



CATÓLICA
FACULTY
OF LAW

ESCOLA DO PORTO



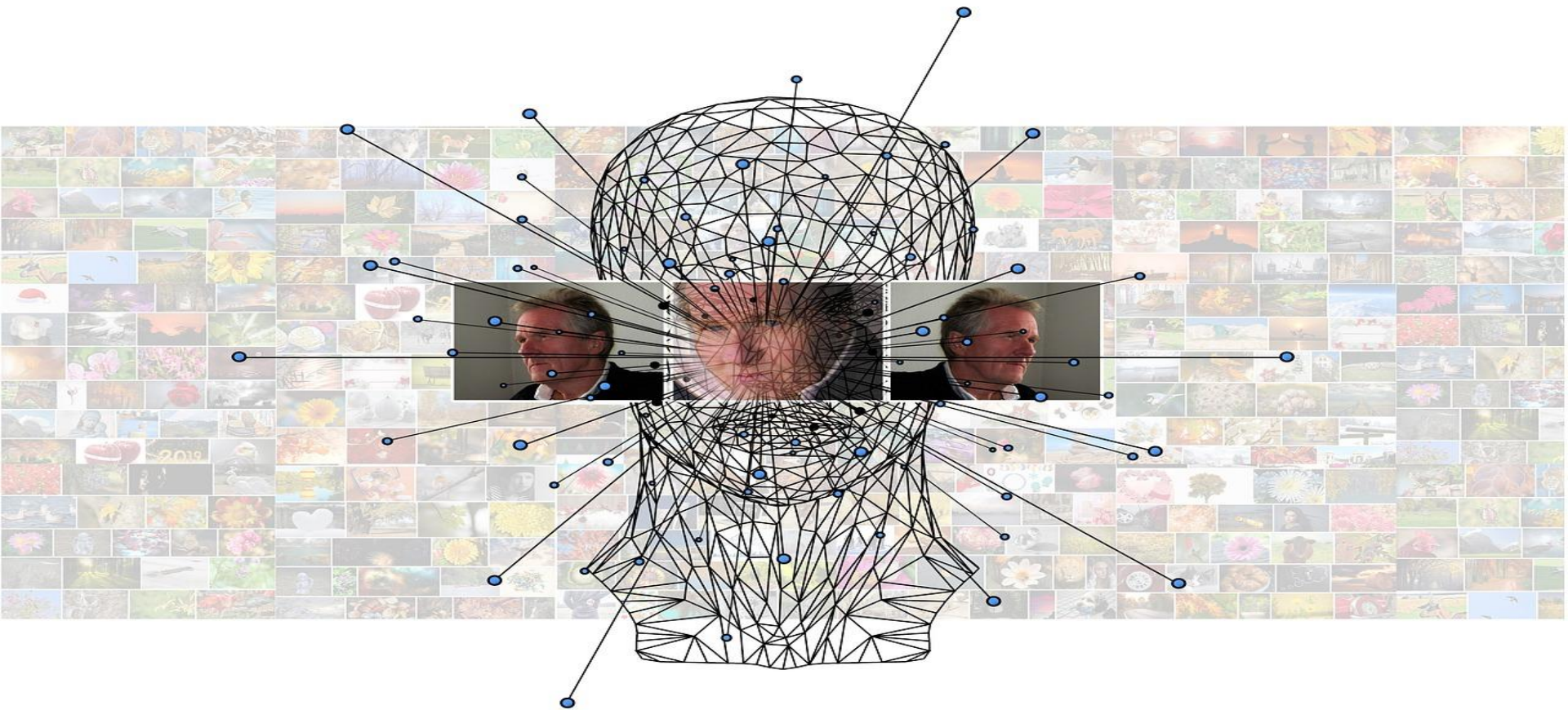
Funded by
the European Union

La fin de la tromperie ? - Contrecarrer la discrimination algorithmique à l'ère du numérique

