Éthique & IA au quotidien

Loyauté des Algorithmiques d'Apprentissage Automatique

PHILIPPE BESSE & CÉLINE CASTETS RENARD, AURÉLIEN GARIVIER, JEAN-MICHEL LOUBES

Université de Toulouse – INSA

Intelligence Artificielle (IA) au quotidien

- Pas de Science Fiction : transhumanisme, singularité technologique, lois d'Azimov
- Pas de Sociologie : destruction des emplois qualifiés, *big data big brother*
- Algorithmes de décision ou aide automatique à la décision
- Sous-ensemble de L'IA : apprentissage automatique (machine learning)
- Entraînés sur des bases de données : apprentissage statistique (*statistical learning*)
  - Risque de défaut de paiement, de comportement à risque, de rupture de contrat
  - Risque de récidive, de passage à l'acte
  - Risque de défaillance ou panne d'un système industriel
  - Diagonostic et bases d'images (*deep learning*)
  - Profilage professionnel (CV), publicitaire
  - ...
- Régression, k-ppv, Arbres, SVM, *random forest, boosting*, réseaux de neurones...

Battage médiatique de l'Intelligence Artificielle

- Convergence entre apprentissage automatique, données massives & Puissance de calcul, stockage
- Succès médiatisés de l'apprentissage profond (*deep learning*) Reconnaissance d'images, véhicules autonomes, jeu de go...
- Données confidentielles et fort impact personnel
- Enjeux sociétaux & financiers considérables

Pourquoi se préoccuper d'Éthique en IA ?

- Rapport Villani (03-2018) : *L'IA au service de l'Humain*
- Vide entre Lois et Technologies : responsabilité des acteurs
- Confiance et acceptabilité des nouvelles technologies
- Entreprises philanthropiques et altruistes ?


THE WALL STREET JOURNAL.

Subscribe Now | Sign In

SPECIAL OFFER: JOIN NOW

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate

Oil at One-Year High on Falling Stockpiles

U.S. Stocks Rise on Oil Rally, Bank Earnings

Platinum Partners' Flagship Hedge Fund Files for Bankruptcy

MARKETS

U.S. Government Uses Race Test for $80 Million in Payments

Checks are ready for minority borrowers allegedly discriminated against on Ally Financial auto loans

By ANNAMARIA ANDRIOTIS and RACHEL LOUISE ENSIGN

Updated Oct. 29, 2015 9:32 p.m. ET

Recommended Videos


c|net

Rechercher sur CNET

News Meilleurs Tests Culture

CNET France › News › Internet › Facebook plonge en bourse, Zuckerberg perd 16,8 milliards de dollars en deux

Facebook plonge en bourse, Zuckerberg perd 16,8 milliards de dollars en deux heures

Mark Zuckerberg a de nouveaux soucis, l'activité publicitaire de Facebook est en repli et l'action chute de 24% après la publication des résultats trimestriels en deçà des attentes. La fortune personnelle du patron aurait dégringolé de 16,8 milliards...

Cinq questions d'éthique

1. Propriété, confidentialité (*privacy*) des données personnelles et rôle de la CNIL (RGPD)
2. Entraves à la concurrence : moteurs de recherche, comparateurs *Pricing* automatique et compétition virtuelle (Ezrachi A., Stucke M. 2016)
3. Biais & Discrimination de décision algorithmiques
4. *Explicabilité*, transparence des algorithmes
5. Qualité des prévisions donc des décisions

Loyauté des algorithmes

- *Trustworthiness* : Mériter la confiance : fiabilité, crédibilité, non discriminatoire
- *Accountability* : responsabilité, capacité à rendre compte

