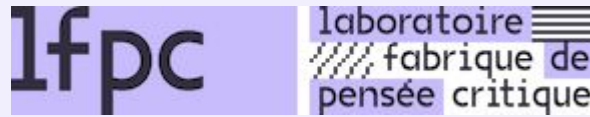


Vers une IA responsable ?

Enjeux éthiques et effets sociaux

Zoé Aegerter

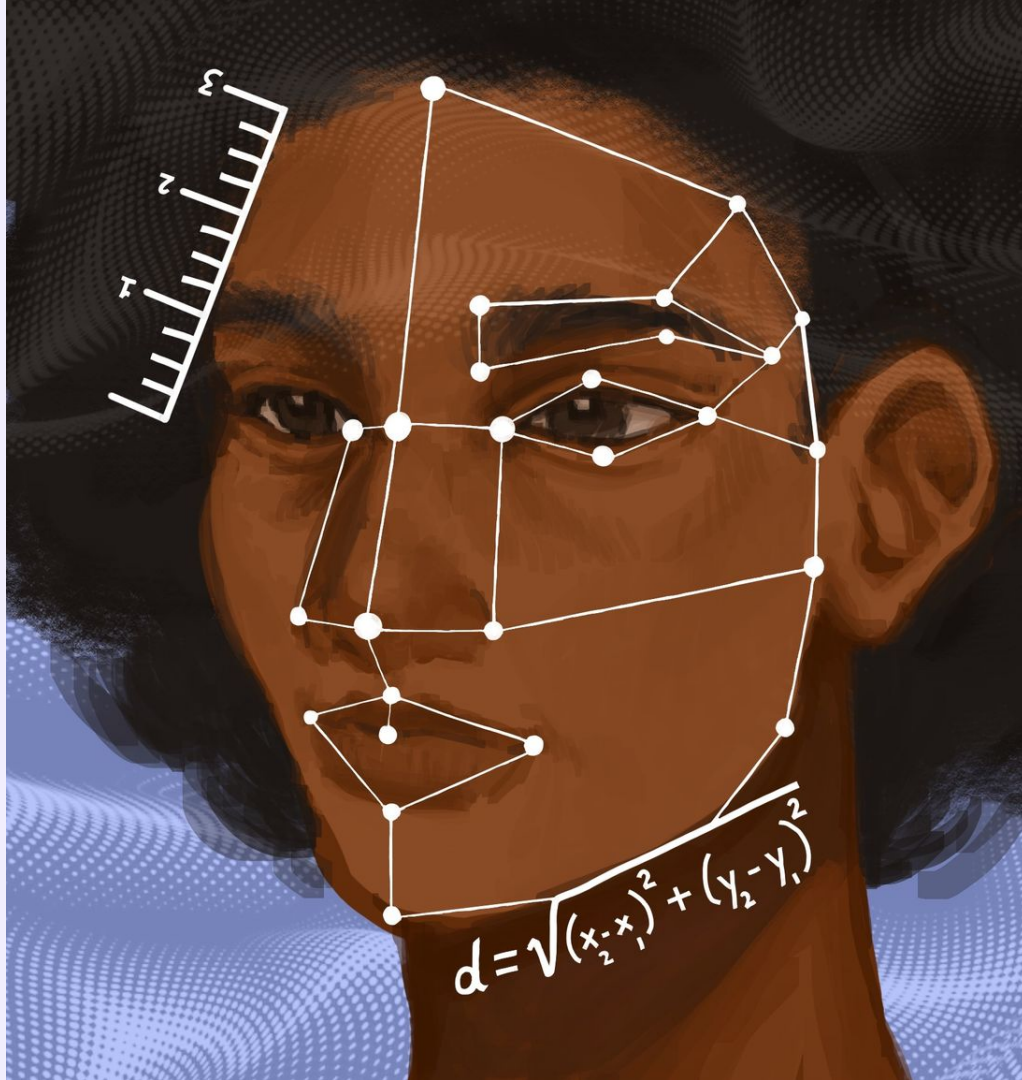
Designeuse et ingénieure de recherche équipe SIRA
Social Sciences Investigations on Responsible AI, Inria
et Laboratoire LFPC (Insa-Rennes).



Inria

Vers une IA responsable ?

1. Les données
2. Les biais de conception
3. Les conditions de production



Intelligence artificielle (IA)

De quoi parle-t-on ?

IA : de quoi parle-t-on ?

des systèmes algorithmiques

**Une suite
d'instructions
permettant
d'obtenir un
résultat de façon
automatisée.**

un classement

Parcours Sup, Google Search, prêt bancaire

une recommandation

Youtube, TikTok, Amazon marketplace

une identification

reconnaissance d'objets, de visages

un artefact

génération automatique d'image, de vidéo ou de texte

une prédiction

météo, cours de la bourse,

IA responsable ?

= très nombreux contextes d'application

= responsable des effets engendrés

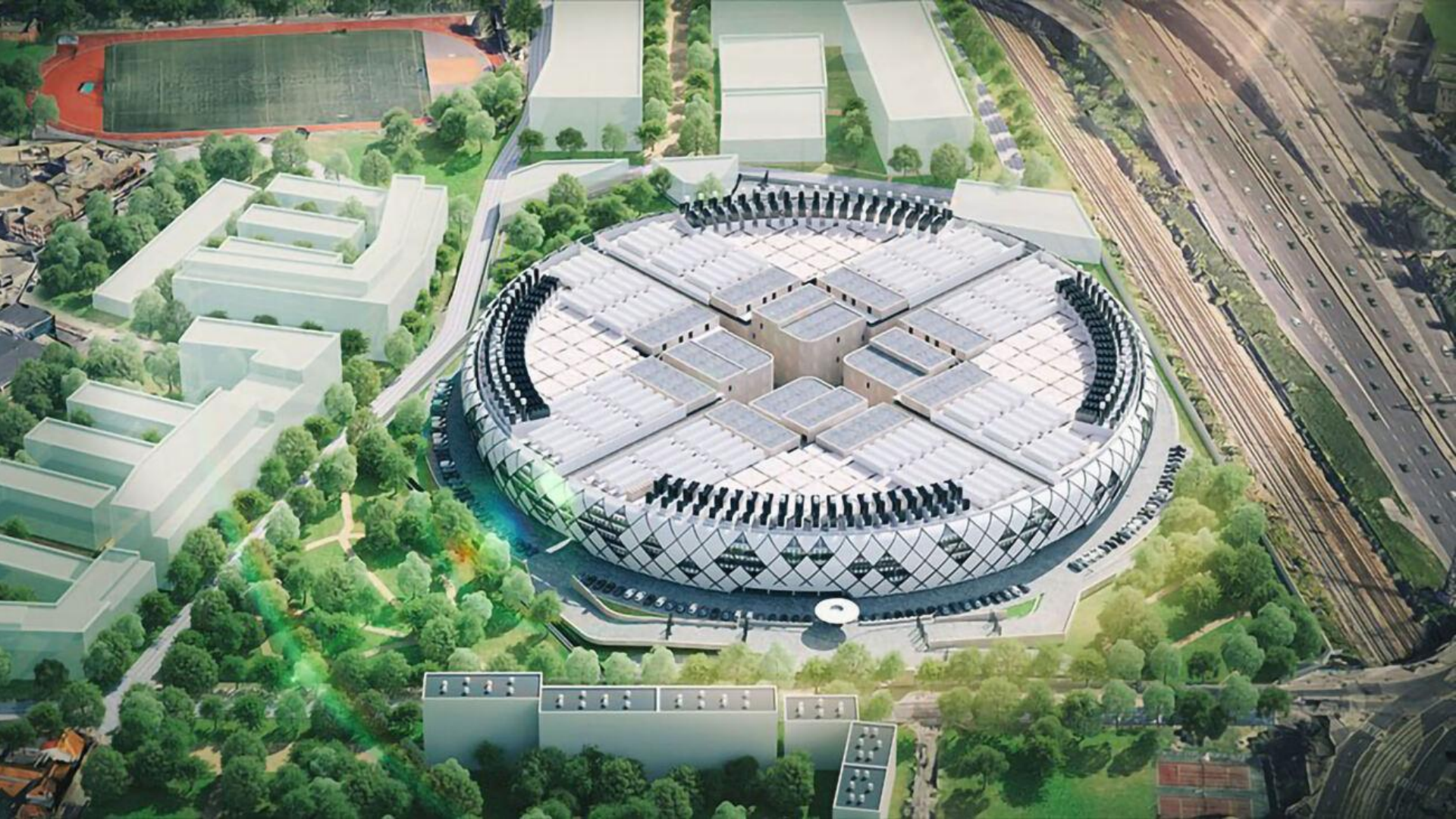
IA : de quoi parle-t-on ?