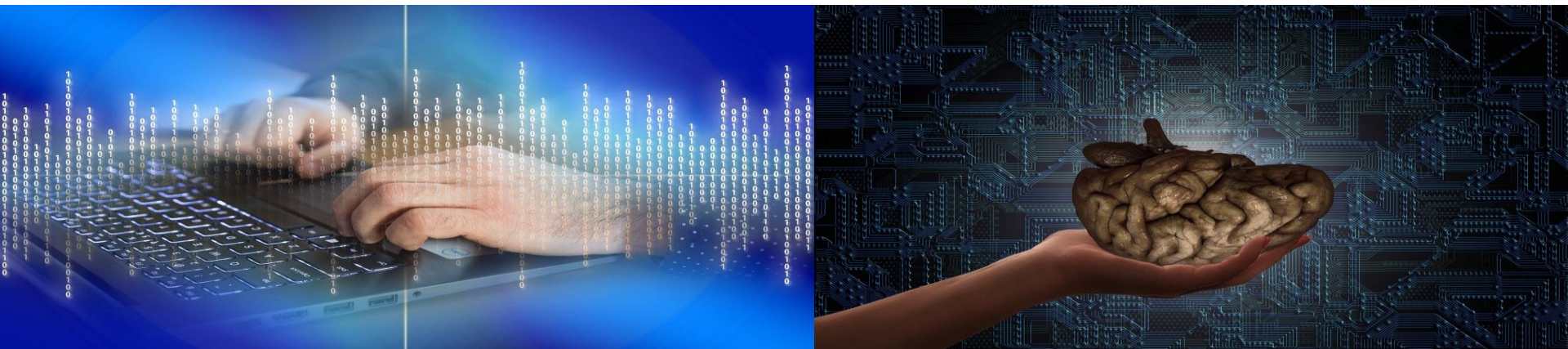
Catarina Santos Botelho
cbotelho@ucp.pt

Qu'est-ce qu'un algorithme ?

- Les algorithmes sont des procédures codées par lesquelles les données d'entrée, suivant une séquence logique et mathématique, se transforment en sortie.



- À ce stade, les algorithmes sont complexes et défient l'entendement humain.
- Les algorithmes peuvent être classés en algorithmes basés sur la connaissance et en algorithmes d'apprentissage automatique ("machine-learning" = **ML**).
- Les systèmes à **base de connaissances** sont les algorithmes traditionnels à organigramme qui nécessitent un codage manuel.
- Les **ML** sont des algorithmes qui alimentent de grandes quantités de données avec des variables de sortie pour que l'algorithme "apprenne sans être explicitement programmé".



Pourquoi ce sujet est-il important ?

Les processus décisionnels automatisés sont une réalité croissante



- dispositions relatives aux soins de santé,
- allocation des vaccins,
- peines de prison,
- examen de la police,
- sélection de l'audit de l'IRS,
- dépouillement des votes,
- détermination des prêts,
- demandes d'emploi,
- répartition du travail,
- demandes de prestations sociales,
- octroi de visas d'immigration,
- sélections pour l'entrée à l'école maternelle ou à l'université,
- prix,
- surveillance des consommateurs,
- marketing microciblé.



Quelles sont les grandes questions ?

- Qu'en est-il de la démocratie et de la protection des droits de l'homme dans une nouvelle "architecture socio-technique de l'internet" ?
- Les algorithmes peuvent-ils intensifier le rejet de la démocratie ?
- L'intervention de l'homme dans les processus décisionnels est-elle toujours essentielle pour la protection des droits de l'homme ou s'agit-il d'une précaution inutile ?



Discrimination directe / discrimination indirecte

- **Direct** : lorsqu'une personne est traitée moins favorablement qu'une autre en raison d'une caractéristique protégée dans un secteur protégé.
- **Indirect** : lorsqu'une disposition, un critère ou une pratique apparemment neutre désavantage une personne appartenant à un groupe protégé sans justification acceptable.