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

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

Discrimination individuelle *vs.* de groupe

- Loi française et RGPD ont une approche individuelle de la discrimination
- *Quid* d'une forme implicite de discrimination :
  - biais structurel ou bases de données biaisées
  - variable sensible absente et prévisible
  - prévision auto-réalisatrice
- Rapport Villani & mesure de groupe de la discrimination :
  Discrimination Impact Assessment (DIA)
- Jurisprudence américaine (*Title vii 1964 Civil Rights Act*) & *disparate impact* (DI)
  Impact disproportionné pour un groupe $DI = \frac{P(Y=1|S=0)}{P(Y=1|S=1)}$
- Question : estimation du DI (cf. Besse et al. 2018 & wikistat)
- Autres mesures de discrimination
  - Taux d'erreur et représentativité de la base (e.g. reconnaissance faciale)
  - Égalité d'opportunité (Hardt et al. 2016 ; What-if-tools) ou
    Conditional Procedure Accuracy Equality (Besse et al. 2018 ; wikistat)
  - ... cf. Zliobaité (2015)

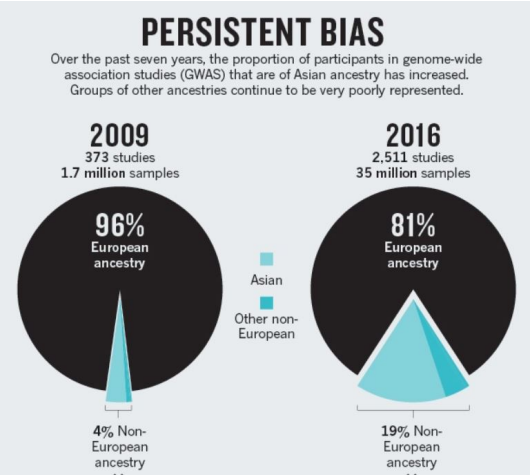
Testing et preuve d'une discrimination individuelle

- *SOS racisme* : pratique jugée déloyale mais reconnue
  - Arrêt de la Cours de Cassation (2012)
  - *article 225-3-1 code pénal* Les délits prévus par la présente section sont constitués... dès lors que la preuve de ce comportement est établie
- Discrimination à l'embauche :
  - Observatoire des Discriminations
  - DARES (Direction de l'Animation, des Études, de la Recherche et des Statistiques)
  - ISM Corum
- Accès à l'assurance, au crédit : L'Horty et al. (2017)

Biais en Santé

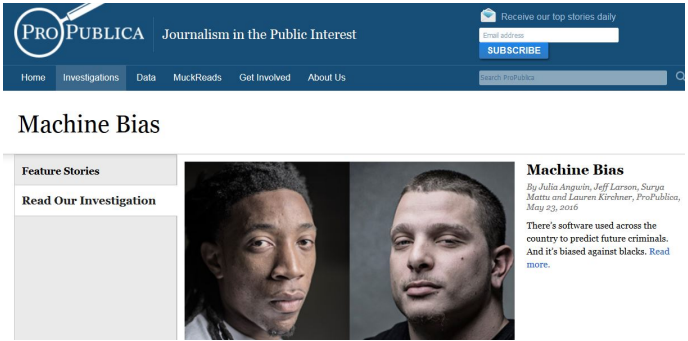
- Médecine de précision *vs.* de population
- Traitement personnalisé : sommes nous tous égaux ?
- *GWAS Genome Wide Association Studies*
  Bases d'associations pangénomiques
- Liaisons entre variants génétiques (SNPs) et traits phénotypiques
- Biais
  - Ethnique : population d'ascendance blanche européenne *Genomics is failing on diversity* (Popejoy et Fullerton, 2016)
  - Âge et environnement : bases transversales et pas longitudinales
  - Genre : Chang et al. (2014), Pulit et al. (2017)

GWAS : Biais ethnique



Popejoy et Fullerton (2016)

Justice prédictive : ProPublica *vs.* equivant (NorthPointe Inc.)



Angwin et al. (2016)

ProPublica vs. Equivant (NorthPointe Inc.)

