Good in Tech
Enjeux éthiques de l'IA et responsabilité numérique des organisations

Christine BALAGUE

Professeur, IMT-BS

Titulaire de la Chaire Good in Tech www.goodintech.org

Membre du groupe d'experts du CSA sur la désinformation en ligne

Membre du Comité d'Ethique de la Défense

Membre de la commission impact des recommandations de la Haute Autorité de Santé

Ex VP du Conseil National du Numérique

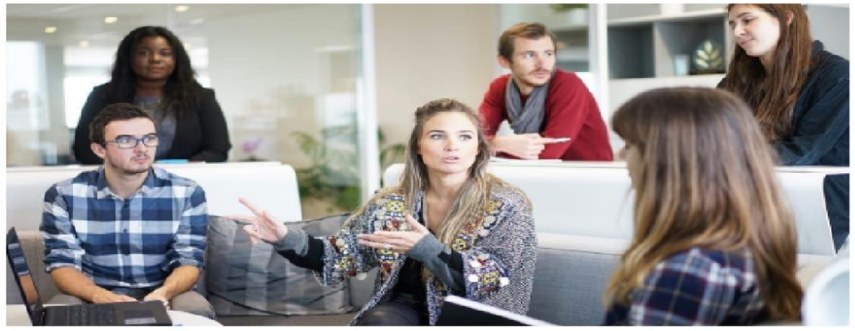
Good in Tech: quatre axes de recherche

Axe 1 : innovation numérique responsable : quelles mesures ?

Quelles sont les dimensions et les mesures de l'innovation numérique responsable ?
Comment intégrer l'innovation numérique responsable dans la responsabilité sociale numérique des entreprises ?

Axe 2 : comment développer des technologies responsables « by design » ?

Comment éviter les biais liés aux données ?
Comment développer des technologie responsables by design ? (plus explicables, redevables, équitables et respectant la régulation sur les données personnelles)



Axe 3 : réinventer les futurs : Quelle société pour demain dans un monde numérique

Quelle société pour demain dans un monde numérique ?
Comment réinventer les futurs dans une perspective fidèle aux Lumières, de préservation du principe d'égalité appliqué au monde connecté ?

Axe 4: gouvernance de l'innovation et des technologies responsables

Quels sont les mécanismes de gouvernance de l'innovation numérique responsable ?
Quels sont les niveaux pertinents de cette gouvernance : Europe, Nation, Entreprise ?

Good in Tech: quelques chiffres sur les réalisations 2020-2021 dans un contexte de pandémie

10 projets de recherche par chercheurs permanents des 2 institutions, financés en 2020

18 articles de recherche soit publiés dans des journaux classés, soit acceptés dans des conférences internationales de recherche, soit soumis à des journaux classés, soit présentés dans des workshops

2 thèses dont une soutenue en décembre 2021 / **3 post-doctorants** (équivalent TP)

12 conférences & webinars (présence internationaux)

10 rapports dont un soumis au prix Marie-Dominique Hagelsteen (ARPP)

3 outils développés:

Trade Off IA logiciel développé pour les data scientists pour mitiger les risques de l'IA. Logiciel testé avec une soixantaine d'utilisateurs data scientists

2 cartographies (acteurs de l'IA en France/ acteurs publicité numérique responsable)

Prix Good in Tech : 2 éditions 2020 (lauréat Datafarm) et 2021 (lauréat JobRepublik), dans le cadre du prix IMT-Bercy (visibilité auprès de plus de 200 start-ups/ DGE/ entreprises membres du jury)

Partenariat international avec l'Obvia au Canada

940 étudiants avec cours sur Innovation Numérique Responsable et Ethique de l'IA

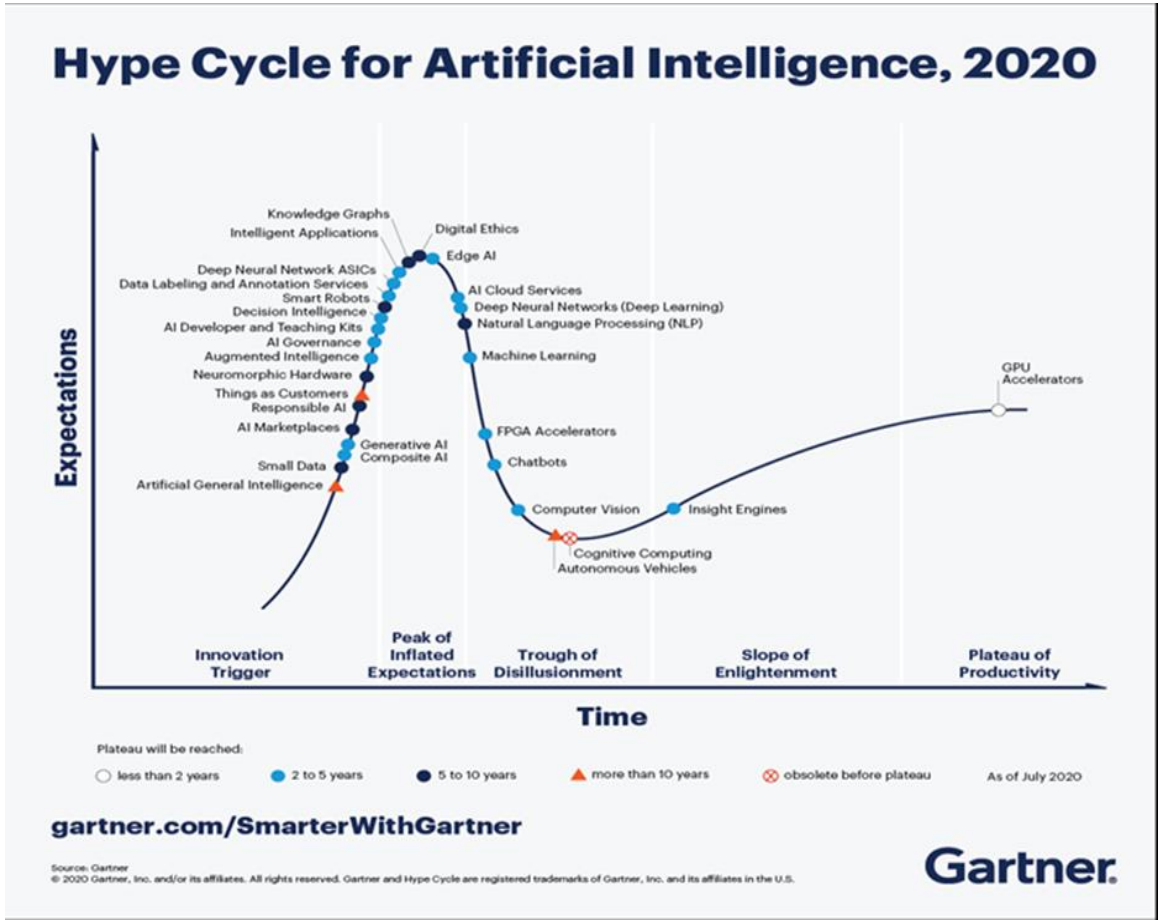
Invited Speaker dans plus de 20 conférences professionnelles de plus de 100 personnes

Etude et sensibilisation auprès des **1000 membres entreprises de Cap Digital**

Matinales partenaires 2021-2022

Pourquoi parler d'éthique de l'IA?
L'IA est partout…

- Lieux: *villes intelligentes, smart home, voiture, digital (moteur de recherche, réseaux sociaux)…*
- Secteurs économiques: *banque, assurance, agriculture, e-commerce, automobile, industrie du futur, éducation, santé, industrie culturelles et créatives, militaire…*
- Métiers: *finance, ressources humaines, marketing, publicité, production…*



Pourquoi parler de l'éthique de l'IA?
L'IA est déployée massivement par des plateformes digitales…

