN · 18 DECEMBER 2019

Rise of Robot Radiologists

Deep-learning algorithms are peering into MRIs and x-rays with unmatched vision, but who is to blame when they make a mistake?

Computer Science > Machine Learning

Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, Christopher Ré

(Submitted on 27 Sep 2019 (v1), last revised 15 Nov 2019 (this version, v2))

Confiance & Transparence : biais, explicabilité, robustesse

- Biais et discrimination : Norme ”réglementaire” d’une définition statistique de la discrimination indirecte ?
- Explicabilité *vs.* Intervention humaine (RGPD) : nommer les responsabilités
- Processus Qualité & Certification : boucle vertueuse (FDA, HAS)

En chantier

Auditabilité (Villani, 2018) : liste d'évaluation & contrôle des algorithmes (# AIPD) ?

- Normes : ANSI, IEEE, ISO ?
- Certification en santé : *Food and Drug Administration*, Haute Autorité de Santé
- Recherche très active dans l'industrie : DEEL, ANITI
  - Détecter, Corriger un biais discriminatoire :
    Sans pénaliser la précision *fair learning* et jusqu'à l'obtention d'un biais explicable
  - Explicabilité individuelle ou industrielle (certification)
  - Qualité et meilleur compromis

Références

- Barocas S. , Selbst A. (2016). Big Data's Disparate Impact, *California Law Review* (104), 671.
- Besse P. del Barrio E., Gordaliza P., Loubes J.-M. (2018). Confidence Intervals for testing Disparate Impact in Fair Learning, arXiv preprint arXiv :1807.06362.
- CNIL, Défenseur des Droits (2012). Mesurer pour progresser vers l'égalité des chances, Guide méthodologique à l'usage des acteurs de l'emploi.
- Challe L., Chareyron S., L'Horty Y., Petit P. (2020). Discrimination dans le recrutement des grandes entreprises: une approche multicanal, *rapport de recherche TEPP* 2020-01.
- Code of Federal Regulations (1978). Title 29 - Labor PART 1607—UNIFORM GUIDELINES ON EMPLOYEE SELECTION PROCEDURES (1978), CHAPTER XIV - EQUAL EMPLOYMENT OPPORTUNITY COMMISSION.
- Commission Européenne (2016). Règlement Général sur la Protection des Données.
- Commission Européenne (2018). Ethics guidelines for trustworthy AI.
- Commission Européenne (2020). White Paper on Artificial Intelligence: a European approach to excellence and trust.
- Défenseur des Droits, CNIL (2012). Mesurer pour progresser vers l'égalité des chances. Guide méthodologique à l'usage des acteurs de l'emploi.
- De-Arteaga M., Romanov A. et al. (2019). Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting, in FAT'19.
- Hemamou L., Wajntrob G., Martin J.-C., Clavel C. (2018). Entretien vidéo différé: modèle prédictif pour la pré-sélection de candidats sur la base du contenu verbal, Workshop sur les Affects, Compagnons Artificiels et Interactions.
- Raghavan M., Barocas S., Kleinberg J., Levy K. (2019) Mitigating bias in Algorithmic Hiring : Evaluating Claims and Practices, arXiv :1906.09208.
- Riach P.A., Rich J. (2002). Field Experiments of Discrimination in the Market Place, *The Economic Journal*, Vol. 112 (483), pp F480-F518.
- 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.
- Zliobaité I. (2017). Measuring discrimination in algorithmic decision making. *Data Min Knowl Disc* 31, 1060–1089. doi.org/10.1007/s10618-017-0506-1.