- Absence de discrimination selon NorthPointe Inc.
  - Distributions des scores (m_1 et m_2) donc DI similaires
  - Taux d'erreur $(FN + FP)/n$ similaires
- Discrimination selon ProPublica : *Conditional Procedure Accuracy Equality*

Matrice de confusion

Observation / Récidive	Score: Faible	Score: Élevé	
Oui	FN	VP	q_1
Non	VN	FP	q_2
	m_1	m_2	n

  - Taux de faux positifs$= FP/q_2$
    afro-américain (45%) *vs.* caucasiens (25%)
- Taux de récidive afro-américain plus élevé (Chouldechova, 2016)
- Taux d'erreur très élevé (40%)

Rapport Villani (2018)

L"ouverture des boîtes noires" de l'IA est un "enjeu démocratique"

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

Jongler entre RGPD et lois nationales

- 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
- Réforme de la loi informatique et libertés LIL 3 : loi n° 2018-493 du 20 juin 2018
- Code des relations entre le public et les administrations
- Conseil Constitutionnel Décision n° 2018-765 DC du 12 juin 2018

Droit à l'explication en France (LIL3)

- Aucune décision de justice impliquant une appréciation sur le comportement d'une personne ne peut avoir pour fondement un traitement automatisé de données à caractère personnel destiné à évaluer certains aspects de la personnalité de cette personne
- Aucune décision produisant des effets juridiques à l'égard d'une personne ou l'affectant de manière significative ne peut être prise sur le seul fondement d'un traitement automatisé de données à caractère personnel, y compris le profilage
- Exception : à condition que les règles définissant le traitement ainsi que les principales caractéristiques de sa mise en œuvre soient communiquées par le responsable de traitement à l'intéressé s'il en fait la demande (sauf secret protégé par la loi)
- Loi n° 2018-493 Le responsable de traitement s'assure de la maîtrise du traitement algorithmique et de ses évolutions afin de pouvoir expliquer, en détail et sous une forme intelligible, à la personne concernée la manière dont le traitement a été mis en œuvre à son égard
- En résumé : identifier une responsabilité humaine plus qu'un droit à l'explication ou l'interprétation.

Décret du 16/03/2017 Art. R. 311-3-1-2.

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 :

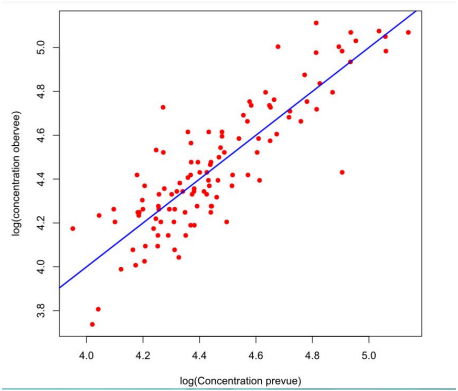
1. Le degré et le mode de contribution du traitement algorithmique à la prise de décision ;
2. Les données traitées et leurs sources ;
3. Les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé ;
4. Les opérations effectuées par le traitement.

Décision du Conseil Constitutionnel du n°2018-765 DC du 12 juin 2018, Pt 71

En dernier lieu, le responsable du traitement doit s'assurer de la maîtrise du traitement algorithmique et de ses évolutions afin de pouvoir expliquer, en détail et sous une forme intelligible, à la personne concernée la manière dont le traitement a été mis en œuvre à son égard. Il en résulte que ne peuvent être utilisés, comme fondement exclusif d'une décision administrative individuelle, des algorithmes susceptibles de réviser eux-mêmes les règles qu'ils appliquent, sans le contrôle et la validation du responsable du traitement

Explicabilité : interprétation d'un modèle linéaire du "siècle dernier"