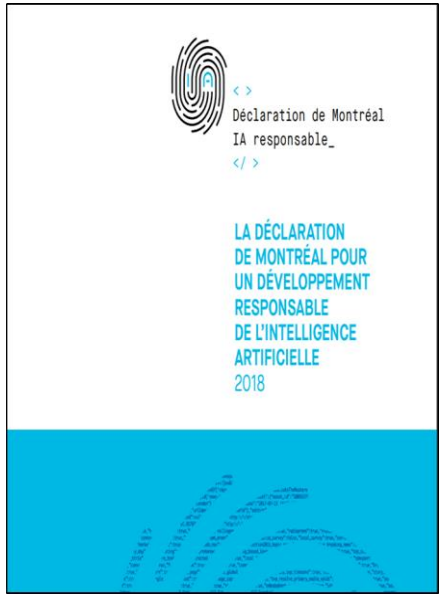
- Entreprises mondialement dominantes
- Collecte massive de données *(modèles bi-faces)*
- Capacité de recrutement des meilleurs data scientists et chercheurs mondiaux
- Financement considérable en R&D *(149 milliards $ en 2021 pour les MAAMA, plus que le Pentagone)*
- Compétition US-Chine

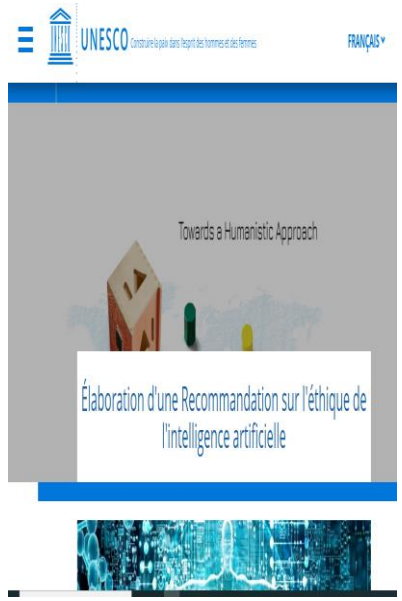
Exhibit 3 - The 50 Most Innovative Companies of 2021

Rank: 1–10	Rank: 11–20	Rank: 21–30	Rank: 31–40	Rank: 41–50
1 Apple	11 Siemens	21 Toyota	31 Xiaomi	41 Inditex
2 Alphabet	12 LG	22 Salesforce	32 IKEA	42 Moderna
3 Amazon	13 Facebook	23 Walmart	33 Fast Retailing	43 Philips
4 Microsoft	14 Alibaba	24 Nike	34 Adidas	44 Disney
5 Tesla	15 Oracle	25 Lenovo	35 Merck & Co.	45 Mitsubishi
6 Samsung	16 Dell	26 Tencent	36 Novartis	46 Comcast
7 IBM	17 Cisco	27 Procter & Gamble	37 Ebay	47 GE
8 Huawei	18 Target	28 Coca-Cola	38 PepsiCo	48 Roche
9 Sony	19 HP	29 Abbott Labs	39 Hyundai	49 AstraZeneca
10 Pfizer	20 Johnson & Johnson	30 Bosch	40 SAP	50 Bayer

Source: BCG Global Innovation Survey 2020 and 2021.

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





Les problèmes d'éthique posés de l'IA
L'IA peut discriminer: les biais des données

- Biais « Garbage in, garbage out » (GIGO)
- Biais de variable omise (données incomplètes)
- Biais de sélection
- Biais d'endogénéité
- Les données représentent-elles correctement le monde?

Sociologie:

Digital Labor et micro travail

- Sécurité et stockage des données
- Anonymisation et ré-identification

Des données biaisées génèrent des algorithmes biaisés, potentiellement discriminants

Les problèmes d'éthique posés de l'IA
L'IA peut discriminer: le cas du système de santé aux US

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.

Sciences, 366, 447-453 (2019)

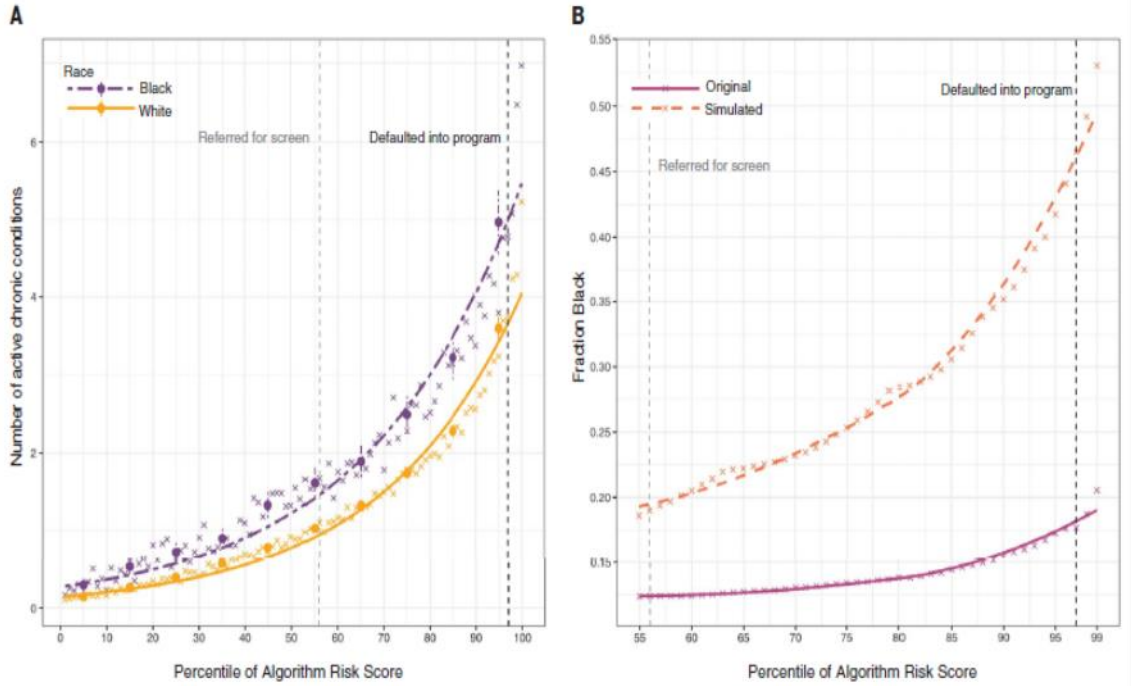
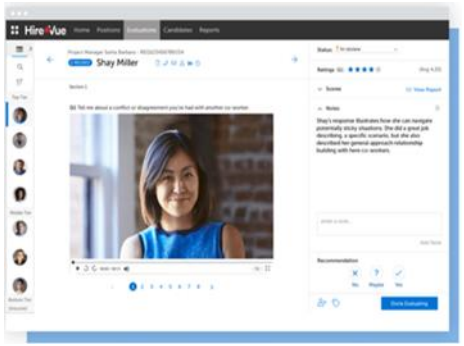


Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (**A**) Mean number of chronic conditions by race, plotted against algorithm risk score. (**B**) Fraction of Black patients at or above a given risk score for the original algorithm ("original") and for a simulated scenario that removes algorithmic bias ("simulated": at each threshold of risk, defined at a given percentile on the *x* axis, healthier Whites above the threshold are replaced with less healthy Blacks below the threshold, until the marginal patient is equally healthy). The × symbols show risk percentiles by race; circles show risk deciles with 95% confidence intervals clustered by patient. The dashed vertical lines show the auto-identification threshold (the black line, which denotes the 97th percentile) and the screening threshold (the gray line, which denotes the 55th percentile).

Les problèmes d'éthique posés de l'IA
L'IA peut discriminer: le cas des algorithmes de recrutement



It's pseudoscience. ***it's a license to discriminate****" Meredith Whittaker, co-founder of the AI now Institute, research center in New York.*

Variables latentes:
Compétence du candidat
Intelligence du candidat
Employabilité du candidat
Meilleur candidat
….

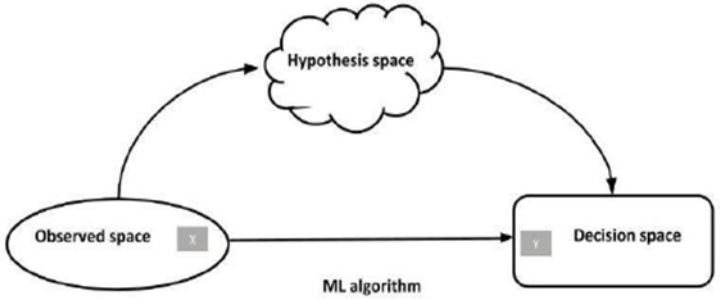


Figure 1: The three spaces of the ML process

Score (ex: d'employabilité)

Classement (liste des candidats retenus ou liste des rejetés)

Recommandation de parcours de carrière

…..