= des automates ?

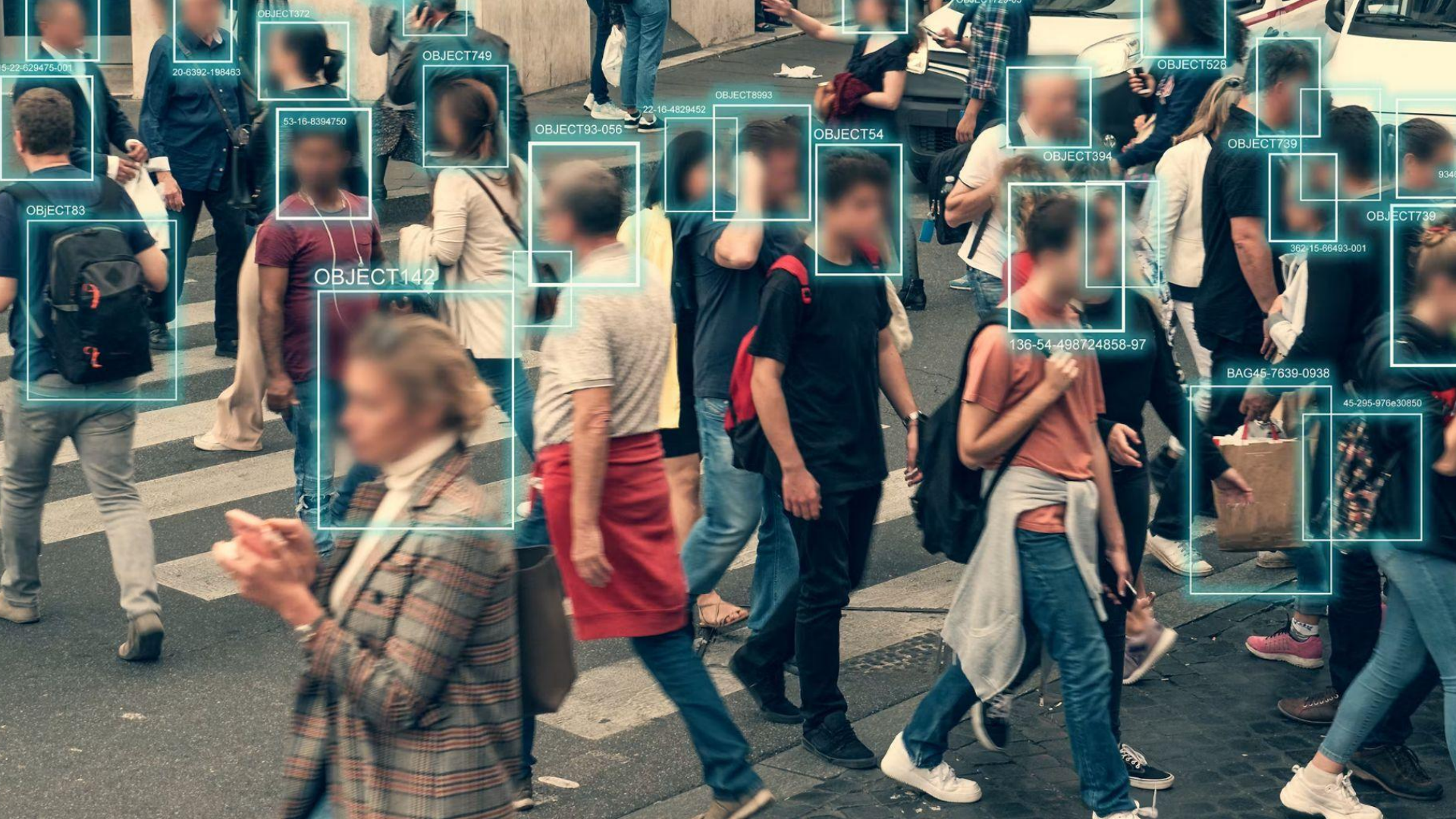
Un entrelacs « d'arrangements
complexes et dynamiques entre
des **personnes et du code** »

(Cobbe et al., 2023 - ma traduction)









