De l'éthique de l'IA à la future régulation de l'IA act: principes et étude des pratiques

Christine Balagué

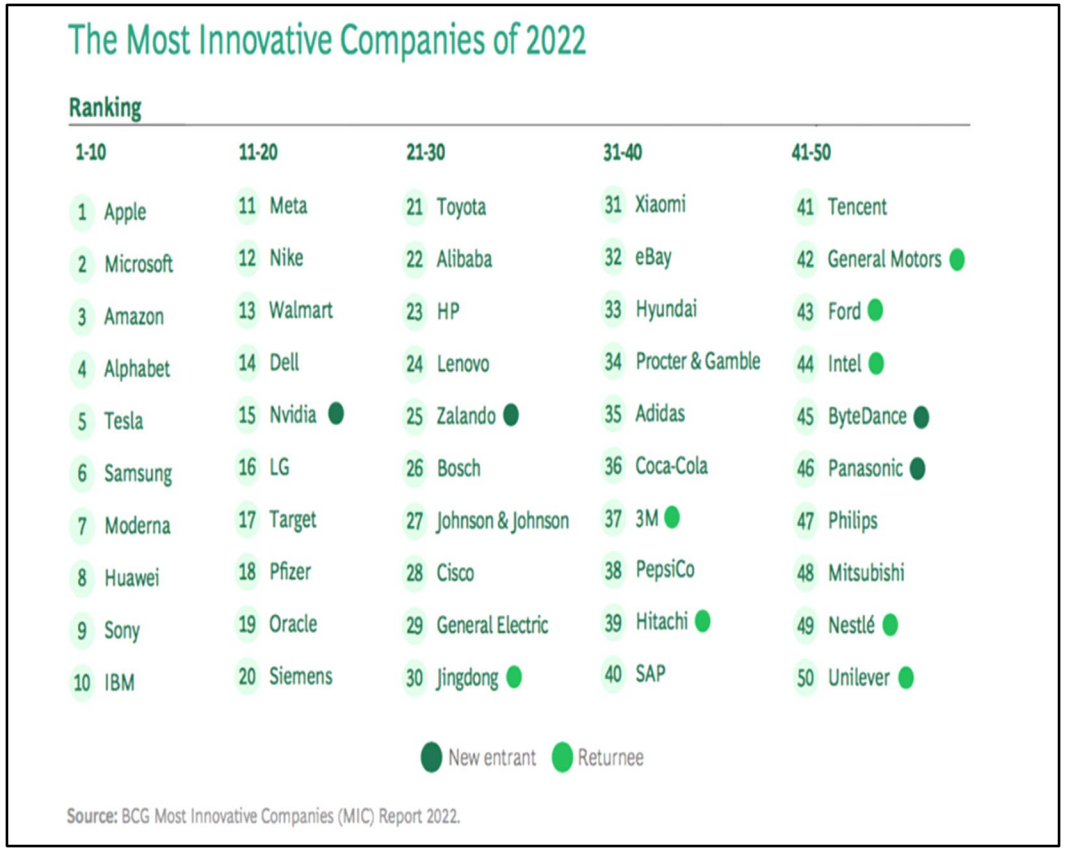
Professeur des Universités à IMT-BS, et fondatrice du réseau de recherche Good in Tech

Institut Mines-Télécom

FONDATION DU RISQUE Louis Bachelier

SciencesPo

Opportunités et risques de l'IA



The Most Innovative Companies of 2022

Ranking

1-10	11-20	21-30	31-40	41-50
1 Apple	11 Meta	21 Toyota	31 Xiaomi	41 Tencent
2 Microsoft	12 Nike	22 Alibaba	32 eBay	42 General Motors ●
3 Amazon	13 Walmart	23 HP	33 Hyundai	43 Ford ●
4 Alphabet	14 Dell	24 Lenovo	34 Procter & Gamble	44 Intel ●
5 Tesla	15 Nvidia ●	25 Zalando ●	35 Adidas	45 ByteDance ●
6 Samsung	16 LG	26 Bosch	36 Coca-Cola	46 Panasonic ●
7 Moderna	17 Target	27 Johnson & Johnson	37 3M ●	47 Philips
8 Huawei	18 Pfizer	28 Cisco	38 PepsiCo	48 Mitsubishi
9 Sony	19 Oracle	29 General Electric	39 Hitachi ●	49 Nestlé ●
10 IBM	20 Siemens	30 Jingdong ●	40 SAP	50 Unilever ●

● New entrant ● Returnee

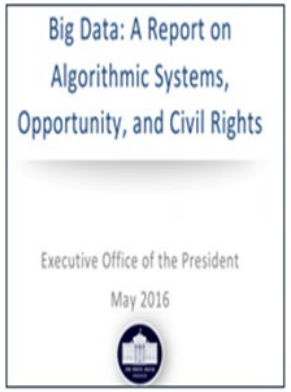
Source: BCG Most Innovative Companies (MIC) Report 2022.