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

Les problèmes d'éthique posés de l'IA
L'IA est opaque et ses résultats difficiles à expliquer

- 1ère période : Systèmes Experts

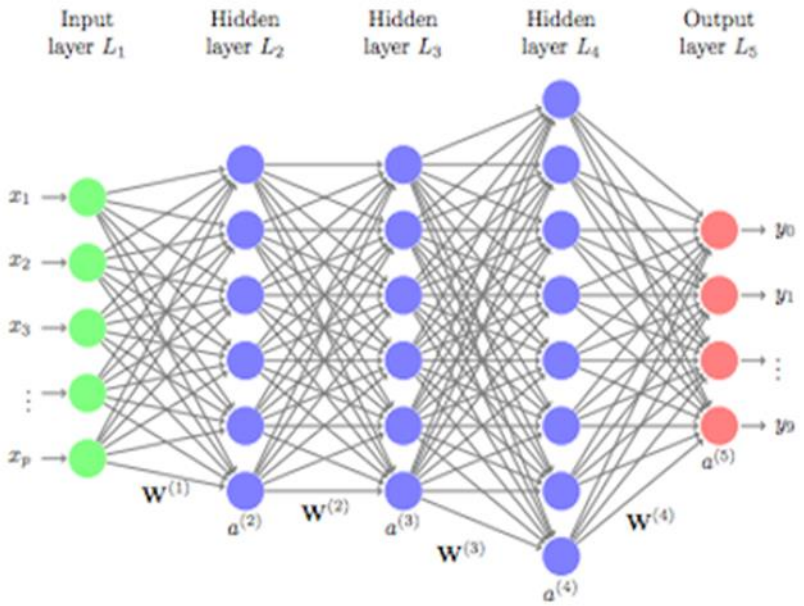
- 2ème période : Apprentissage statistique

Apprendre sans comprendre

- 3ème période: Adaptation contextuelle

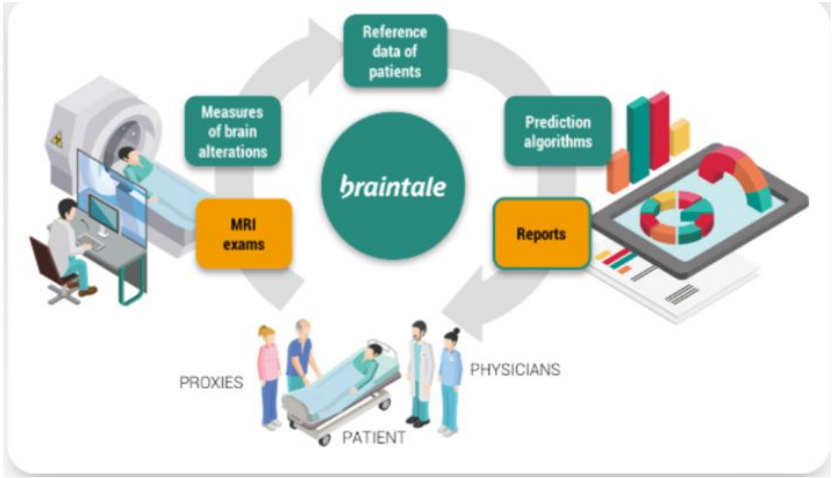
Produire des explications liées au contexte

– *Mais systèmes trop complexes*
– *Opacité*
– *Problème d'explicabilité……*



Les problèmes d'éthique posés de l'IA
L'IA est opaque: risque élevé sur des domaines sensibles

L'agence des produits de santé (FDA) aux États-Unis a approuvé le tout premier algorithme capable d'anticiper les arrêts cardio-respiratoires en milieu hospitalier. Un outil de surveillance qui a déjà sauvé des vies selon les essais cliniques réalisés.



BrainScore is a SaaS solution combining the brain quantification (brainQuant) and artificial intelligence algorithms for predicting coma emergence of patients in coma admitted in intensive care units after severe traumatic brain injuries, cardiac arrest or haemorrhagic stroke.

brain SCORE

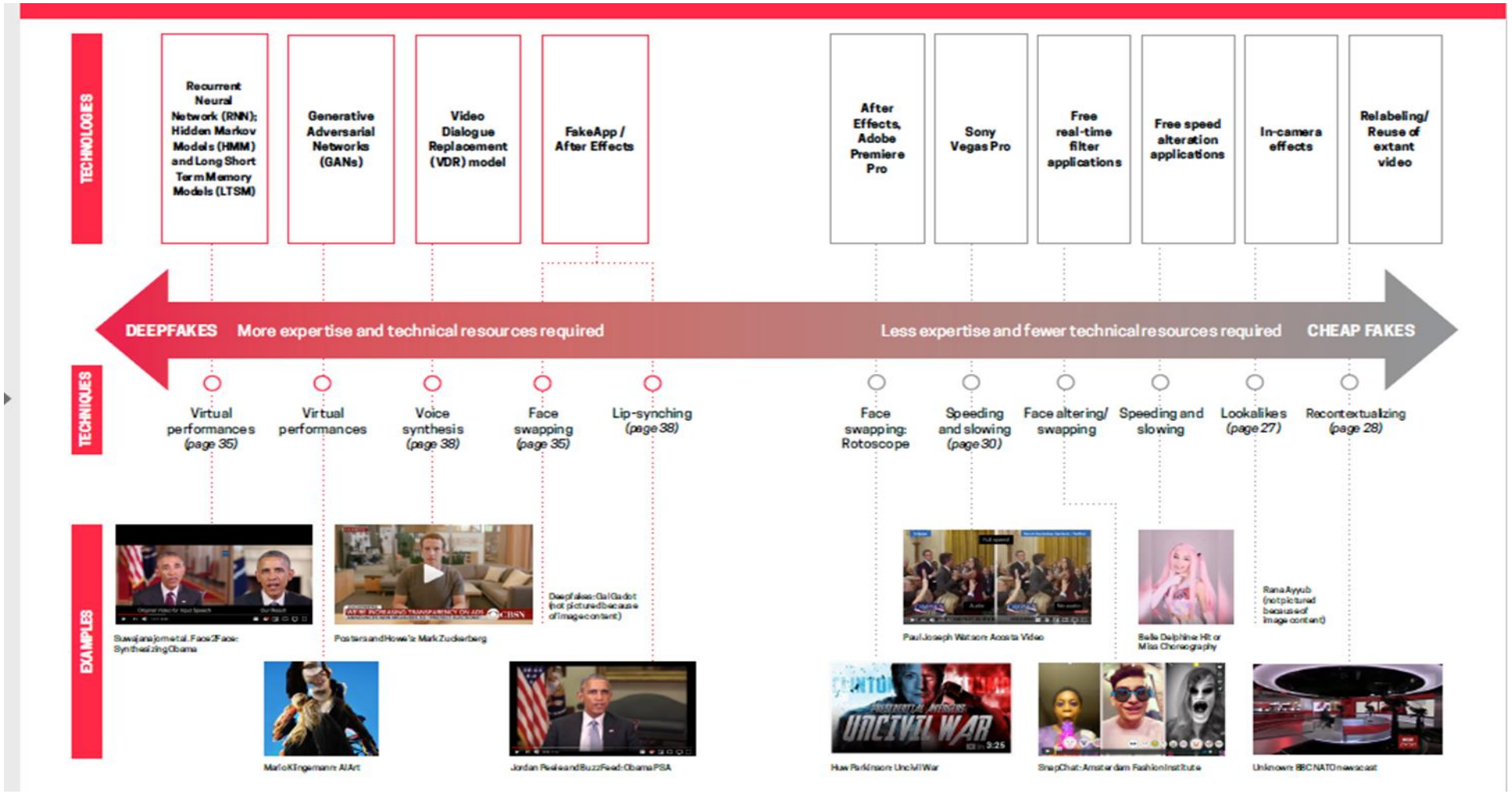
23andMe : le géant des tests ADN vend les données génétiques de ses clients