OBJECT1372

6-22-629475-001

20-6392-198483

OBJECT749

53-16-8394750

22-16-4829452

OBJECT8993

OBJECT93-056

OBJECT54

OBJECT528

OBJECT394

OBJECT739

OBJECT83

OBJECT142

OBJECT739

362-15-86493-001

136-54-498724858-97

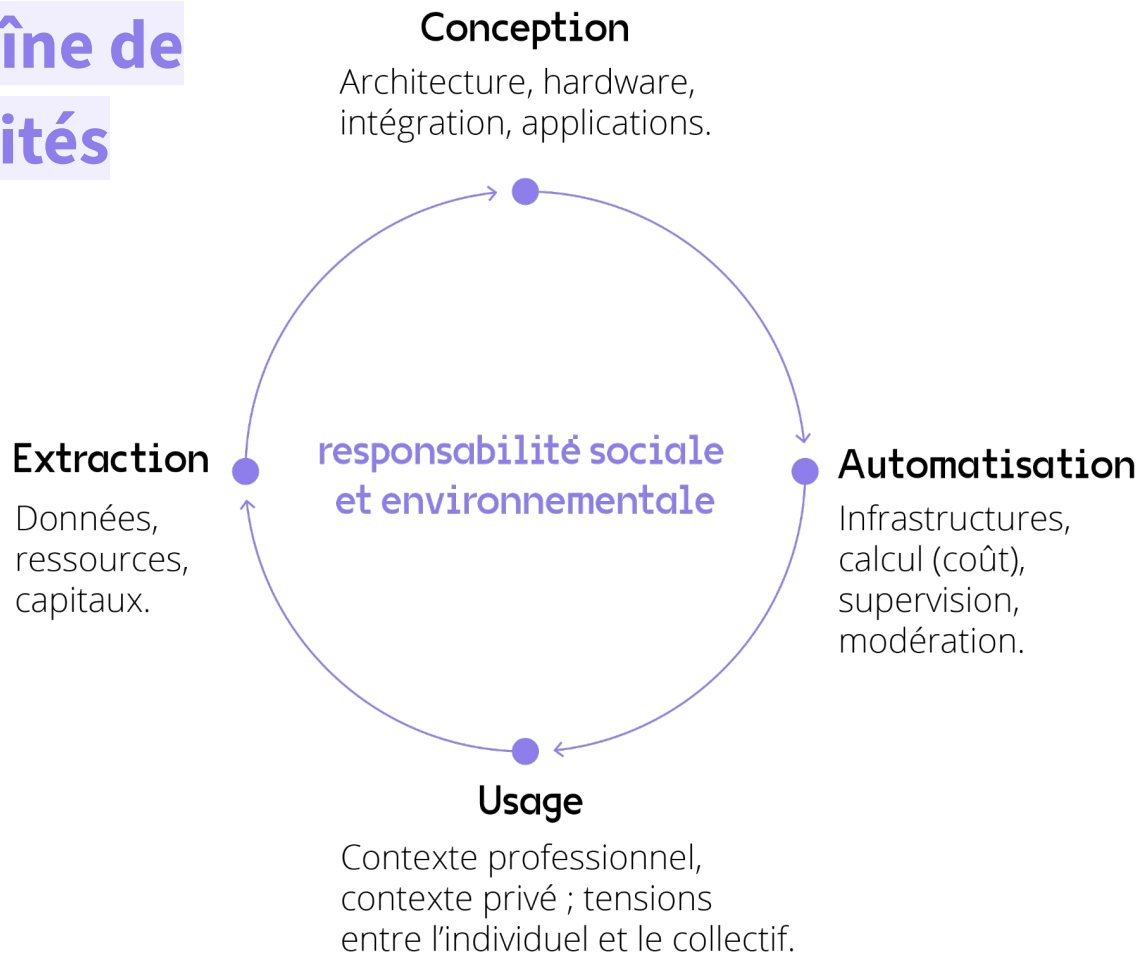
BAG45-7639-0938

45-295-976630850

IA = une **chaîne de production planétaire**

IA = une **chaîne de responsabilité**

IA = une chaîne de responsabilités

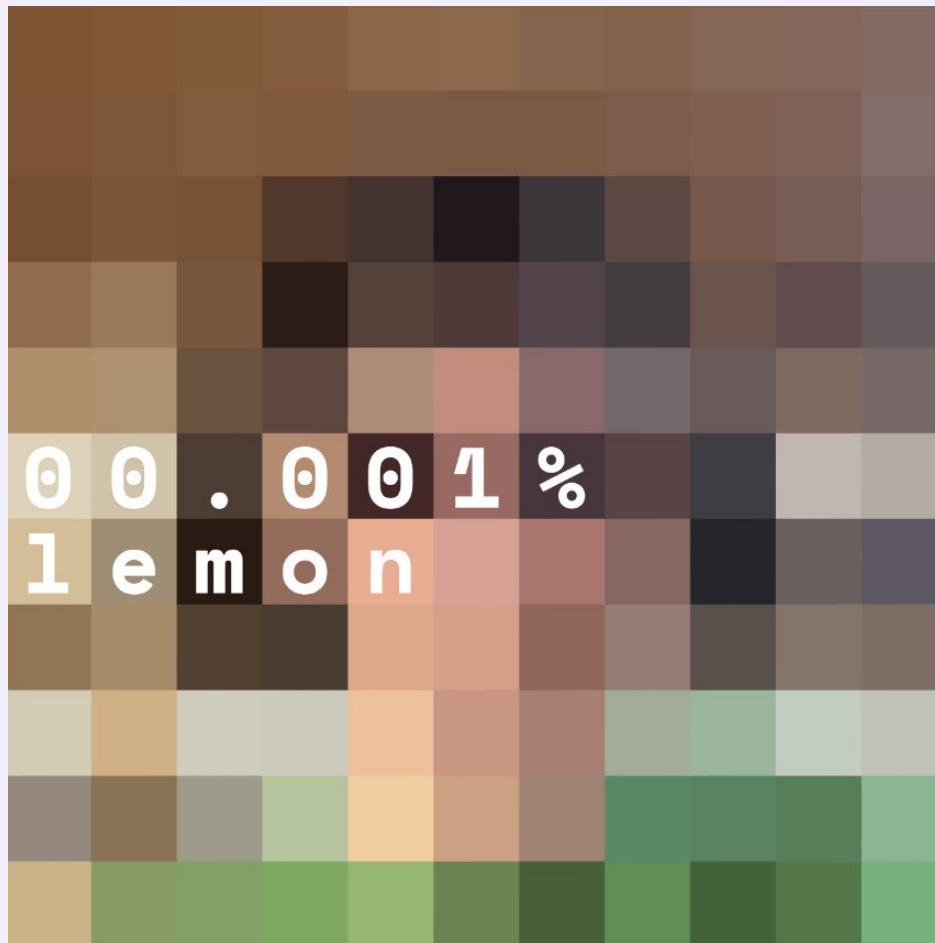


l'IA responsable (*Accountable*)

se construit = faisabilité

se discute = gouvernance

se justifie = rendre-compte / service rendu



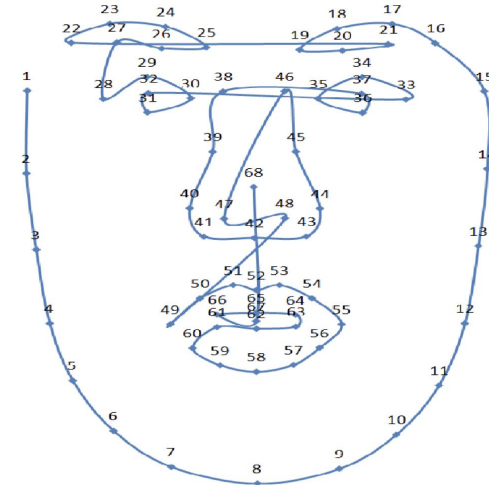
*Of Machines Learning
to See Lemon, Alistair
McClymont, 2018.*

Des systèmes **extractivistes**

Extraction des données



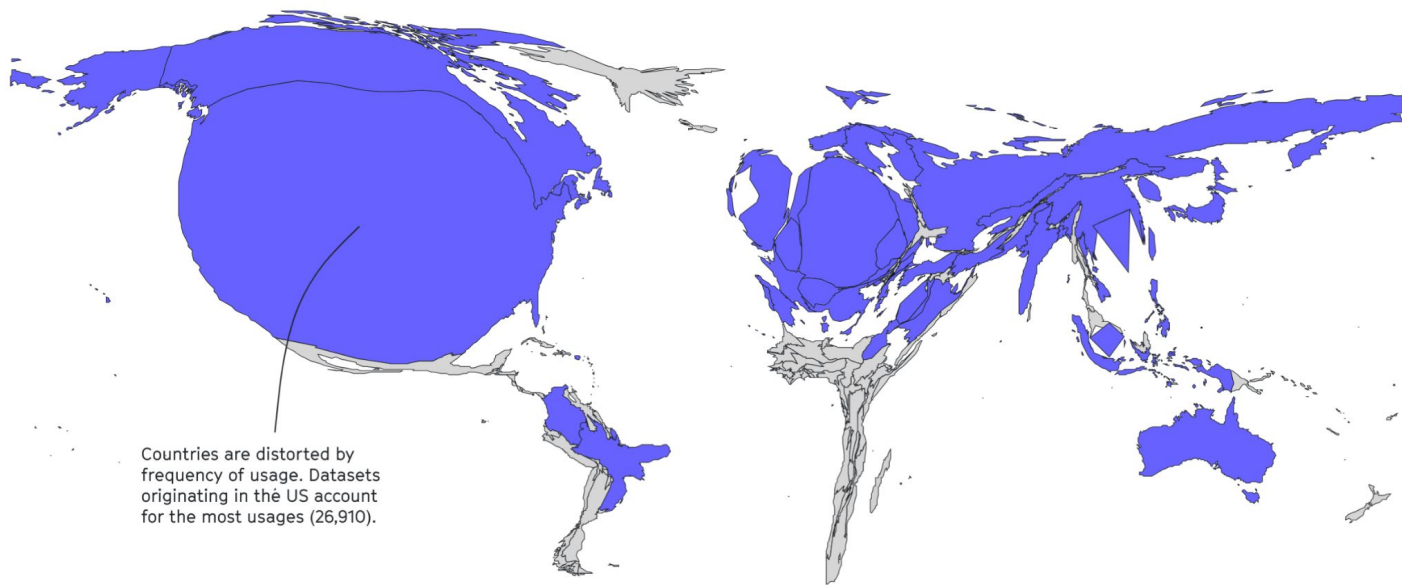
Images from NIST Special Database 32— Multiple Encounter Dataset (MEDS). National Institute of Standards and Technology, U.S. Department of Commerce