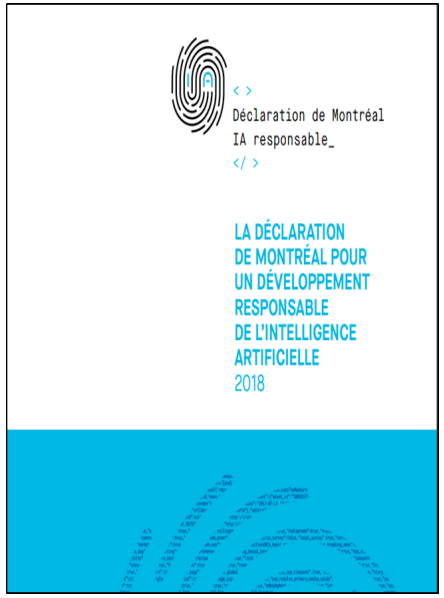
MAAMA R&D expenditures in 2021*:*
149 billions of dollars, more than the Pentagone
(source: *The Economist*)

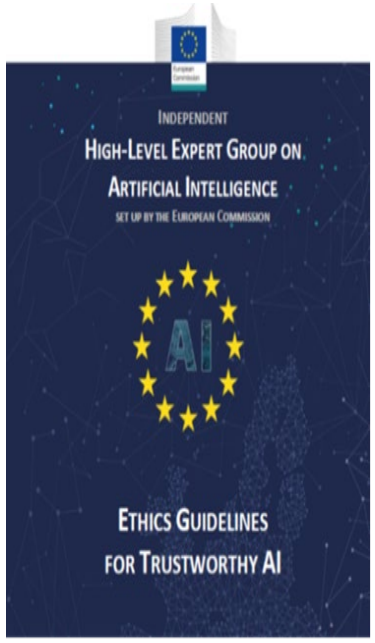
Risks & Dark side of AI (Wirtz et al., 2022)

- Technological, data and analytical AI risks (data bias, violation of privacy, vulnerability to attacks, …)
- Informational and communicational AI risks (manipulation, fake news, censorship, …)
- Economic AI risks (disruption of labour market, automation effects,…)
- Social risks (unemployment, social discrimination, surveillance, …)
- Ethical risks (harm to humans, unfairness,…)
- Legal and regulatory AI risks (who will compensate the victims, unforeseen behaviour of AI, wrong regulation,...)

Pourquoi parler de l'éthique de l'IA?
Un enjeu national et international







Principaux enjeux éthiques de l'IA

Biais et discrimination

Opacité et manipulation

Enfermement des individus vs autonomie

Prise de décision avec des algorithmes porteurs d'opinion

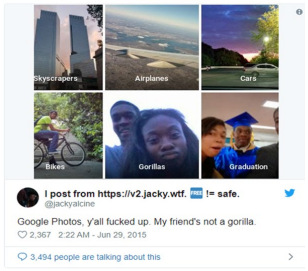
Biais des données:

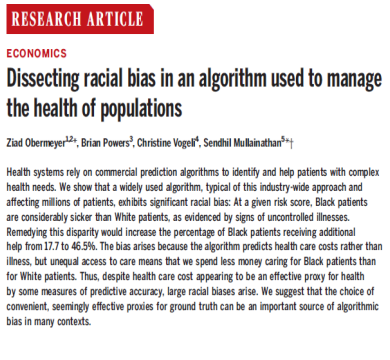
- Biais « Garbage in, garbage out »
- Biais de variable omise
- Biais de sélection
- Biais d'endogénéité

Sécurité et stockage des données

Anonymisation et ré-identification

Discrimination (ex race et genre) par les algorithmes (ex: santé, recrutement, recherche images):




RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

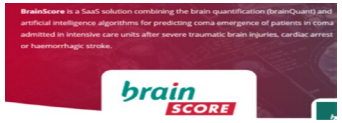
Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

Opacité et manipulation potentielle des moteurs de recherche et des RS:

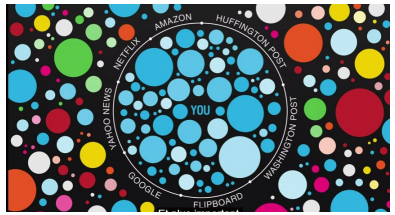
R. Epstein & R.E. Robertson "The search engine manipulation effect (SEME) and its possible impact on the outcome of elections", PNAS, 112, E4512-21, 2015

Opacité des algorithmes et prise de décision:



Opacité des algorithmes et information (deep fakes, fake news)

Bulles filtrantes sur Facebook:

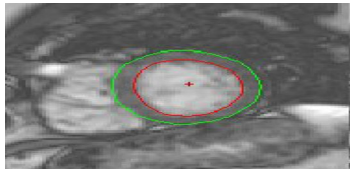


Enfermement par des algorithmes de recommandation:

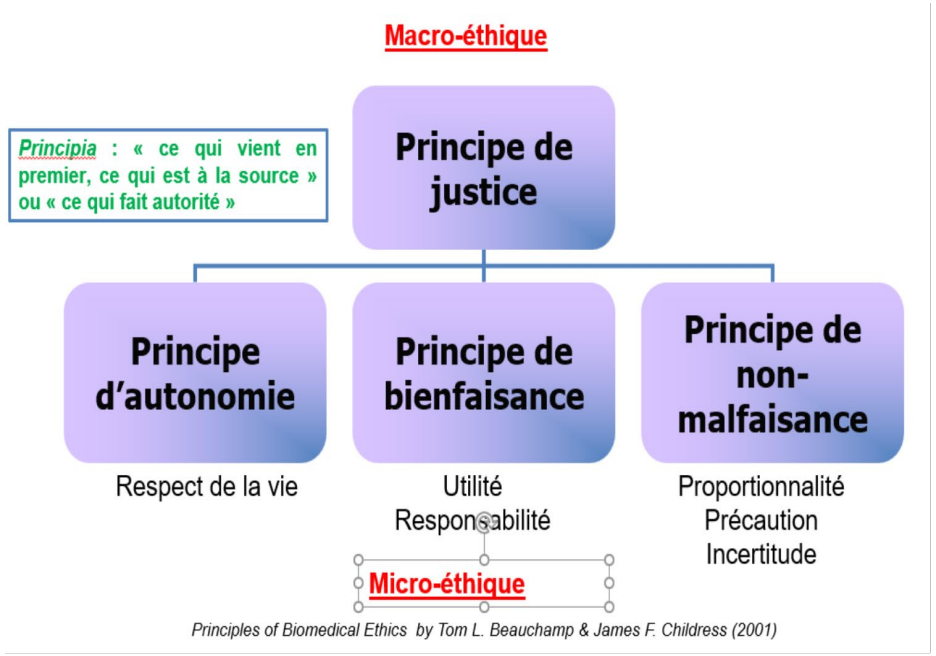
Reproduction des stéréotypes

Minimisation faux positifs ou faux négatifs: ex imagerie médicale

Over 75% of what people watch comes from a recommendation



Phase 1: l'éthique de l'IA fondée sur des grands principes



Principles of Biomedical Ethics by Tom L. Beauchamp & James F. Childress (2001)

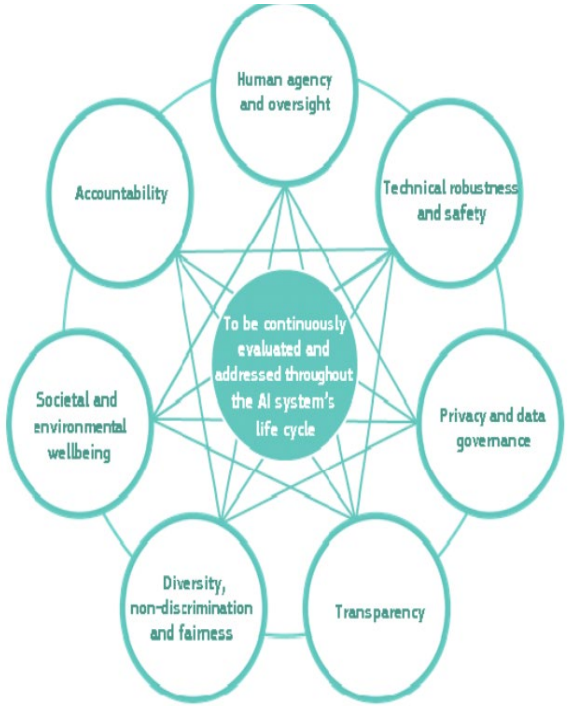


Figure 2: Interrelationship of the seven requirements: all are of equal importance, support each other, and should be implemented and evaluated throughout the AI system's lifecycle