Traitement différencié / Impact différencié

- Titre VII de la loi sur les droits civils de 1964 (États-Unis)
- *Traitement différencié* : lorsque l'employeur traite le salarié différemment en raison d'une caractéristique protégée.
- *Impact différencié* : une pratique qui, à première vue, semble neutre et non discriminatoire, mais qui, en réalité, désavantage de manière disproportionnée un groupe protégé.



Critères relatifs aux groupes protégés

- Les groupes protégés sont-ils fixes ou d'autres "axes d'identité socialement saillants" peuvent-ils être ajoutés ?
- Le statut socio-économique ou le niveau d'éducation peuvent-ils être pris en compte dans les propositions anti-discrimination ? Et qu'en est-il de l'obésité, des maladies chroniques ou du type de navigateur ?
- La conception de la législation antidiscrimination, en particulier les listes de motifs et/ou d'exceptions, doit-elle être ouverte, fermée ou hybride ?



Biais concernant les algorithmes de ML

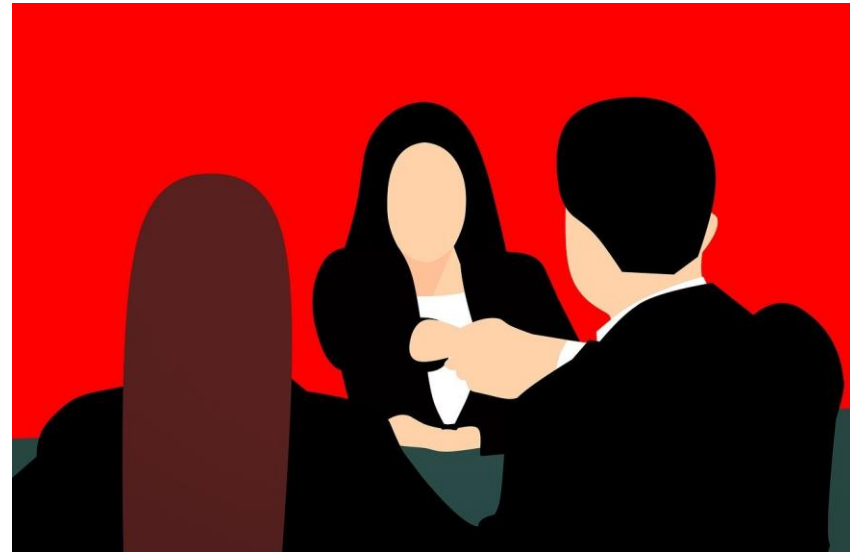
:

- *un biais conséquent*, lorsque les créateurs de l'algorithme sont intentionnellement ou non biaisés, et que ce biais se reproduira dans le mécanisme de traitement des données ;
- *les données d'échantillonnage biaisées*, dans lesquelles l'échantillon fourni ou les données d'apprentissage sont biaisés ;
- *le biais de résultat* ou de *cercle vicieux*, qui se produit même lorsque l'algorithme est entraîné avec des données représentatives, en raison des inégalités sociales intégrées dans les données.