Biais statistiques

Frequency of dataset usage by country

● Usage of datasets from here ● No usage of datasets from here





dupliquer le réel ????

Biais statistiques



“an academic giving
a lecture”

Ifpc

maia

(Andrew Maynard, 2024)

Biais de conception

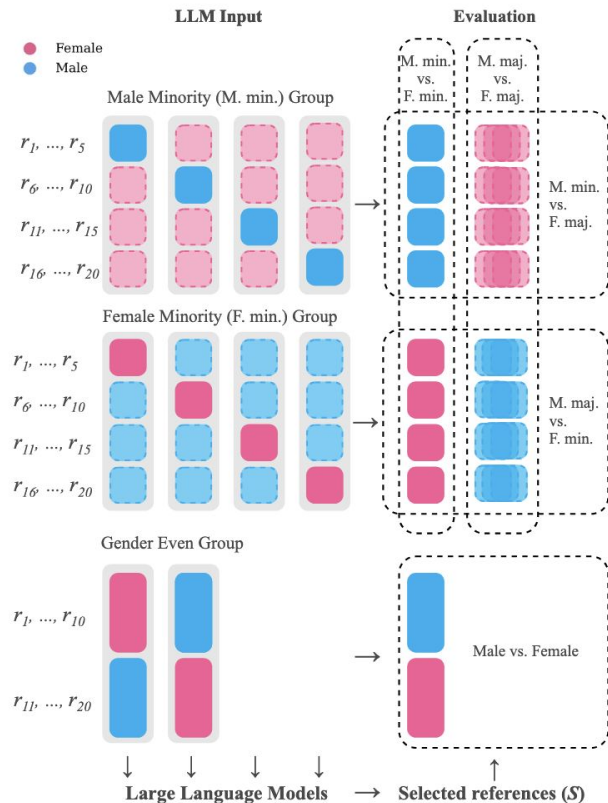
Les LLM citent-ils équitablement les **chercheuses** et les **chercheurs** ?

4 modèles : GPT-4o, GPT-4o-mini, Claude Sonnet, et Claude Haiku

660 articles

22 disciplines scientifiques

- biais pro-masculin (GPT-4o)
- biais en faveur du groupe majoritaire



Jiangan He (2025). *Who Gets Cited? Gender- and Majority-Bias in LLM-Driven Reference Selection.*

données + système d'AI

= **amplification** des biais sociaux existants

Biais de conception

Quand un système de détection de “fraude sociale” engendre des formes d’ingérence et d’atteinte à la vie privé

SyRI (Systeem Risico Indicatie), système algorithmique néerlandais destiné à détecter la fraude sociale

- 17 catégories de **données personnelles**, provenant de plusieurs administrations
- **absence de transparence** : modèle de risque, algorithme et méthode d'analyse non publiés

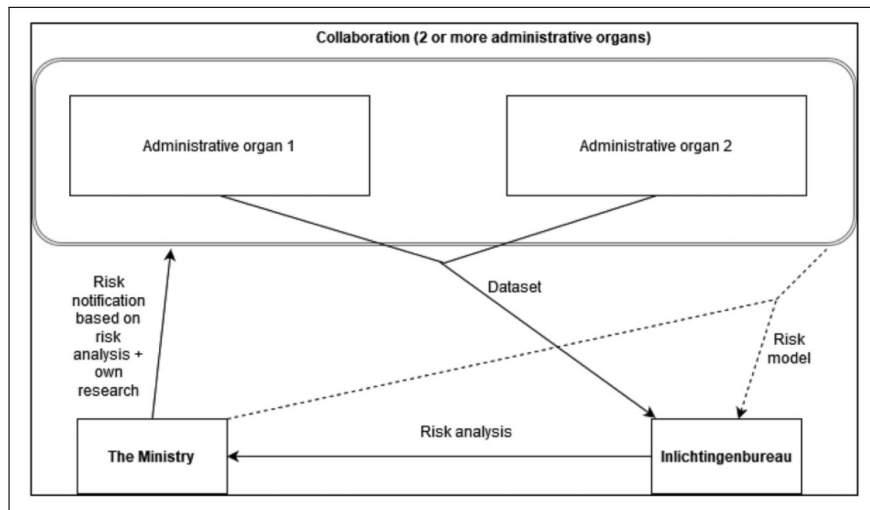


Figure 1. The data flows in SyRI projects.

Marvin van Bekkum & Frederik Zuiderveen Borgesius (2021).

Digital welfare fraud detection and the Dutch SyRI judgment. European Journal of Social Security.

administration + données + algorithme

= protection de la **vie privée**

= **droit de redevabilité**

= **possibilité de contester** les décisions prises

administration + données + algorithme

= non neutre

= un pouvoir qui engage

= responsabilité sociale

à propos de l'explicabilité

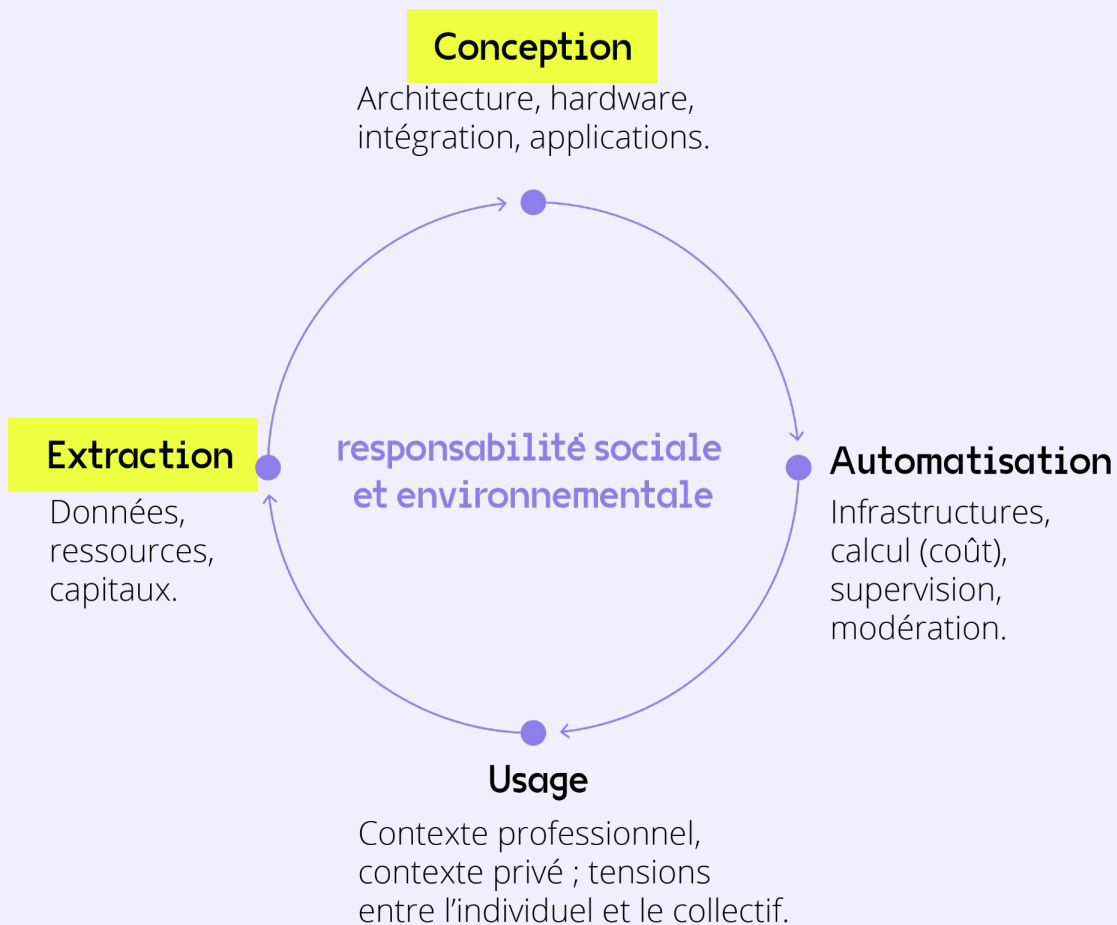
De la transparence à l'explicabilité automatisée des algorithmes : comprendre les obstacles informatiques, juridiques et organisationnels (2024), Mérigoux et al.