- Critique de l'éthique fondée sur des grands principes:
  - Modèle théorique difficilement applicable
  - Pas de prise en compte des particularités des modèles
  - N'intègre pas le contexte de développement
  - Effets induits liés à la prévalence de mesures top-down (Mittelstadt 2019, Powers 2020)

Phase 2: l'éthique de l'IA par des réponses techniques: le cas du fair machine learning

Fairness Metrics	Definition
Statistical parity	Probability of being classified with the favorable label is independent of group membership
Disparate impact	Ratio of probabilities of being classified with the favorable label between protected and unprotected groups is close to one
Equalized odds	Both false positive rates and true positive rates for protected and unprotected groups are the same
Equal opportunity	True positive rate is the same between protected and unprotected groups
Predictive rate parity	Fraction of correct positive predictions is the same for protected and unprotected groups

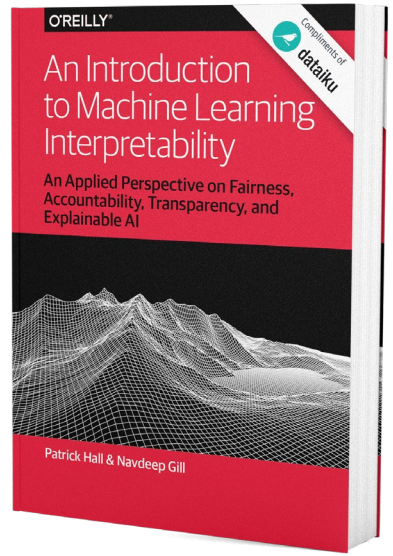
- **Principes:**

  - Solution « sciences dures », pour data scientists
  - Impose à l'algorithme une contrainte pour ne pas discriminer des groupes de populations
  - Pre processing (modification du data set), in processing (contraintes dans le processus d'apprentissage), post processing (changer les seuils de décision)

- **Critiques du Fair ML, solutions techniques pour éviter les discriminations**
  - Introduire des principes dans l'algorithme limite la prise en compte du contexte (Lipton, 2020)
  - Utiliser des méthodes pour corriger la "fairness" peut accentuer les inégalités intra categories (ex: inégalités entre femmes) (Speicher 2018)
  - Pas de prise en compte de l'environnement socio-technique (Selbst 2019).
  - Représentation trop simple des catégories (race, genre par ex)
  - Effets contraires (Fazelpour & Lipton, 2020)
  - Inefficace (Selbst et al., 2019)
  - Trade off entre performance et fairness

Phase 2: les réponses techniques, de l'explicabilité à l'interprétabilité des algorithmes