Bastien L 9 août 2018 Sécurité Ecrire un commentaire

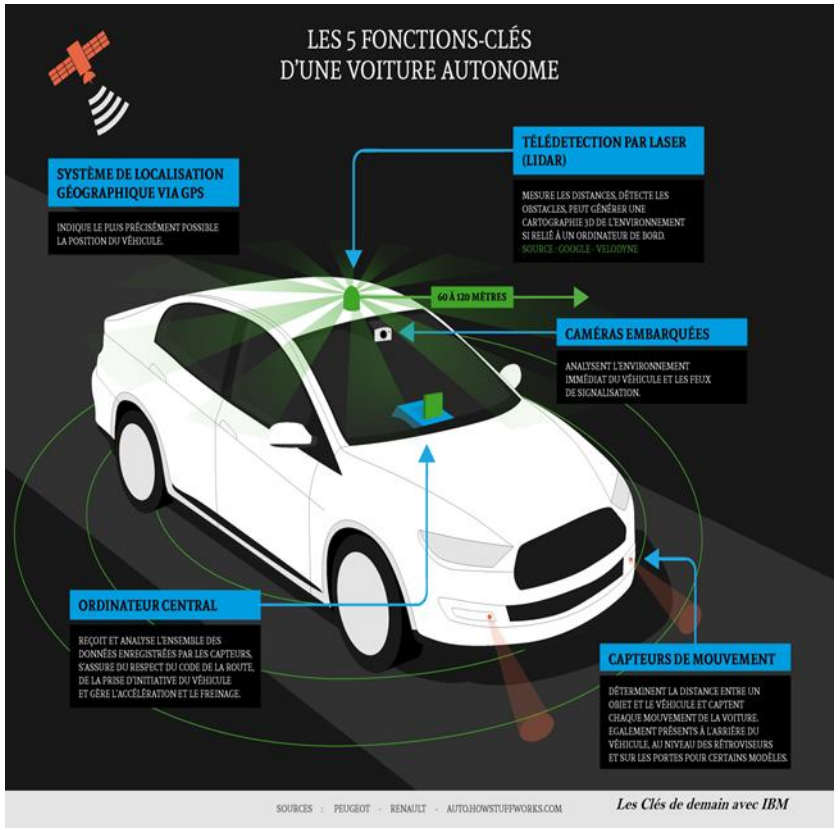
23andMe, l'un des leaders des tests ADN destinés aux particuliers, vient de vendre les données génétiques de ses 5 millions de clients au géant britannique de l'industrie pharmaceutique GlaxoSmithKline pour 300 millions de dollars. Un partenariat qui suscite la controverse.

Fondée en 2006, l'entreprise californienne 23andMe est spécialisée dans les tests d'ADN à destination des particuliers. Elle permet à ses clients de **remonter leur arbre généalogique** de façon fiable pour découvrir qui sont leurs ancêtres lointains.

Les problèmes d'éthique posés de l'IA
L'IA est opaque: risque de manipulation, cas des deep fakes & cheap fakes



Les problèmes d'éthique posés de l'IA
L'IA est opaque: problèmes d'identification de responsabilité



Gola Romain (2017). L'adaptabilité de la règle de droit face à l'émergence des véhicules connectés et autonomes. Revue Lamy Droit de l'Immatériel, no.133, p.57-61.

Les problèmes d'éthique posés de l'IA
L'IA génère des bulles d'enfermement

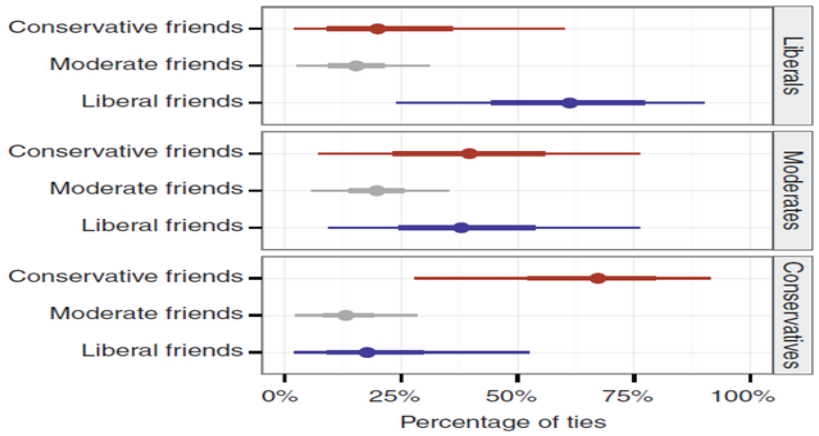
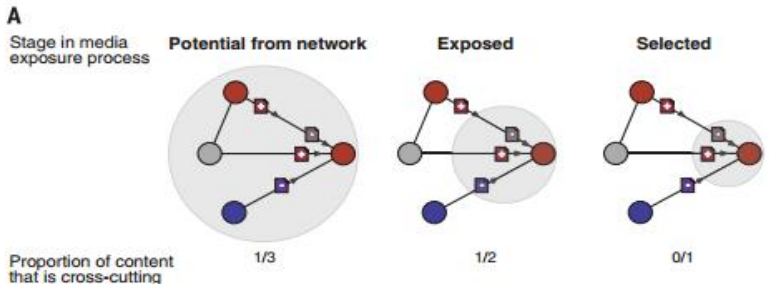
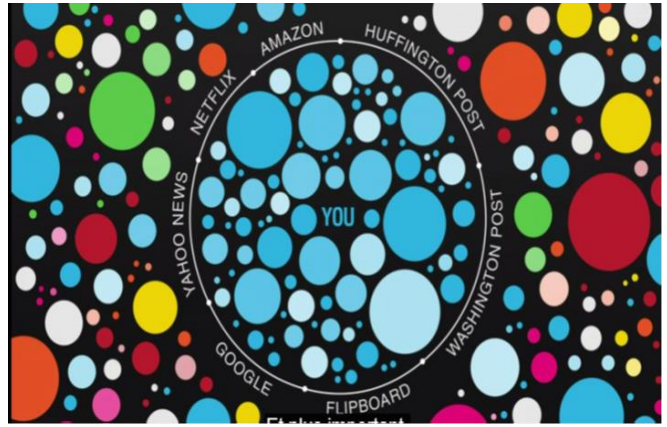
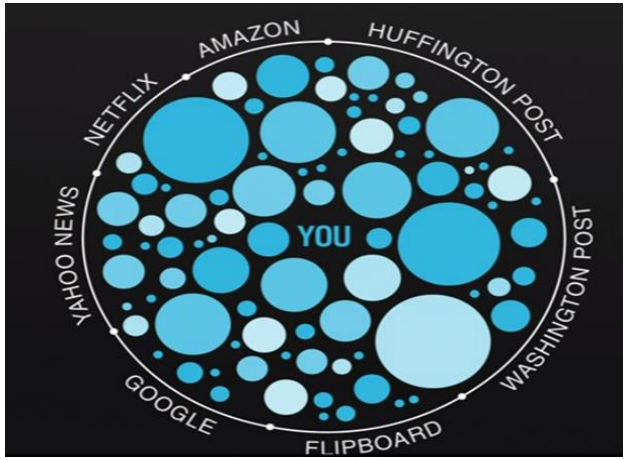


Fig. 2. Homophily in self-reported ideological affiliation. Proportion of links to friends of different ideological affiliations for liberal, moderate, and conservative users. Points indicate medians, thick lines indicate interquartile ranges, and thin lines represent 10th to 90th percentile ranges.