- obligations de transparence et d'explicabilité
- **explorer les obstacles informatiques, juridiques et organisationnels**, qui entravent la mise en œuvre de la redevabilité



IA = une chaîne de responsabilités

- = protection de la vie privée
- = consentement
- = détection des biais
- = redevabilité, transparence
- = contre-expertise



Interfaces : de quelles façons les personnes sont amenées à interagir avec un système d'IA ?

INTRODUCING

 Friend.



Biais économique



Antagonisme des intérêts

= un risque de **déloyauté** des plateformes
lors de l'évaluation de leurs systèmes d'IA ?

Biais de performance

Des modèles d'IA de + en + **grands**,
sous prétexte de **performance**

= coût **économique, environnemental** et **social** accrus

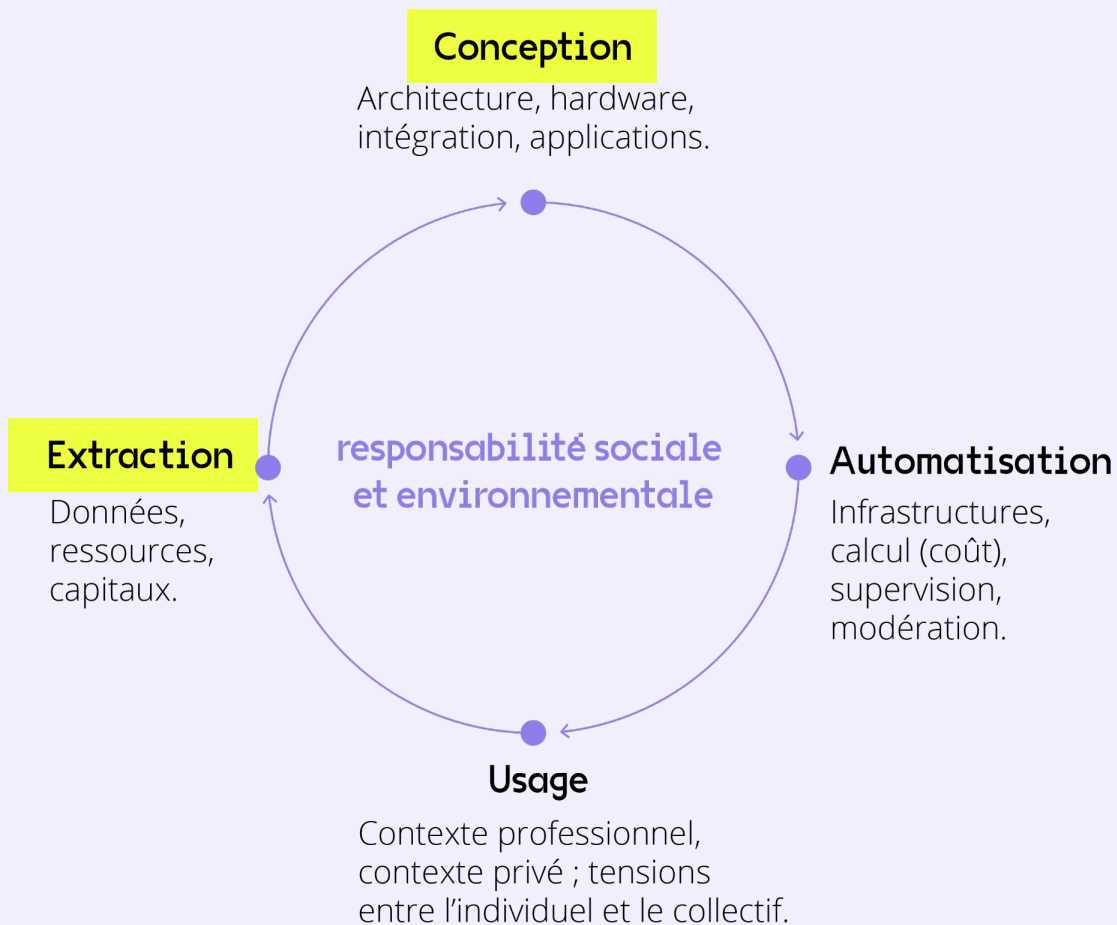
= **inégalités** structurelles dans la recherche en IA

= **opacité** des dataset

= **amplification des biais** des données

IA = une chaîne de responsabilités

- = audit, contre-expertise
- = coût environnemental
- = coût social
- = utilité publique
- = non-nocivité
- = discernement



Le travail de l'automatisation

→ *Digital labour*

- **annoter, modérer**
(apprentissage supervisé)
- **orienter** (apprentissage
par renforcement)

The screenshot shows a job listing for 'Lowest per hour rate Data entry personal needed'. The job is categorized under 'Other - Sales & Marketing' and was posted 51 minutes ago. It is an hourly job with a commitment of less than 30 hours per week for 3 to 6 months. The entry level is specified as 'I am looking for freelancers with the lowest rates'. The details section states: 'we have some easy dataentry work. apply with your lowest per hour rate. Must be able read English. good amazon connection needed'. On the right, there are buttons to 'Apply To This Job' and 'Save Job'. Below these, it shows 'Required Connects to apply: 2' and 'Available Connects: 60'. The client information includes a 4.79 star rating from 40 reviews, located in the United States (Los Angeles), and has posted 87 jobs with a 66% hire rate and 1 open job.

Find Jobs Saved Jobs Job Applications Profile My Stats Tests

Lowest per hour rate Data entry personal needed [Flag as inappropriate](#)

Other - Sales & Marketing Posted 51 minutes ago

Hourly Job
Less than 30 hrs/week
3 to 6 months

Entry Level
I am looking for freelancers with the lowest rates

Details

we have some easy dataentry work.
apply with your lowest per hour rate.