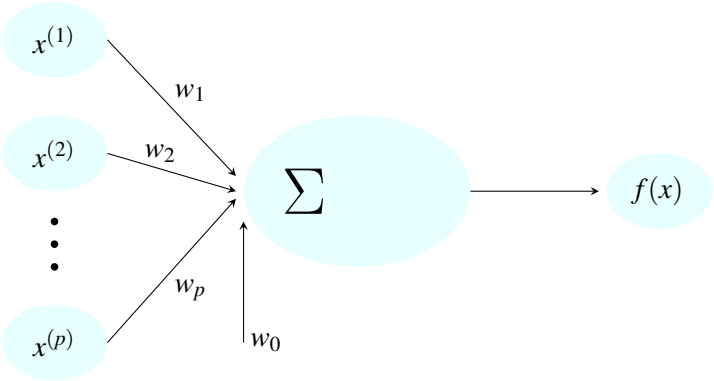
Prévoir la Concentration en Ozone



$$\begin{aligned}\log(\text{ConcODemain}) &= 2,4 + 0,35 \times \log(\text{ConcOJour}) + 0,05 \times \text{Sec} + \\ &+ 0,03 \times \text{T12} - 0,03 \times \text{Ne9} + 0.1 \times \text{Vx9}\end{aligned}$$

Modèle / Neurone Linéaire

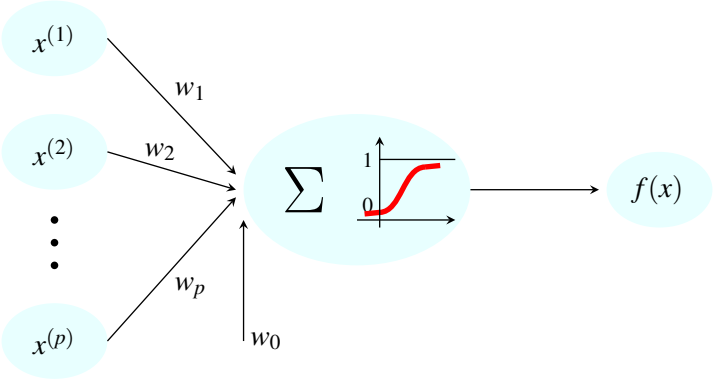
Modéliser / prévoir une variable quantitative



$$f(x) = w_0 + w_1 \times x^{(1)} + w_2 \times x^{(2)} + \cdots + w_p \times x^{(p)}$$

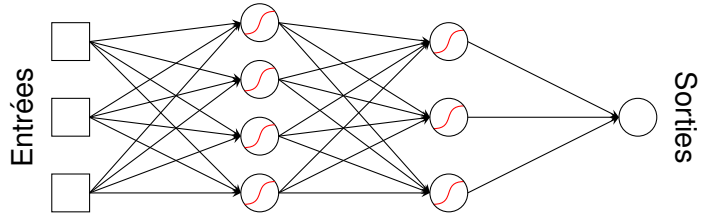
Modèle / **Neurone** *logistique*

Variable binaire : Maladie, Panne, Départ, Faillite...



Exemple en épidémiologie : interpréter, évaluer les facteurs de risque

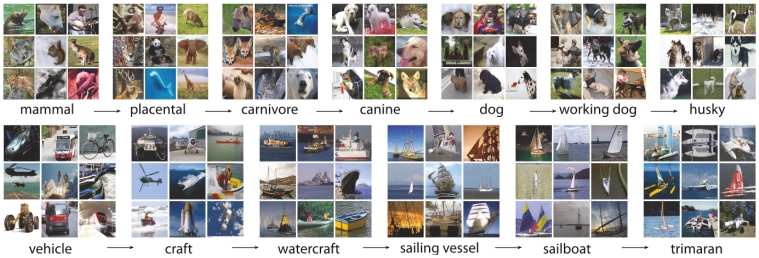
Explicabilité : réseau de neurones : (Perceptron)



- Interprétation impossible : *Boîte Noire*
- Idem pour k-p.p.v., SVM, *boosting, random forest*...
- Quelle explication ?