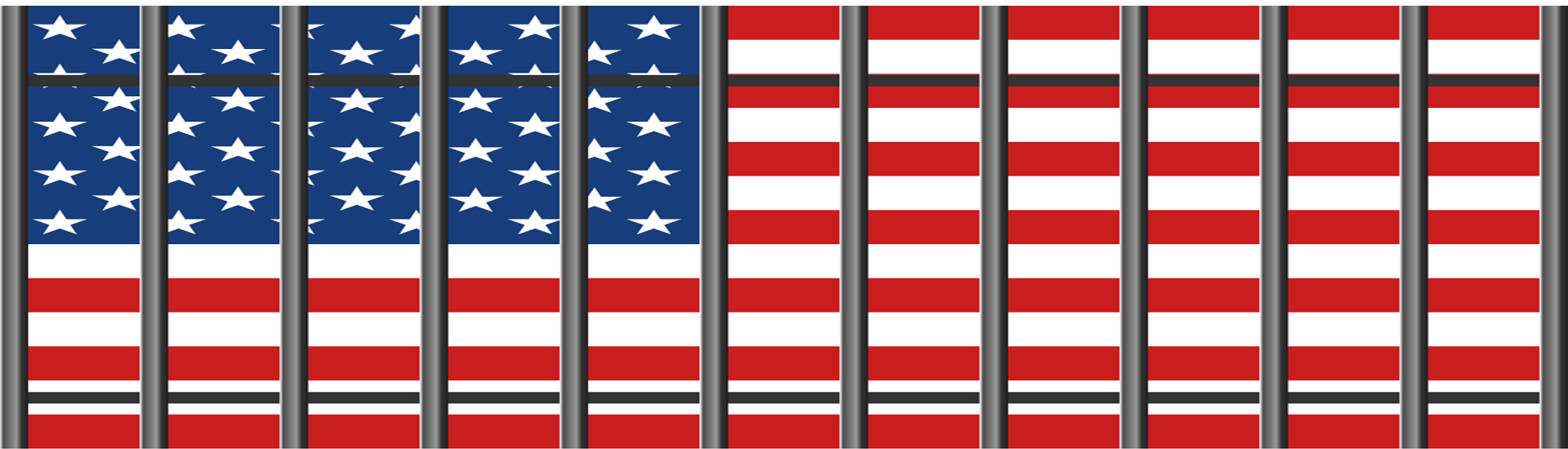


Exemples d'impact différencié algorithmique et de discrimination statistique

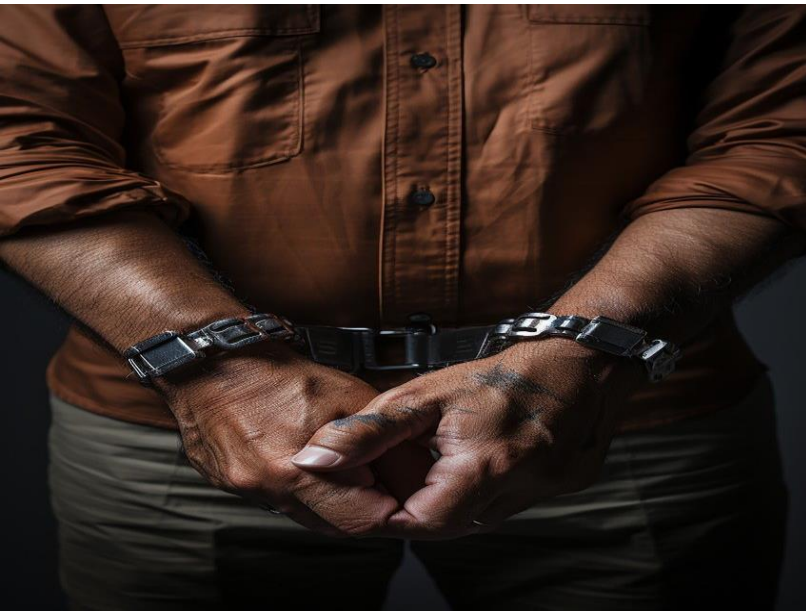
- ***Affaire Amazon :***
- Amazon a développé un système de ML pour identifier les *futurs employés des développeurs de logiciels*. L'algorithme a classé les candidats à l'emploi en fonction de la productivité attendue. Comme l'algorithme a été formé à l'aide des données d'embauche existantes - qui reflètent la prédominance masculine dans l'industrie technologique - les candidats dont le curriculum vitae comporte des indicateurs féminins (par exemple, club de tennis féminin) ont été discriminés, quelles que soient leurs qualifications pour le poste.