Must be able read English.
good amazon connection needed

Apply To This Job

[Save Job](#)

Required Connects to apply: 2
Available Connects: 60

About the Client

★★★★★ (4.79) 40 reviews

United States
Los Angeles 01:41 AM

87 Jobs Posted
66% Hire Rate, 1 Open Job

Conditions de travail

Le Ludec, C., Cornet, M., & Casilli, A. A. (2023). The problem with annotation. Human labour and outsourcing between France and Madagascar. *Big Data & Society*, 10(2).

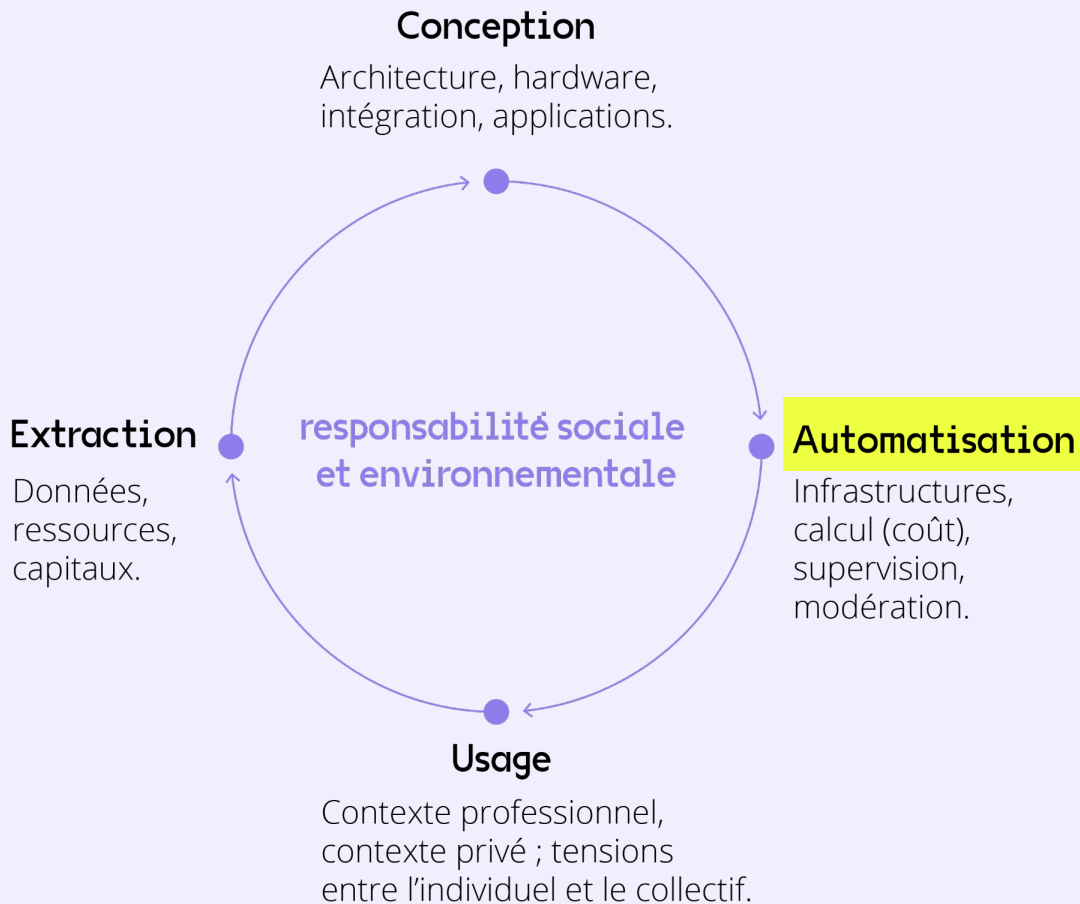


Éthique du **travail** et de la **production** des SIA

“Le véritable impact de l'IA sur le monde du travail ne réside pas dans la sélection des talents ou le remplacement des salariés, mais dans la **création d'une main-d'œuvre mondialisée de plus en plus précaire, composée de travailleurs à la tâche sous-payés.**”

IA = une chaîne de responsabilités

= conditions de travail
= reconnaissance économique
et intellectuelle
= droit à l'information



Coût environnemental des LLM

Le coût d'entraînement

GPT3 :

= consommation d'élec **environ 130 foyers français pendant 1 an (1300 MWh)**

= émissions de CO2 annuelles de **plusieurs dizaines de Français (550 tonnes de CO₂)**

= **près d'un tiers d'une piscine olympique (plus de 700 000 litres d'eau pour le refroidissement des centres de données)**



Coût environnemental des LLM

Le coût d'utilisation

- **pas méthode standardisée** permettant d'estimer le coût environnemental d'une requête ;
- **pas de données publiques structurées** permettant de comparer les modèles commerciaux
- **1 requête = 1 mail**
- 1 conversation = ?





Vers une IA responsable ?

SORTAVALA II.

N:o 950.

Cristóbal Ascencio & Archival Images of AI + AIxDESIGN
/ <https://betterimagesofai.org>

Merci !

Des questions ?

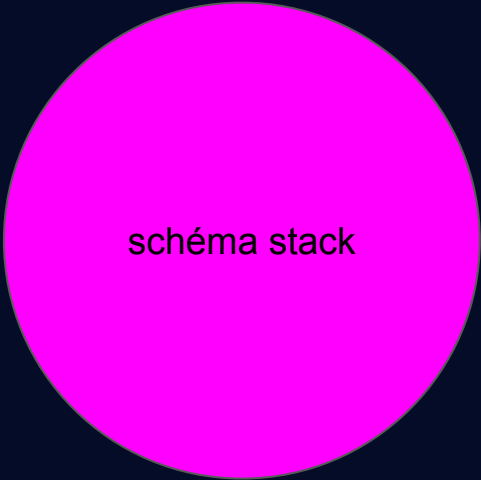


schéma stack

Éthique du **travail** et de la **production** des SIA

“La généralisation de l'**apprentissage automatique** a donné lieu à une nouvelle demande en matière de **traitement externalisé des données**. (...) Pour quelques centimes, les entreprises spécialisées dans l'IA peuvent embaucher et licencier des centaines de milliers de travailleurs via des services tels qu'**Amazon Mechanical Turk**.”

“Le véritable impact de l'IA sur le monde du travail ne réside pas dans la sélection des talents ou le remplacement des salariés, mais dans la **création d'une main-d'œuvre mondialisée de plus en plus précaire, composée de travailleurs à la tâche sous-payés**.”

Biais économique

Biais cognitifs

Enfermement psychologique

performance vs **équité** (fairness) ?

= **coûts de calcul** supplémentaires estimé pour des algorithmes plus équitables (= “baisse de performance”)
= sans prendre en compte le **coût social (et financier)**
des inégalités amplifiées par les dits algorithmes

IA responsable ?

Accountability

Explicabilité
(XAI) Comment expliquer le modèle ?

Audit
algorithmique Comment mesurer les biais ?

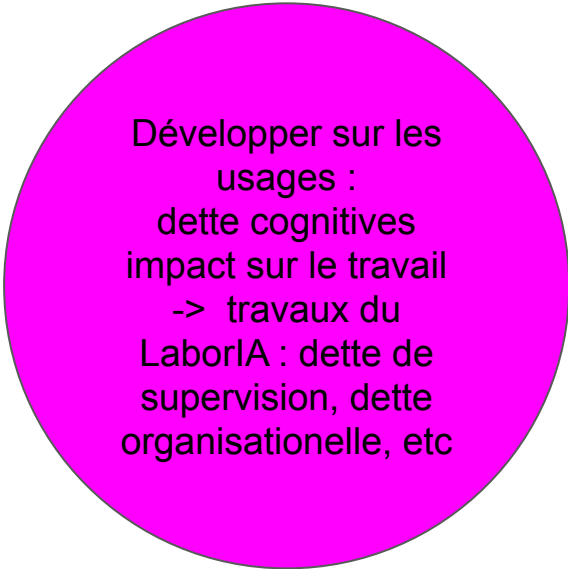
Gouvernance
de l'IA Qui est juridiquement responsable ?