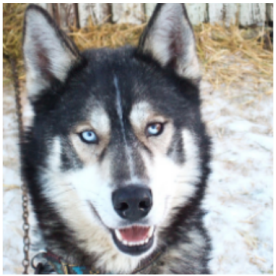
Explicabilité : Deep Learning

Exemple : base de données ImageNet :
15 millions d'images, 22000 catégories
2016 : 152 couches et mieux que l'expert humain



Ouvrir les boîtes noirs pour l'acceptabilité (robustesse)

- Enjeux majeurs sociétaux et industriels (projet franco-canadien DEEL)
- Recherche très active
- Interprétation impossible, quelle explication ?
- Explication fonctionnelle, détection d'artefact *vs.* explication individuelle
- Approximation locale linéaire : package LIME de R, What-if-tools & TensorFlow



Éthique ”industrielle”

Amazon, Google, Facebook, IBM, Microsoft, Apple...

1. We believe that artificial intelligence technologies hold great promise for raising the quality of people's lives and can be leveraged to help humanity address important global challenges such as climate change, food, inequality, health, and education.

...

7. We believe that it is important for the operation of AI systems to be understandable and interpretable by people, for purposes of explaining the technology.

Loi et qualité des décisions

- Qualité et confiance envers une décision
- Apprentissage statistique : qualité de décision = qualité de prévision
- RGPD et lois françaises muettes
- Loi sur la publication des sondages d'opinion
- *L'efficacité prédictive sera d'autant plus grande qu'elle sera le fruit de l'agrégation de données massives* in *La Gouvernementalité Algorithmique* (Rouvroy et Berns, 2013)
- **Vrai** et **Faux**
- Taux d'erreur de 3% en image vs. 30 à 40% pour le risque de récidive
- Ne pas confondre estimation / prévision d'une moyenne (*loi des grands nombres*) et celle d'un processus physique, d'un comportement individuel humain
- Domaine d'application : pertinence et représentativité de la base d'apprentissage (*Google flue trend 2008-2015*)
- Obligation de moyen mais pas de résultat

IA au quotidien et éthique

- Biais et discrimination : *DIA* ou définition statistique de la discrimination de groupe
- Le droit à l'Explication oblige à une responsabilité humaine, pas à une interprétation / explication
  indispensable mais spécifique à chaque domaine d'application
- Qualité de décision / prévision : vide juridique sur l'obligation d'informer (cf.sondages)

Recherche très active

- Corriger un biais (*fair learning*) et discrimination positive
- Explicabilité industrielle ou individuelle
- Qualité et meilleur compromis

Éthique individuelle

- *Hippocratic oath*
- Code of ethics
- Serments d'Hippocrate du *data scientist*
- Charte européenne et américaine du statisticien
- Consentement libre et éclairé !

Éthique collective

- CNIL, Autorité de la concurrence
- Contre pouvoir citoyens
- Audit des algorithmes (Rapport Villani)
- Proposition de labels

Références

- Angwin J., Larson J., Mattu S., Kirchner L. (2016). How we analyzed the compas recidivism algorithm. ProPublica, en ligne consulté le 28/04/2017.
- 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.
- 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).
- Chouldechova A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, arXiv pre-print.
- Hardt M., Price E., Srebro N. (2016). Equality of Opportunity in Supervised Learning, *30th Conference on Neural Information Processing System (NIPS)*.
- L'Horty Y., Bunel M., Mbaye S., Petit P. , du Parquet L. (2017). Discriminations dans l'accès à la banque et à l'assurance : Les enseignements de trois testings, *TEPP Research Report 2017-08*, TEPP.
- Popejoy A., Fullerton S. (2016). Genomics is failing on diversity, *Nature*, 538, 161-164.
- Pulit S., Karaderi T., Lindgren C. (2017). Sexual dimorphisms in genetic loci linked to body fat distribution, *Bioscience Report*, 37(1).
- Rouvroy A., Berns T. (2013). Gouvernementalité algorithmique et perspectives d'émancipation, *Réseaux*, 177, 163-196.
- Zliobaité I. (2015). A survey on measuring indirect discrimination in machine learning. arXiv pre-print.


MOI JE DIS : L'INTELLIGENCE ARTIFICIELLE, C'EST BIEN
SAUF QUAND ELLE SE MET AU SERVICE DE LA CONNERIE NATURELLE