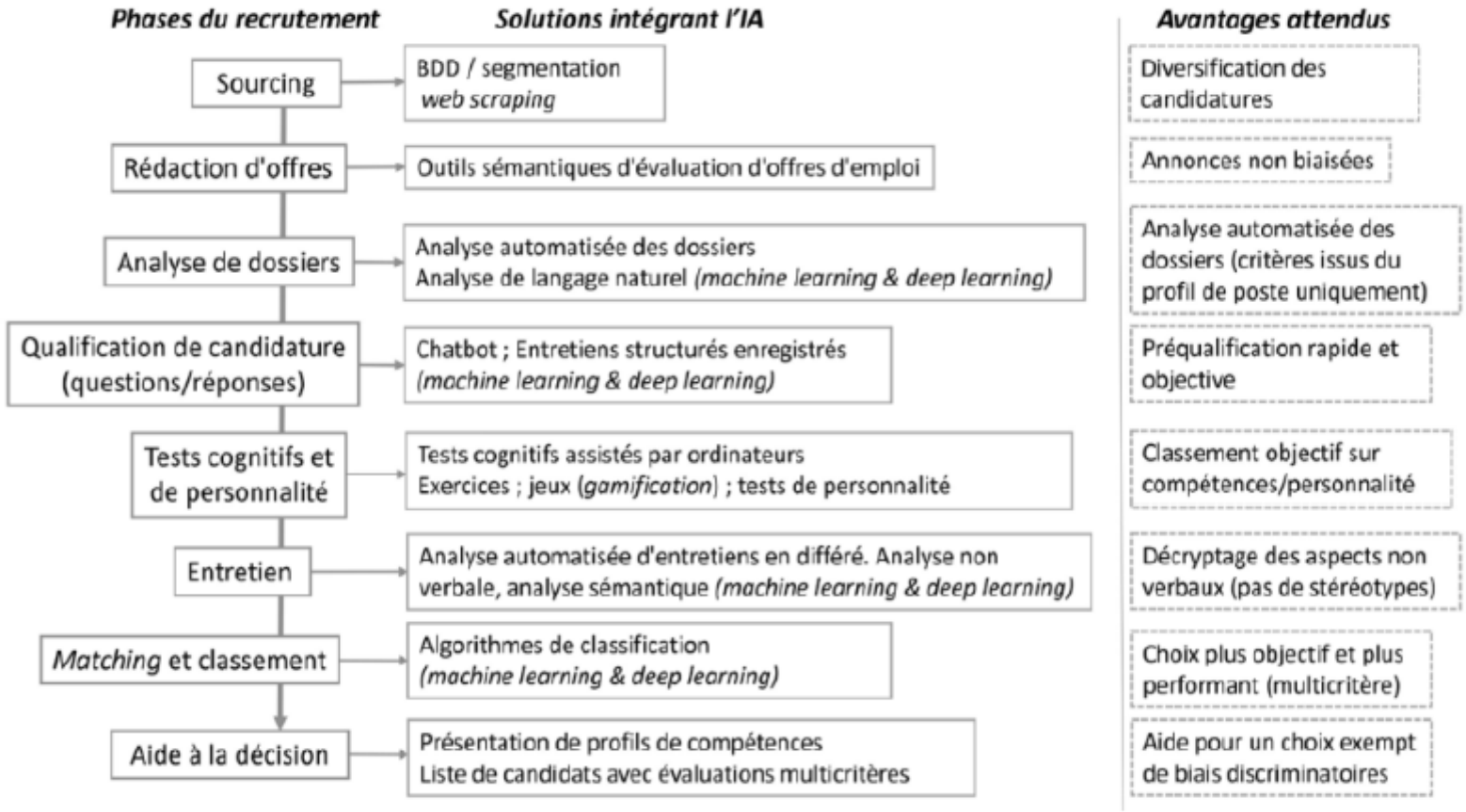
- ***Comprehensibility:*** the user must understand why the algorithm give him these results
- ***Actionability:*** the user must be able through his actions to modify the algorithm results
- ***Generalizability:*** the user must be able to generalize the results with his own case results
- ***Complexity (ou simplicity):*** too much information limits the user's understanding



Jean Marie John Matthews (2021) Critical Empirical Study on Black-box Explanations in AI. International Conference on Information Systems

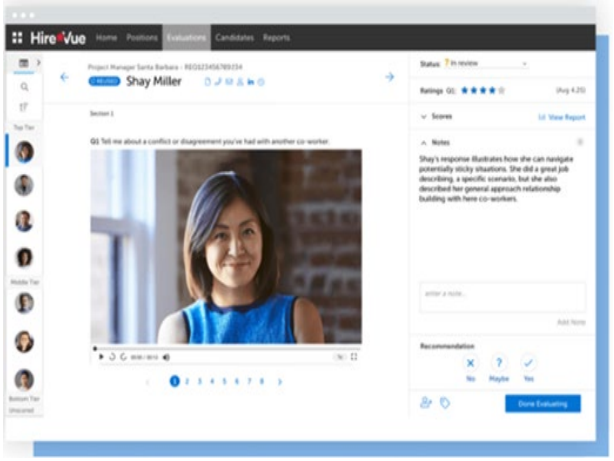
Phase 3: approches pluridisciplinaires computer science/SHS
Repenser l'IA dans sa conception (*Matthews Jean-Marie, Cardon Dominique, Balagué Christine, 2022)*

Le cas des algorithmes de recrutement (1)



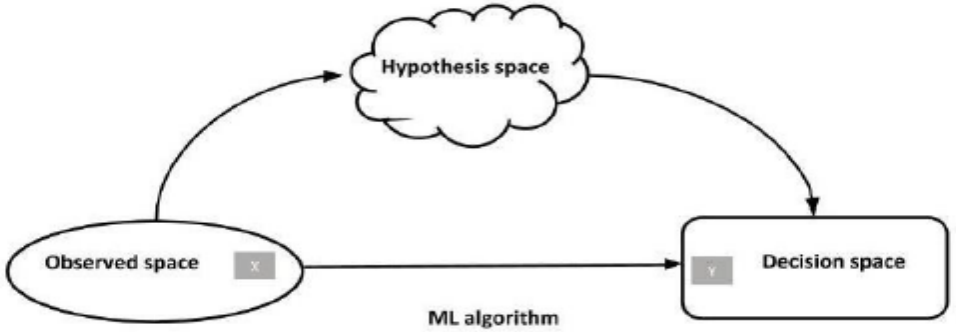
La place de l'IA dans le processus de recrutement (Lacroux, Martin-Lacroux, 2021).