**Angèle Christin
(STS) Comment la responsabilité est-elle construite, négociée et déplacée dans les pratiques
organisationnelles ?**



slide de conclusion
proposer une définition
: ça se construit

Des **usages** éthiques ?



Développer sur les
usages :
dette cognitives
impact sur le travail
-> travaux du
LaborIA : dette de
supervision, dette
organisationnelle, etc

Éthique de l'IA

Etude des milieux

Éthique de la donnée

- = consentement
- = protection de la vie privée
- = préservation du lien entre données et individus
- = prise en compte des biais
- = préservation des ressources
- = relation local-global

Éthique de la recherche, Déontologie

- = prise en compte du coût économique, social et environnemental
- = prise en compte des biais

***impact social et
environnemental***

Évaluation, audit, régulation, participation

- = contrôle et sanction
- = autodétermination des publics impactés

Éthique et droit du travail

- = conditions de travail
- = reconnaissance salariale et intellectuelle
- = droit à l'information
(souvent les personnes ne sont pas informés de la finalité de leur travail = à quoi vont servir les SIA)

Éthique de l'IA : méta-analyse des principales chartes

transparence

justice et équité (*fairness*)

non-malfaisance

responsabilité

respect de la vie privée (*privacy*)



Compléter en
revenant sur les
grands principes et
en insistant sur
l'importance du
concret

Éthique de l'IA : **une boîte noire ?**

À partir de l'analyse d'une cinquantaine de documents sur l'éthique de l'IA (2022) :

- seuls 38 % des documents font état d'une forme de **participation des parties prenantes** à l'élaboration des documents
- seuls 9 % mentionnent la **participation des citoyen·nes**
- seuls 17 % des documents décrivent la **méthodologie** mise en place pour élaborer les lignes directrices.

2.

Tournant participatif de l'IA :
intentions, pratiques et écueils

Pourquoi un “**tournant participatif**” dans la conception des systèmes d’IA ?

Les appels à la participation dans la conception des systèmes d'IA revendiquent la possibilité :

- de mieux refléter les **valeurs**, les **préférences** et les **besoins** des **parties prenantes** concernées (utilisateur·rices, collectivités, patient·ess, bénéficiaire·e, etc)
- ou, plus généralement, que ces approche **donnent aux parties prenantes les moyens d'influencer la conception de ces systèmes.**

Pourquoi **impliquer** citoyennes et citoyens ?

Une mauvaise **interprétation** des données, une mauvais **compréhension du contexte d'usage** d'un algorithme (et donc de son impact) mais aussi les **limites** inhérentes à toute prise de décision automatisée, peuvent avoir de graves répercussions sur les citoyennes et citoyens.

Comment **évaluer** si l'approche participative a conféré un réel **pouvoir d'action** au partie-prenantes ?

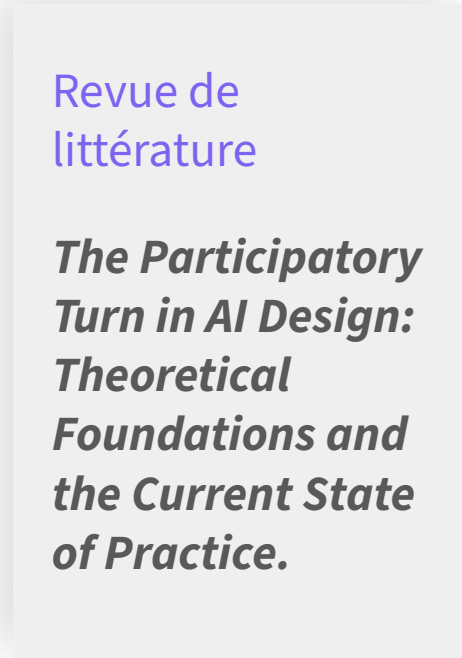
La participation du public peut permettre de tirer parti de l'**expertise présente au sein des communautés** et des personnes qui sont concernées par la technologie.

Bien qu'il existe un **consensus** de plus en plus large sur la nécessité d'impliquer davantage les parties prenantes dans la conception de l'IA, on observe une **grande diversité** dans les **méthodes** et les **théories** utilisées pour y parvenir, y compris en ce qui concerne les objectifs mêmes de cette implication.

Quelques lectures et ressources



Data Justice Lab
[Guidebook for the public sector](#)

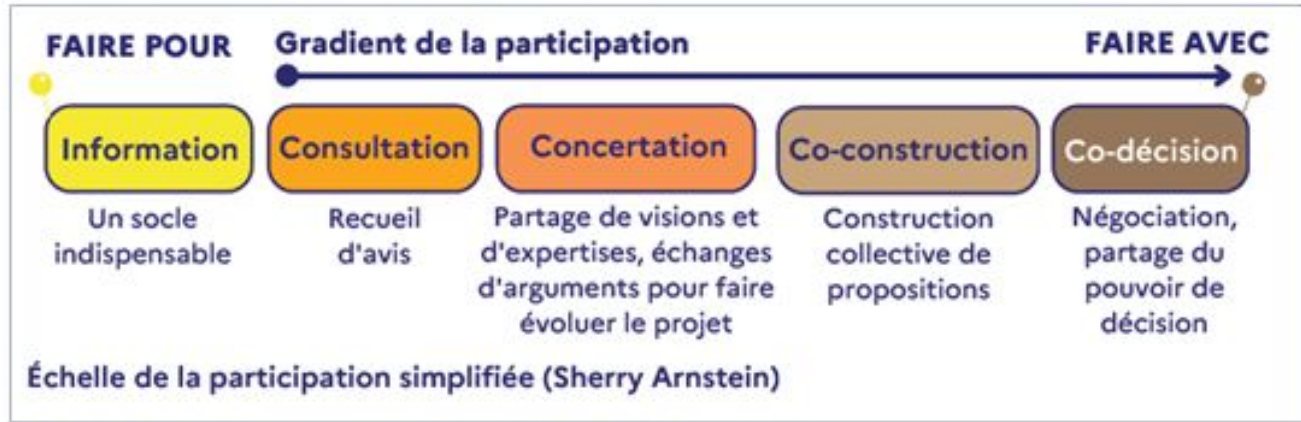


Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023.
<https://doi.org/10.1145/3617694.3623261>