- **Le cas COMPAS (Correctional *Offender Management Profiling for Alternative Sanctions*) :**
- À partir des données d'un comté américain, une société a mis au point un outil d'évaluation des risques, le COMPAS, pour prédire la récidive des détenus et des prévenus.
- Pour la question "cette personne va-t-elle récidiver si elle bénéficie d'une libération conditionnelle ?", l'algorithme a comparé les caractéristiques de la personne avec celles des personnes qui ont récidivé pendant la libération



- Le problème signalé par ProPublica en 2016 était que le COMPAS utilisait sa prédiction de réarrestation comme un substitut de la récidive.
- Le COMPAS a été accusé de produire des résultats biaisés sur le plan racial, car les faux positifs (qui prévoient un risque élevé de récidive alors que le délinquant ne récidive pas) pour les prévenus afro-américains étaient deux fois plus importants que les faux positifs pour les prévenus d'origine caucasienne.



L'illusion de l'objectivité algorithmique

- "L'aversion pour les algorithmes / L'IA en tant que monstre

contre

- Super-pouvoirs herculéens / champions de l'objectivité



Discrimination algorithmique endogène

- La discrimination réside dans la conception même de l'algorithme.
- Plus de transparence : empêcher la migration de décisions humaines discriminatoires vers des décisions algorithmiques.
- La transparence est-elle le nouveau piège du consentement ?



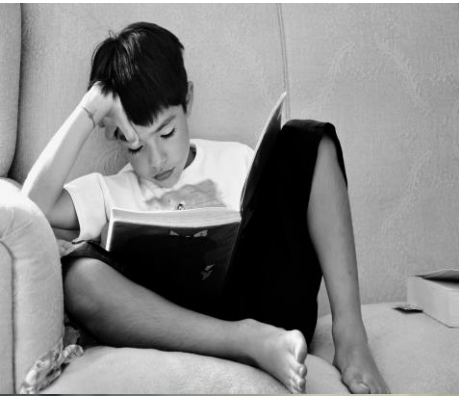
Discrimination algorithmique exogène

- Le problème n'est pas la "boîte noire" elle-même, mais le monde réel dans lequel les algorithmes fonctionnent.
- La discrimination est le résultat de préjugés implicites, de distorsions statistiques ou de représentations historiques erronées de groupes qui se sont frayé un chemin dans des ML supposées aseptisées.
- L'idée est de "passer d'algorithmes simplement *non discriminatoires* à des algorithmes activement *antidiscriminatoires*" (Bornstein).



L'égalité réelle

- ***La discrimination positive algorithmique*** peut être réalisée par plusieurs mesures, telles que l'incorporation de mesures d'équité dans les processus de sélection des modèles, ou par l'équité de groupe en utilisant des approches de "parité démographique" ou de "parité statistique".



Régulation et débiaisement des algorithmes

- Les algorithmes étant incompréhensibles pour les humains, un effort supplémentaire est nécessaire pour découvrir les résultats discriminatoires.
- Dans ces conditions, les décideurs politiques, les tribunaux, les agences, les autorités publiques et les employeurs ne devraient pas adopter une position romantique consistant à faire confiance aux algorithmes de manière excessive et sans esprit critique.
- Une vigilance active est donc fortement recommandée.



1. Contrôle algorithmique



- L'Acte européen sur l'IA vise à mettre en place une IA éthique et centrée sur l'humain, dans laquelle les résultats des systèmes d'IA sont traçables, explicables et transparents.
- Les systèmes d'IA sont classés en fonction de leur niveau de risque potentiel (article 5) :
 - Interdit a priori
 - Risque élevé
 - Faible risque

- Test de proportionnalité.
- Pour minimiser l'impact potentiel des systèmes d'IA à haut risque, les déployeurs doivent procéder à des **évaluations de l'impact sur les droits fondamentaux (FRIA)** avant de déployer un système d'IA à haut risque (article 27).