Le cas des algorithmes de recrutement (2)



Variables latentes:

Compétence du candidat
Intelligence du candidat
Employabilité du candidat
Meilleur candidat
….



Techniques:

- Analyse du texte (NLP)
- Reconnaissance d'image
- Identification des blocs dans le CV
- Matching offre/CV par bloc
- Algorithmes de distance entre concepts (ex finance proche de BNP)
- Algorithmes de scoring multi critères

Score (ex: d'employabilité)

Classement (liste des candidats retenus ou liste des rejetés)

Recommandation de parcours de carrière

…..

Matthews Jean-Marie, Cardon Dominique, Balagué Christine (2022). From Reality to World. A Critical Perspective on AI Fairness. Journal of Business Ethics.

Enjeux éthiques des algorithmes de recrutement

- Tensions entre optimisation des processus de recrutement et complexification de ses modalités
  - Impact sur le ressenti des candidats
  - Opacité du système et difficultés d'explicabilité des méthodes d'IA
- Tensions entre efficacité prédictive de l'algorithme et discriminations
- Tensions entre précision des données collectées et limites légales
- Tensions entre optimisation du processus de recrutement et risque d'endogénéité
- Tensions sur la responsabilité des acteurs

Etude des pratiques: perception des risques
Méthodologie: 20 entretiens concepteurs d'algorithmes de recrutement (start ups et grandes entreprises)

Risque de discrimination :

- Risque fort dans le matching
- Discrimination potentielle à partir de certaines données:
  - année de naissance
  - sexe
  - quartier
  - turn over dans le CV
  - trous dans le CV
  - femmes enceintes
  - handicap
  - maladie
- L'algo accentue le biais de discrimination du recruteur
- Risque de rejet des profils atypiques
- Privilégie pour des postes écrits en français des personnes à profil en français (les modèles préfèrent faire une comparaison au sein d'une langue plutôt qu'entre différentes langues)

Protection des données personnelles:

- Risque de non respect de l'identité car on manipule des CV
- Risque de non respect du RGPD
- Risques liés au stockage des données

Biais des données

- Biais GIGO (mauvaise qualité de photo dans le CV, données pas fiables)
- Biais qualité des annotateurs extérieurs
- Biais de variable omise
- Biais de sélection: insuffisance en quantité et représentativité statistique des cas
- Biais d'endogénéité: on s'appuie sur le passé : et si de nouveaux métiers apparaissent ?

Clonage : l'algorithme reproduit des biais passés.

- Risque de reproduction des biais de parcours professionnels précédents

RISQUES PRESENTS DANS LA LITTÉRATURE MAIS NON MENTIONNES:
sécurité des données/ risques liés à l'anonymisation/ opacité des systèmes/ opinions encapsulées dans les algorithmes

Etude des pratiques: solutions proposées
Méthodologie: 20 entretiens concepteurs d'algorithmes de recrutement (start ups et grandes entreprises)

Discrimination :

- Travail spécifique et forte attention sur ce sujet
- on exclut des critères de la base de décision:
  - age
  - sexe
  - Nationalité
  - quartier
- certains interlocuteurs vont plus loin (turn over dans le CV, trous dans le CV, femmes enceintes, maladie, handicap)
- On bride l'IA: on interdit à l'algorithme d'apprendre sur des variables discriminantes (nettoyage de la base d'apprentissage)
- Au lieu de passer par des critères booléens (oui/non), on donne un score des candidatures (accès à une plus grand diversité)
- Possibilité pour le recruteur de traiter les candidatures de façon anonyme

Protection des données personnelles

- Respect du RGPD, de la loi
- Forte sensibilisation des recruteurs sur ce sujet, on en parle avec eux au début du projet
- Pas de stockage du CV s'il n'est pas anonymisé
- Respect des cookies et de l'identité
- Suppression de données (photo, adresse, téléphone, genre)
- Consentement des candidats quant à l'utilisation et le stockage de leurs données anonymisées/ Notification au candidat sur le stockage
- Consultation de la CNIL pour connaître les obligations et les droits
- Stockage sur un cloud français ou européen

Biais des données:

- Biais de sélection: on essaie d'avoir une BDD large en quantité et en types de données (décisions positives négatives, motifs de refus, raisons, base de trajectoires, base de décision, feedback candidats)
- Données normées et comparables