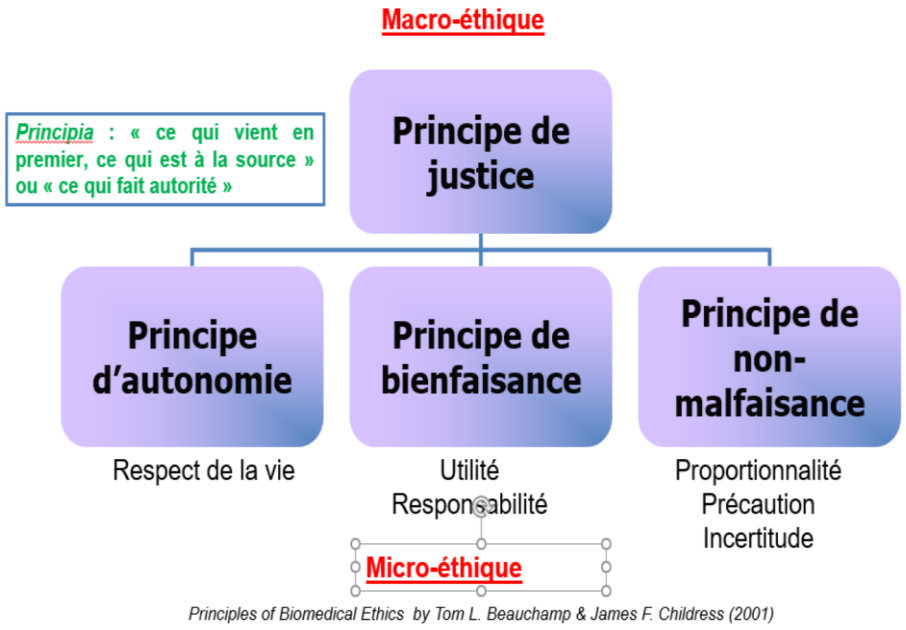






Bakshy E., Messing S., Adamic L.A. (2015), Exposure to ideologically diverse news and opinion on Facebook, Science, vol. 348, issue 239

Les solutions aux enjeux éthiques de l'IA
Phase 1: l'IA doit respecter des grands principes



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

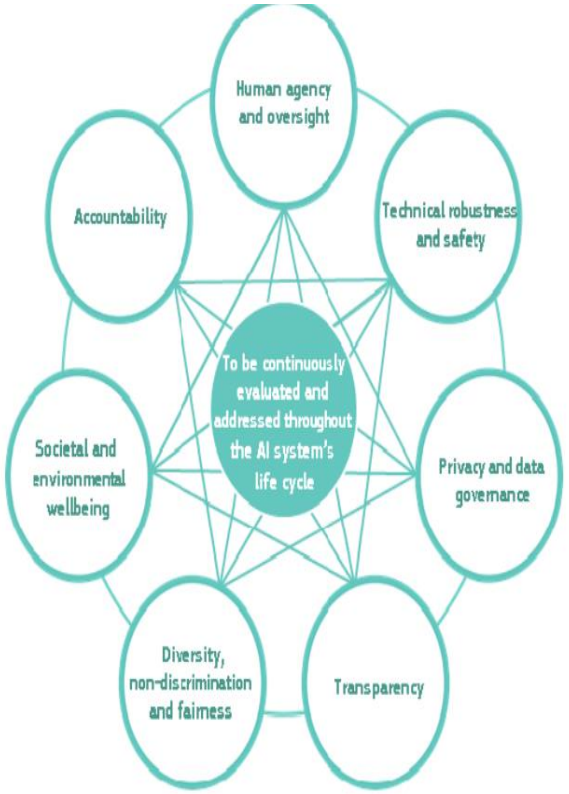


Figure 2: Interrelationship of the seven requirements: all are of equal importance, support each other, and should be implemented and evaluated throughout the AI system's lifecycle

Les solutions aux enjeux éthiques de l'IA
Phase 2: les 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

- **Principe:**

✓ 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)

Les solutions aux enjeux éthiques de l'IA
Phase 2: les réponses techniques, le cas de l'explicabilité des algorithmes

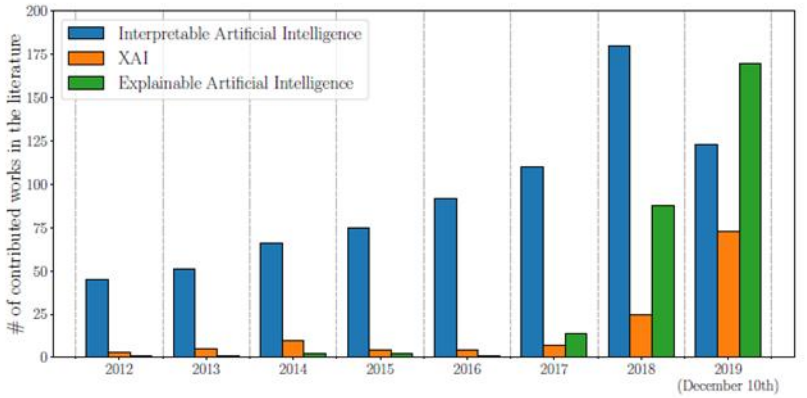


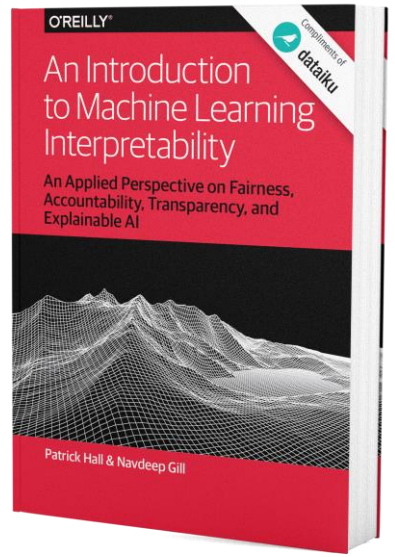
Figure 1: Evolution of the number of total publications whose title, abstract and/or keywords refer to the field of XAI during the last years. Data retrieved from Scopus® (December 10th, 2019) by using the search terms indicated in the legend when querying this database. It is interesting to note the latent need for interpretable AI models over time (which conforms to intuition, as interpretability is a requirement in many scenarios), yet it has not been until 2017 when the interest in techniques to explain AI models has permeated throughout the research community.

Source: Barredo Arieta et al. (2019)

- ***Understandability (intelligibility):*** characteristics of a model to make a human understand its function-how the model works-without any need for explaining its internal structure or the algorithmic means by which the model processes data internally
- ***Comprehensibility:*** ability of a learning algorithm to represent its learned knowledge in a human understandable fashion
- ***Interpretability:*** ability to explain or to provide the meaning in understandable terms to a human
- ***Explainability:*** associated with a notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans
- ***Transparency:*** a model is transparent if by itself it is understandable (simulatable models, decomposable models, algorithmically transparent models).

Les solutions aux enjeux éthiques de l'IA
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



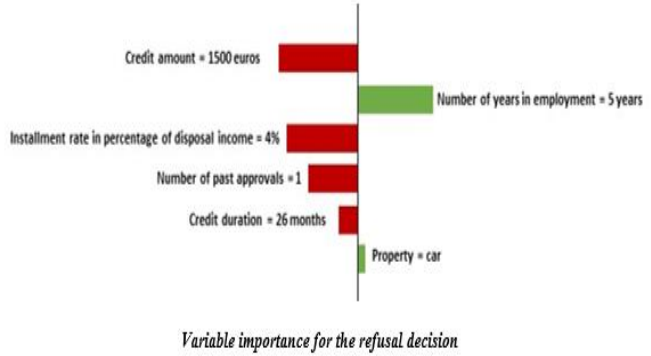
Kosol et al. (2020)

Les solutions aux enjeux éthiques de l'IA
Phase 2: les réponses techniques, de l'explicabilité à l'interprétabilité des algorithmes

Counterfactual explanation

> — If the credit amount was between 3000 and 4000 euros, your credit application would have been accepted by the algorithm. As a reminder, your credit amount is currently 1500 euros.
>
> — If the credit duration was reduced to less than 12 months, your credit request would have been accepted by the algorithm. As a reminder, your credit duration is currently 26 months.

Post hoc Shapley explanation



			RECEPTION			
			Fairness perception	Trust perception	Negative Comments rate	Claim rate
TECHNICAL INTERPRETABILITY	NO EXPLANATION (BASELINE)		3.48	3.38	55%	58%
	TRANSPARENT		3.45	**3.7**	51%	55%
		pvalue	*(0.9)*	***(0.1)***	*(0.6)*	*(0.9)*
	POST-HOC SHAPLEY		3.3	3.3	55%	66%
		pvalue	*(0.3)*	*(0.8)*	*(1)*	*(0.6)*
	POST-HOC COUNTERFACTUAL		3.45	**3.7**	**38%**	**37%**
		pvalue	*(0.9)*	***(0.1)***	***(0.016)***	***(0.1)***

FIGURE 6.3 – Selection of best explanation modes when AI incident is present (t-test with H0 : means equality with the baseline scenario where there is no explanation). Population : 400 individuals corresponding to situations with sexist incidents.