Article 6 (loi européenne sur l'IA)

3. Par dérogation au paragraphe 2, un système d'IA n'est pas considéré comme étant à haut risque s'il ne présente pas **de risque significatif** d'atteinte à la santé, à la sécurité ou aux droits fondamentaux des personnes physiques, y compris en n'influençant pas sensiblement le résultat de la prise de décision. C'est le cas lorsqu'une ou plusieurs des conditions suivantes sont remplies :

(a) le système d'intelligence artificielle est destiné à exécuter une tâche procédurale limitée ;

(b) le système d'IA est destiné à améliorer le résultat d'une activité humaine déjà réalisée ;

(c) le système d'IA est destiné à détecter des modèles de prise de décision ou des écarts par rapport à des modèles de prise de décision antérieurs et n'est pas destiné à remplacer ou à influencer l'évaluation humaine effectuée précédemment, sans contrôle humain approprié ; ou

(d) le système d'IA est destiné à effectuer une tâche préparatoire à une évaluation pertinente aux fins des cas d'utilisation énumérés à l'annexe III.

2. Le blocage de l'information : "L'équité par l'aveuglement ?"

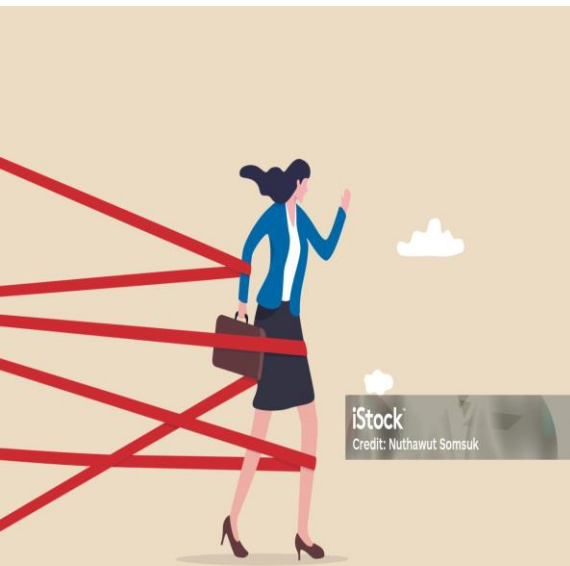
- Le blocage des informations personnelles concernant les catégories protégées dans les données de formation empêche-t-il les résultats discriminatoires ?
- Pour obtenir des résultats sans distinction de race, de handicap, de sexe, de revenu ou de code postal, il n'est peut-être pas suffisant (ni même judicieux) de simplement supprimer les caractéristiques sensibles.



- Les systèmes de ML est excellent dans la "détection de modèles, ainsi que dans l'identification et l'exploitation de substituts non évidents pour l'omission de variables sensibles".



- Comment pouvons-nous divulguer "des couches et des substituts", et comment savoir quels *substituts* (substituts fonctionnels) contiennent des informations précieuses ?
- Un autre facteur de complication est que la parité en matière de race et de sexe peut encore négliger les *différences intersectionnelles* au sein de ces groupes selon d'autres dimensions, telles que les handicaps, l'âge ou la nationalité.