Clonage : l'algorithme reproduit des biais passés

- Utilisation d'algorithmes qui ne s'appuient pas sur des données historiques (ex on exclut la connaissance préalable des différentes mobilités)

SOLUTIONS NON MENTIONNEES: FAIR ML, pas de réflexion sur les catégories, pas d'explicabilité

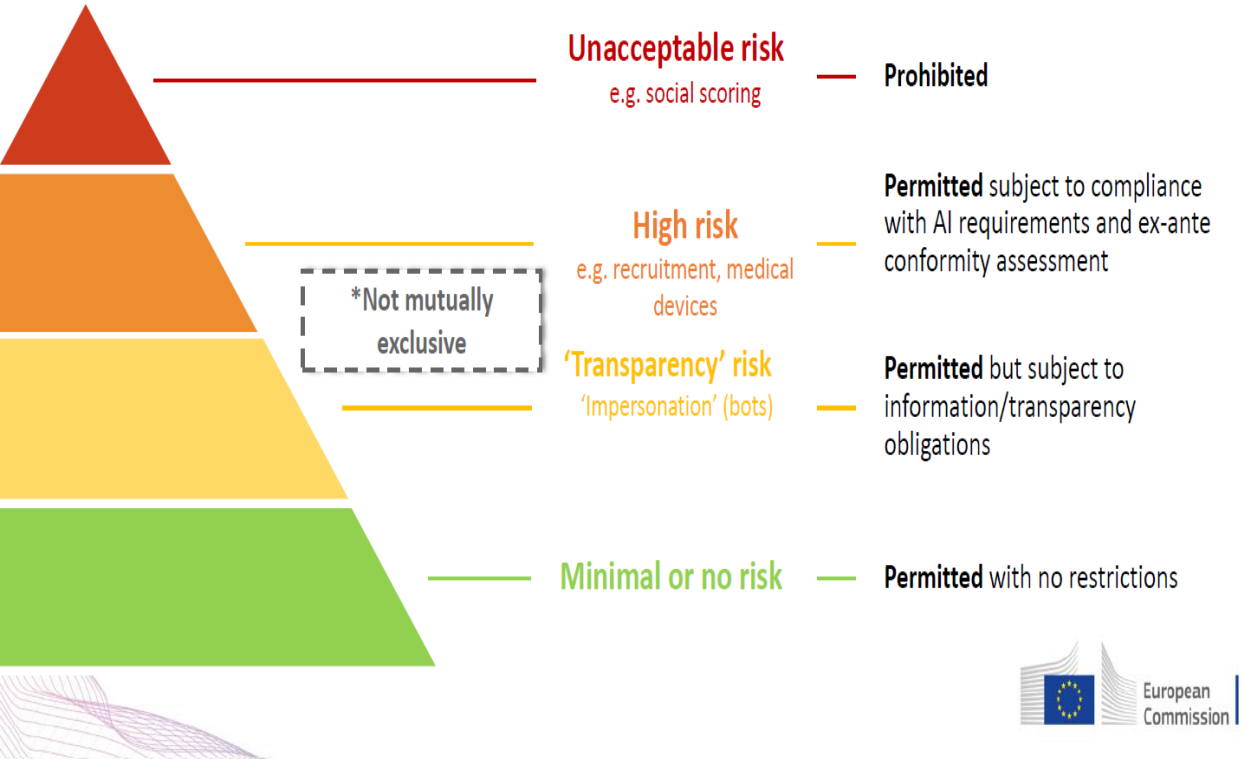
Etude des pratiques: un renvoi des responsabilités
Méthodologie: 20 entretiens concepteurs d'algorithmes de recrutement (start ups et grandes entreprises)

- **C'est le concepteur de l'algorithme qui doit être responsable:**
  - Le concepteur doit brider l'algorithme car le recruteur a des biais inconscients
  - Le recruteur peut demander des choses au concepteur et celui-ci doit vérifier certaines demandes auprès de son avocat (ex blacklister des individus)

- **C'est le recruteur qui doit être responsable:**
  - C'est au recruteur de savoir ce qu'il souhaite anonymiser
  - Le recruteur doit être bienveillant sur les CV et ne pas discriminer de son côté
  - La mesure de la sympathie du candidat relève du recruteur, pas du concepteur
  - Ce sont les recruteurs qui peuvent ne pas accepter des candidats au parcours atypique, qui définissent une liste des motifs de refus acceptables, qui décident d'un feedback aux candidats