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

Les solutions aux enjeux éthiques de l'IA
Phase 2: les réponses techniques, des interfaces pour les data scientists

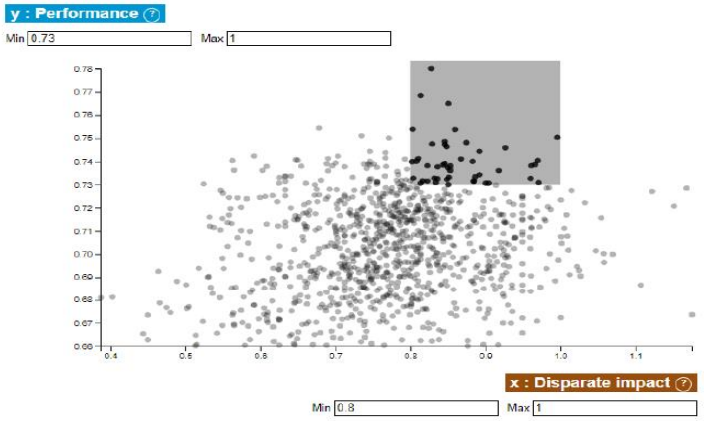


FIGURE 4.2 – Screenshot of a scatter plot of classifiers with respect to Performance and Disparate Impact metrics.
Note : The user selects here the subset of classifiers corresponding to the grey area

	Performance	Disparate Impact	Error distribution	Privacy (number of private features)	Interpretability (number of features)
With the interface	0.74 (<0.01)	0.91 (<0.01)	0.14 (0.35)	-5.8 (<0.01)	-9.4 (<0.01)
Without the interface	0.82	0.78	0.15	-6.5	-21

FIGURE 4.5 – Average metrics of the algorithms with and without the interface
P-value with t-test on H0 : means equality with the metrics without the interface

Interface accessible sur le site www.goodintech.org:
https://medialab.github.io/exp-ai/

2 niveaux de critiques des phases 1 et 2

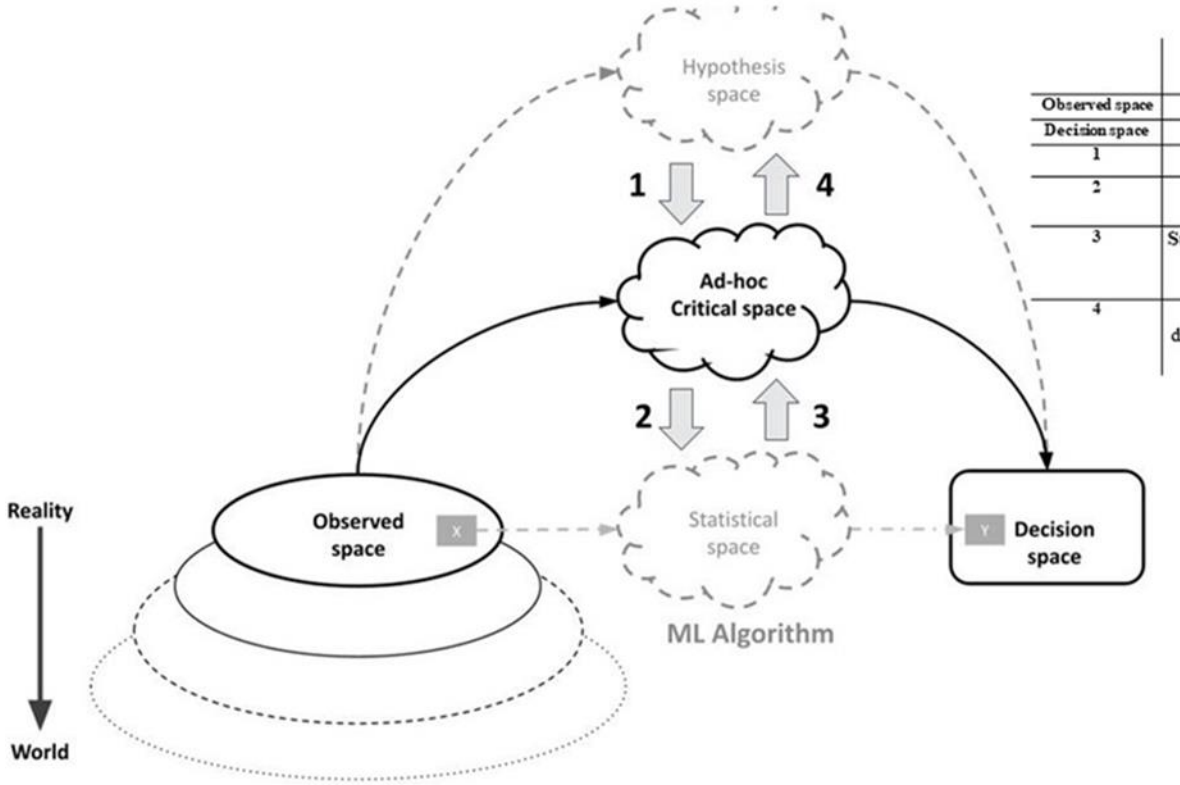
Discriminations liées à l'IA (unfairness)

- Modèles prédictifs en justice (Larson et al. 2016)
- Reconnaissance faciale (Buolamwini and Gebru 2018)
- Moteurs de recherche (Kay et al. 2015)
- Publicité (Sweeney 2013; Datta et al. 2015)
- Reconnaissance de la parole (Tatman 2016)
- Modèles d'IA pour recruter (Leicht- Deobald et al. 2019)
- Modèles prédictifs en santé (Obermeyer et al. 2019)

Critiques des initiatives sur l'IA responsable

- Ethique fondée sur des grands principes (Mittelstadt 2016, 2019; Trustworthy AI CE; Principes OCDE et Unesco)
  - 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)
- 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

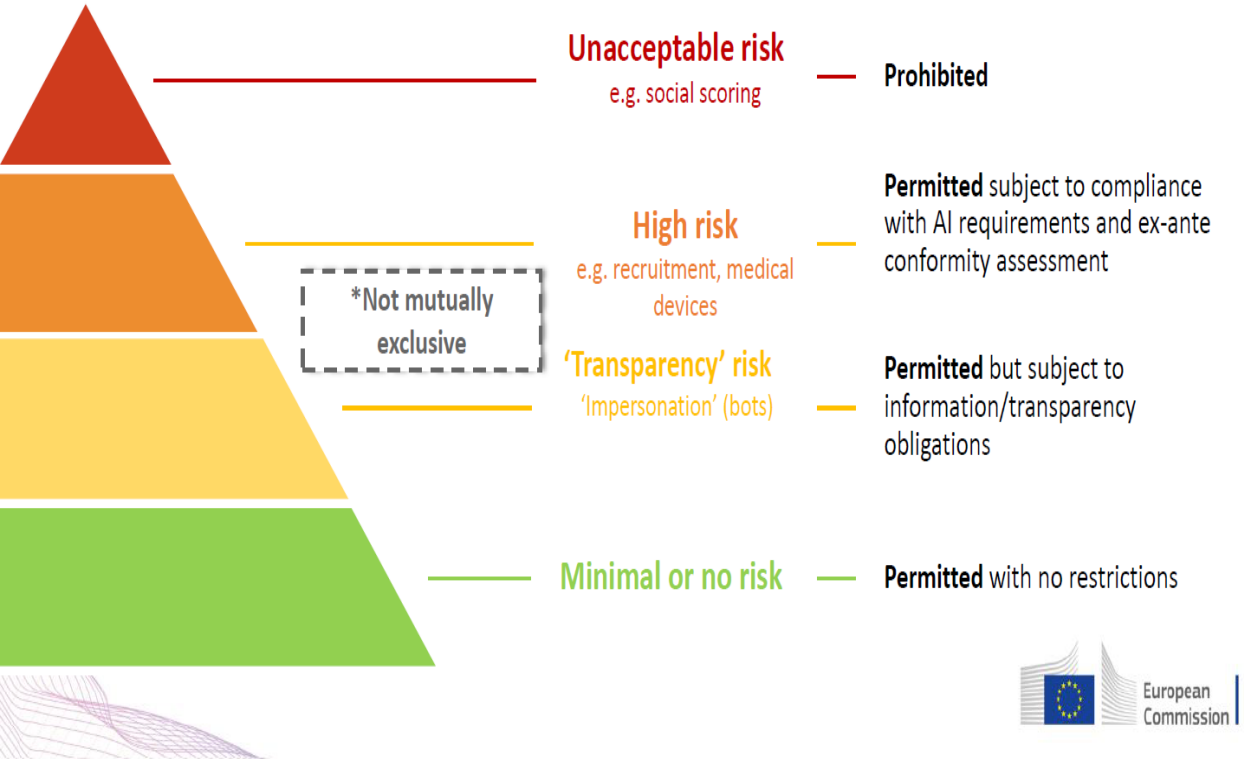
Les solutions aux enjeux éthiques de l'IA
Phase 3: apport des sciences humaines et sociales et de la RSE