3. Réguler et façonner les données par la conception algorithmique

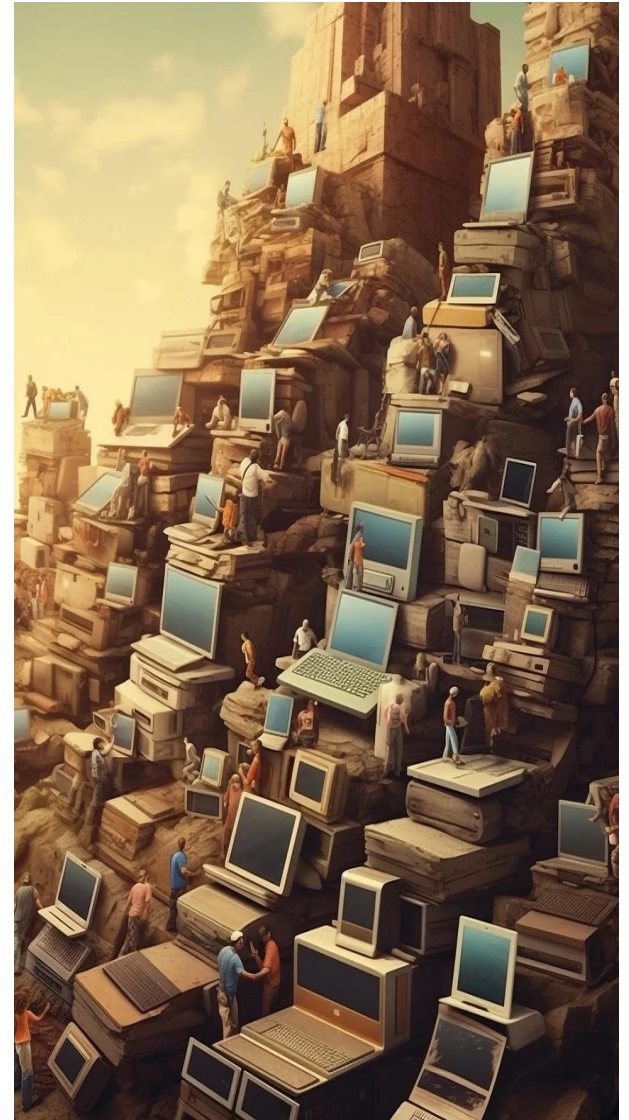
- Le règlement général sur la protection des données (RGPD) exige que les activités de traitement des données soient transparentes (considérants 39, 58 et 78 du RGPD et articles 5, paragraphe 1, point a), et 12 du RGPD), accorde le droit d'accès et le droit d'être informé de l'existence d'une décision automatisée (articles 13 à 15), ainsi que le droit de ne pas être soumis uniquement à une décision automatisée (article 22), et a introduit l'analyse d'impact relative à la protection des données, afin d'évaluer les risques pour les droits et libertés des personnes concernées et les risques liés à l'activité de traitement (article 35). Toutefois, ce droit ne s'applique pas lorsque la personne a donné son consentement explicite à la décision automatisée (article 22, paragraphe 2, point b), du RGPD).



À ce stade, il est plus important de régler les données qui alimentent les algorithmes que de régler les algorithmes.

L'objectif est de recueillir des "données équitables" susceptibles de contrebalancer les préjugés de la société.

Au lieu de bloquer les données, certains chercheurs proposent de les *modifier*. Le prétraitement des données peut être réalisé en ajustant les données d'entrée de manière générale ou en ajustant la variable cible (par exemple, le sexe).



4. Stratégies ex post de réparation des discriminations : droits individuels ou droits collectifs ?

- les tribunaux (par le biais du contrôle juridictionnel)
- les autorités chargées de la protection des données (par exemple, par le biais d'évaluations de l'impact sur la protection des données)
- "La FDA (Administration américaine "Food and drugs administration) pour les algorithms" ?
- "Un droit individuel à contester les décisions de l'IA" ?



Les angles morts des algorithmes et l'importance d'une approche centrée sur l'humain

- Existe-t-il un "droit à une décision humaine" ?
- L'IA peut imiter et même surpasser certaines parties de l'intellect humain. Néanmoins, le raisonnement inductif humain, la cognition humaine et la présence humaine dans l'exercice d'un *pouvoir discrétionnaire* (différent de l'arbitraire) ou d'un jugement au cas par cas restent nécessaires.



Pourquoi ?

- Parce que seule la personne humaine peut appréhender pleinement la réalité. L'IA exécute des opérations logiques qui sont des abstractions de la réalité (le monde extérieur ou le monde réel). Ce faisant, l'IA crée sa propre réalité au sein de "la réalité", bien qu'il s'agisse toujours d'une réalité restreinte qui ne peut ontologiquement atteindre son essence. Au contraire, les humains *font l'expérience de la réalité et l'embrassent.*



- Il ne s'agit pas d'une compétition futile entre les humains et les machines. Nous n'avons pas besoin de l'intervention humaine dans l'IA parce que "nous, les humains", sommes meilleurs.
- Nous avons besoin de conseils humains parce que nous ne pouvons pas faire l'économie de l'humanité, de l'éthique et de la morale dans les processus de décision publics et privés susceptibles d'affecter profondément nos vies.