- **C'est l'algorithme qui est éthique et responsable:**
  - Un algorithme n'est pas sensible à l'esthétique
  - Pas de risque de renforcement des biais des recruteurs (ex on supprime l'adresse)
  - Un algorithme traite tous les emplois de la même manière (pas de différence entre agent d'entretien et super expert au Brésil)
  - Un algorithme permet la discrimination positive et la diversité

Les solutions aux enjeux éthiques de l'IA
La régulation: le projet de loi européenne sur l'IA



1 SAFETY COMPONENTS OF REGULATED PRODUCTS

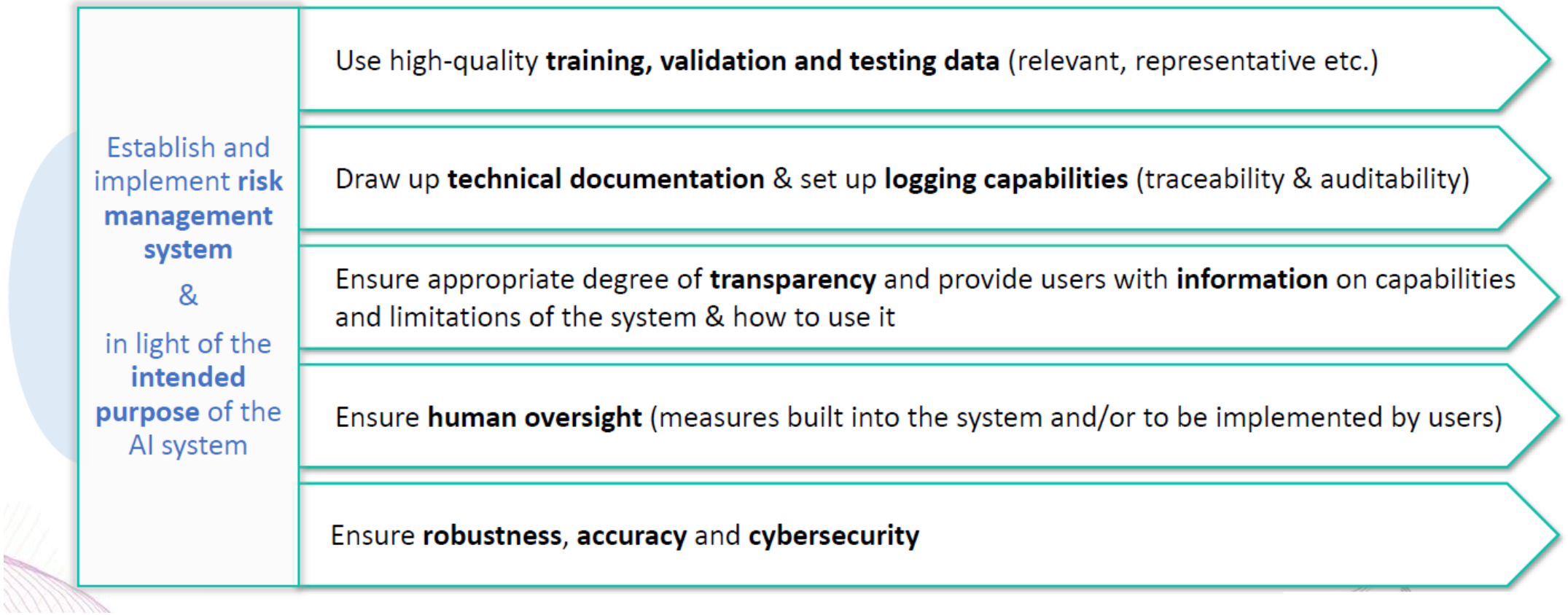
(e.g. medical devices, machinery) which are subject to third-party assessment under the relevant sectorial legislation

2 CERTAIN (STAND-ALONE) AI SYSTEMS IN THE FOLLOWING AREAS

- ✓ Biometric identification and categorisation of natural persons
- ✓ Management and operation of critical infrastructure
- ✓ Education and vocational training
- ✓ Employment and workers management, access to self-employment
- ✓ Access to and enjoyment of essential private services and public services and benefits
- ✓ Law enforcement
- ✓ Migration, asylum and border control management
- ✓ Administration of justice and democratic processes

European Commission

Les solutions aux enjeux éthiques de l'IA
La régulation: les exigences du projet de loi européenne pour l'IA « high-level risk »



Merci de votre attention

Q&A

christine.balague@imt-bs.eu
@balague

Institut Mines-Télécom

FONDATION DU RISQUE Louis Bachelier

SciencesPo