	Example
Observed space	Candidate's video of the job interview and purchasing behavior data
Decision space	Having the best evaluation two years after recruitment
1	**Break down the objective.** Ex: what does it mean to be a "good candidate"?
2	**Radicalizing through experimentation.** Ex: candidates' purchasing behavior can be artificially modified to experiment algorithmic outputs.
3	**Stabilize new realities by aggregating points of view.** Ex: emergence of new categories like "candidates smiling during the interview but with impulsive buying behaviour"
4	**Re-qualify the objective and open-up realities.** Ex: Conduct a multi-stakeholder discussion to assess if the emergent categories can be normative categories for making a recruitment decision

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

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



1 SAFETY COMPONENTS OF REGULATED PRODUCTS

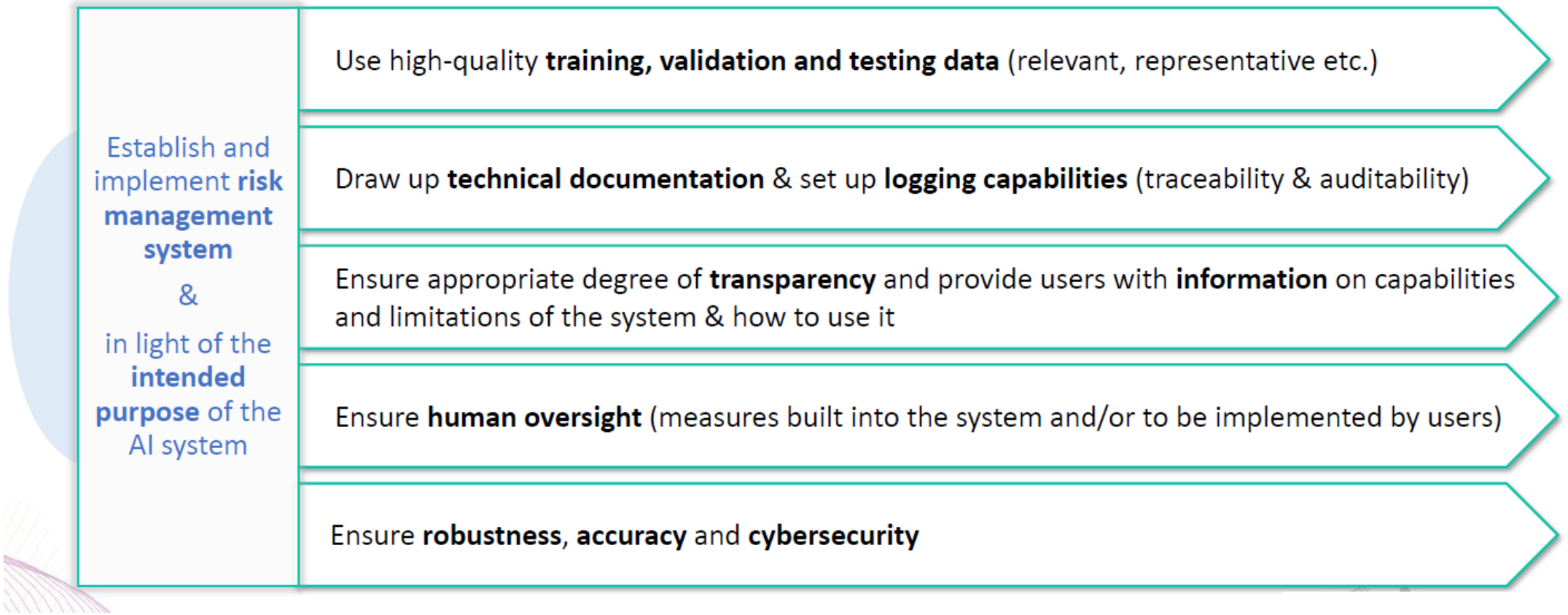
(e.g. medical devices, machinery) which are subject to third-party assessment under the relevant sectorial legislation

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 »



Les solutions aux enjeux éthiques de l'IA
Faire évoluer les pratiques des entreprises: mesurer la Responsabilité Numérique des Entreprises

Rendtorff, J.D. (2019), "The Principle of Responsibility: Rethinking CSR as SDG Management", Philosophy of Management and Sustainability: Rethinking Business Ethics and Social Responsibility in Sustainable Development, Emerald Publishing Limited, Bingley, pp. 205-220.

Buhmann, K. Jonsson, J. and Fisker, M. (2019), "Do no harm and do more good too: connecting the SDGs with business and human rights and political CSR theory", Corporate Governance, Vol. 19 No. 3, pp. 389-403.

Thèse en cours Ahmad Aidar, chaire Good in tech

Travaux de thèse d'Ahmad Aidar sur l'innovation numérique responsable (axe 1 chaire)

1. Soumission Article de recherche: Systematic Literature Review

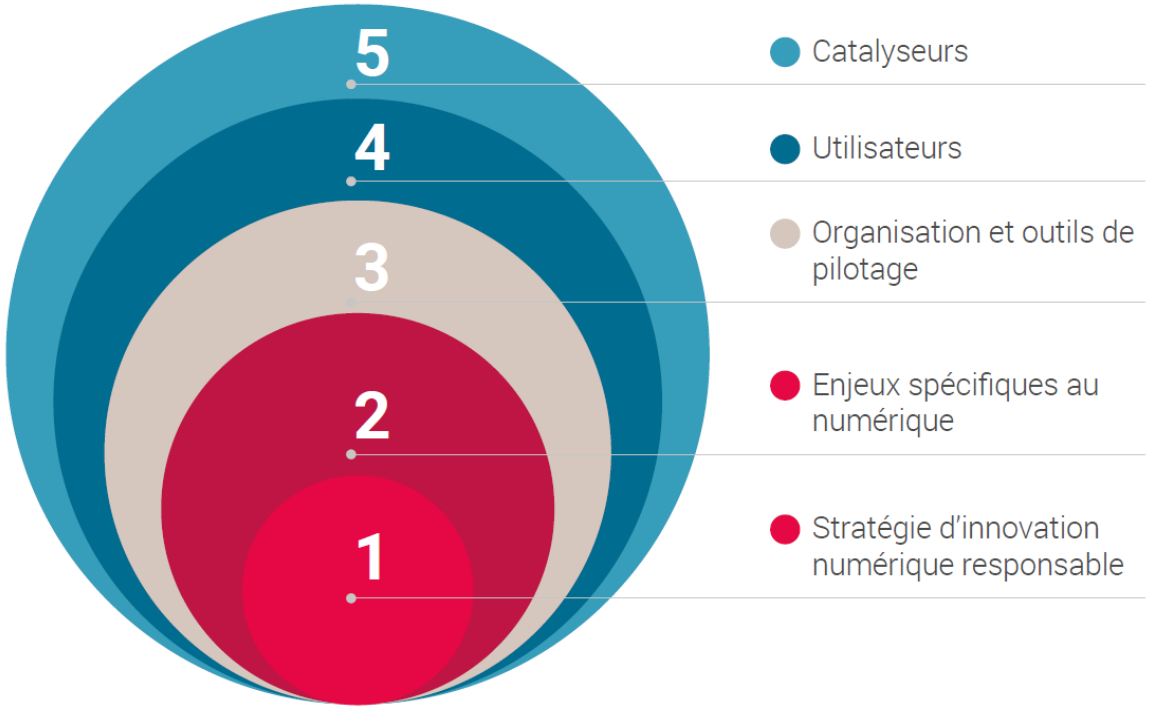
Ahmad Haidar, Christine Balagué: Responsible Innovation in AI: A Systematic Literature Review & Guidance for Future Research, *Special Issue Technovation,* 30 nov 2021

2. Terrain empirique sur des entreprises membres de Cap Digital

1. Etude qualitative auprès de 20 membres entreprises (grands groupes/start ups) membres de Cap Digital: identifier les dimensions de l'innovation numérique responsable

2. Etude quantitative auprès des 1000 membres entreprises de Cap Digital (en cours d'analyse)

A framework for Responsible Digital Innovation (Balagué Christine, Ahmad Haidar, article soumis à Technovation)

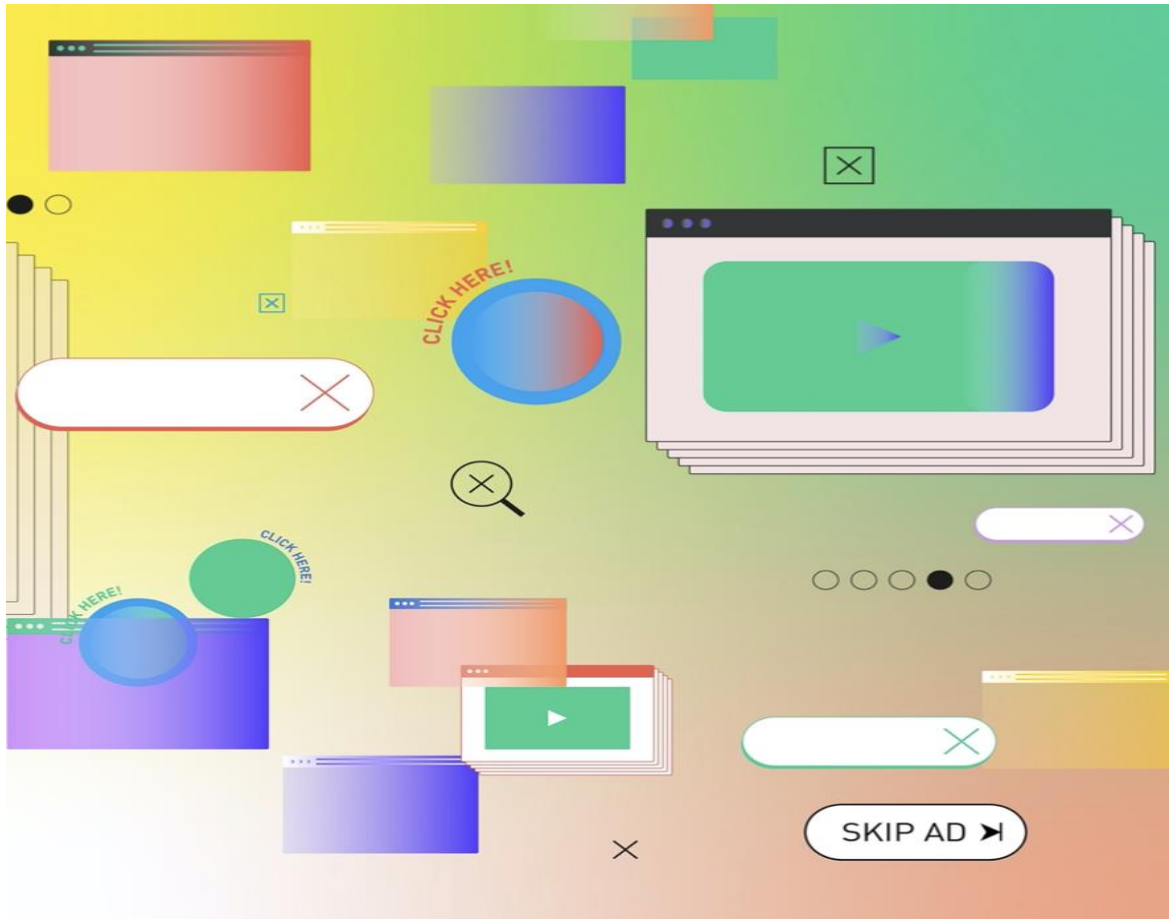


Digital Responsible Innovation (DRI) strategy	• Creating a mission-based status for the company • Integrating digital responsible innovation in CSR • Positioning digital product/service on DRI « by design » • Creating a corporate culture on DRI
Digital Specific Challenges	• Impact of digital innovation on environment • Data protection • Artificial Intelligence and Algorithms • Data and Information Systems security • Digital sovereignty • Inclusiveness
Organization and KPIs	• Collaboration and partnerships • Large-scale training • Selection of partners and suppliers • Human resources and organization • KPIs
Users	• Co-innovation with users • Transparency towards users • Technologies for users • Data collection frugality • Control back to users
Catalysts	• Top management support • Benchmarking • Regulation • Trial/error approach • Fundings

Responsible Digital Innovation Framework

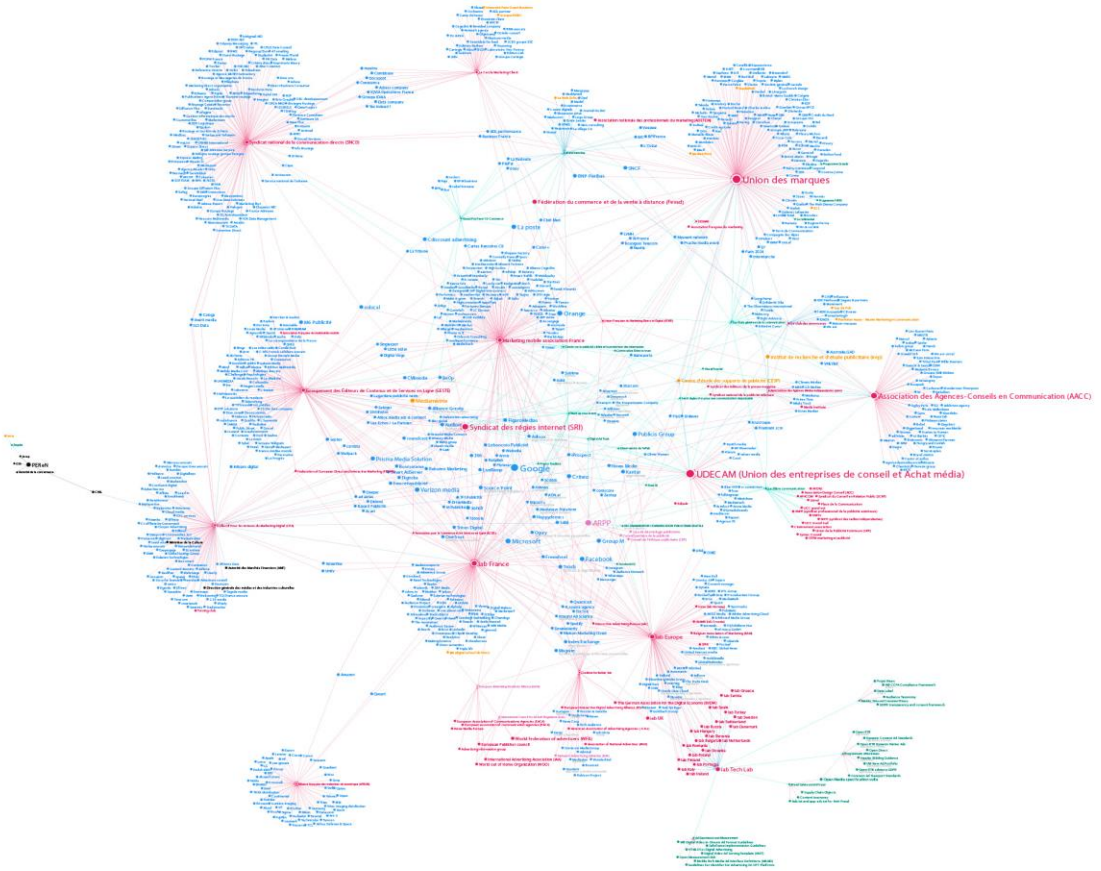
Publicité Numérique Responsable: rapport

Good In Tech



Publicité Numérique Responsable: cartographie des acteurs

https://www.goodintech.org/Cartographies-1.html



MERCI

@balague
christine.balague@imt-bs.eu