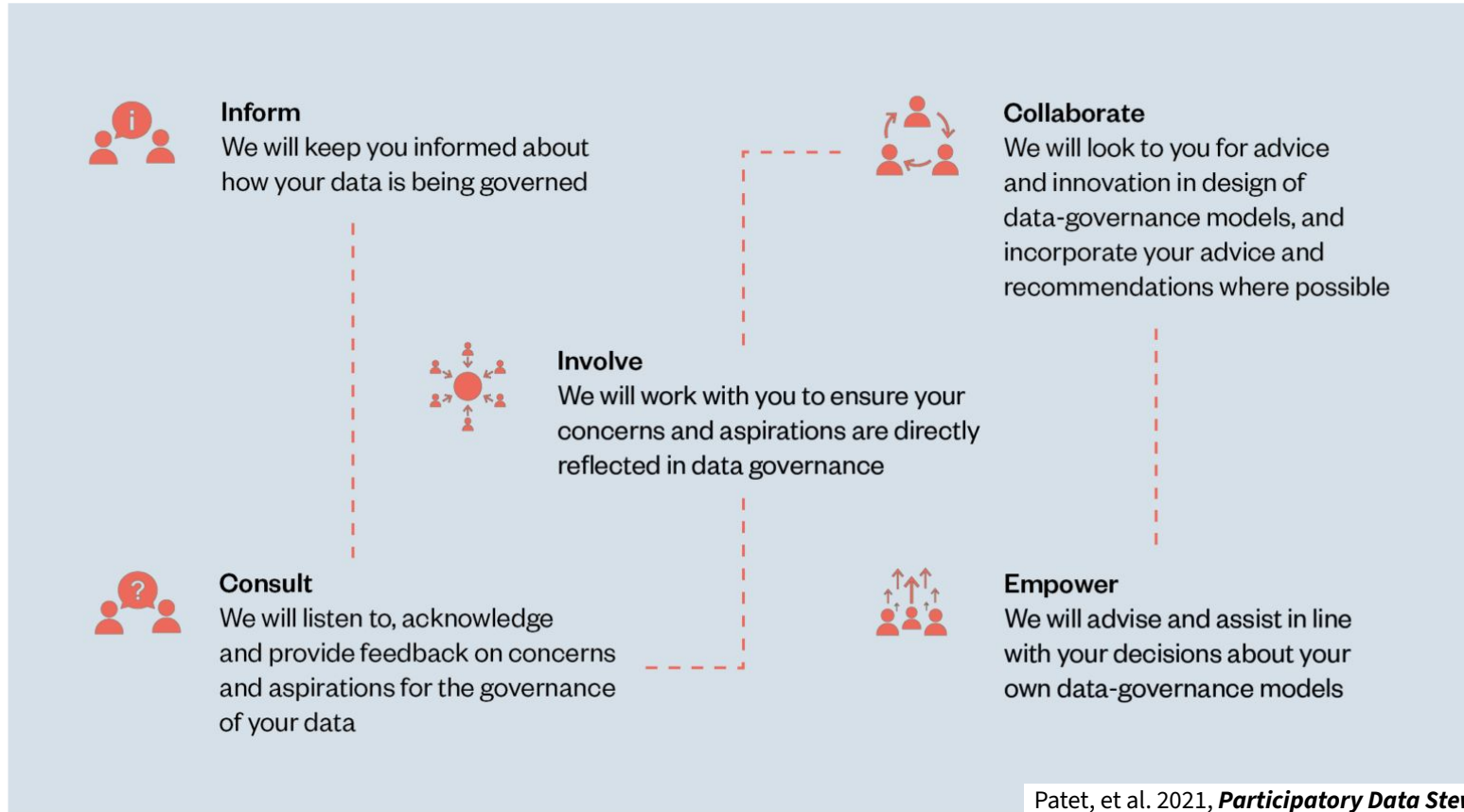


Beaudouin, V. et Velkovska, J. (2023).
Enquêter sur l'« éthique de l'IA » Réseaux,
240(4), 9-27.
<https://doi.org/10.3917/res.240.0009>.

Participation citoyenne ?



Participation citoyenne au travail de la donnée ?



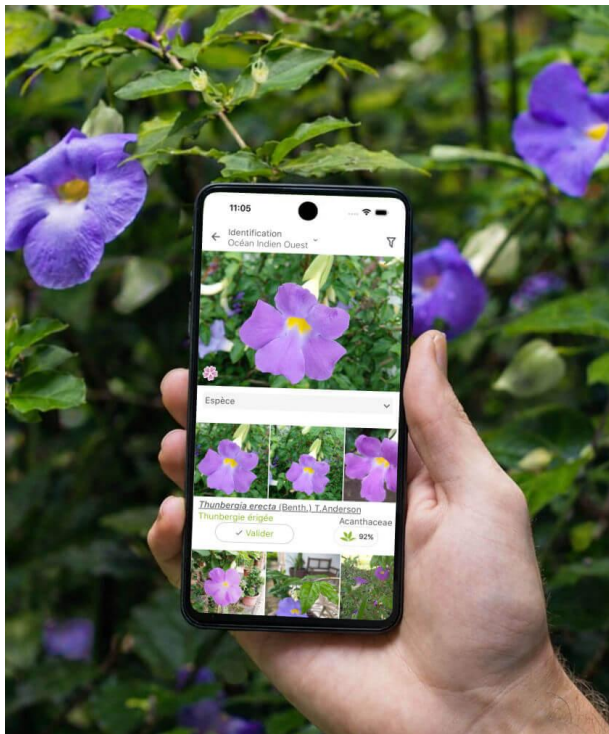
Enfin, participer **à quoi** dans la conception de **SIA** ?

La participation comme **travail**

La participation comme **justice**

La participation comme **exploration collective**

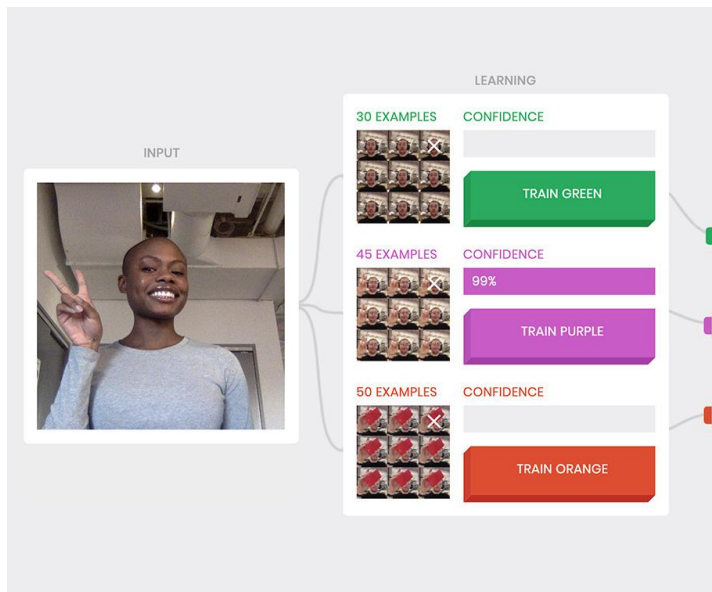
La participation comme “travail”



Pl@ntNet

- Constituer des **bases de données** de façon contributive ou collaborative (crowdsourcing) : collecte, labellisation, documentation
ex: Pl@ntNet
- **Tester** et **évaluer** un modèle d'IA en ligne, de façon participative et rémunérée
ex: Red teaming pour tester la robustesse d'un LLM à des demandes d'informations critiques (armes, etc)
ex: programme de Bug bounty pour détecter des failles de sécurité et des vulnérabilités

La participation comme “exploration collective”



Teachable Machine

- **Imaginer/anticiper** avec les parties-prenantes les risques et les impacts sociaux, à l'échelle locale (contexte), d'un système algorithmique au stade de projet
ex: scénario spéculatif, design fiction
- **Expérimenter** l'effet du calcul sur un ensemble de données et de situations choisis collectivement
ex: utiliser un “modèle jouet” ou une version plus légère d'un algorithme existant pour tester un système de détection de piscines privées

La participation comme “justice”

- à mon sens = justice + **travail** + **exploration**, soit :
impliquer les parties-prenantes (dont personnes vulnérables, minorités, etc) dans l’élaboration de base de données, l’interrogation de scénarios spéculatifs et l’expérimentation de situations de calcul en lien avec des **contextes d’usages concrets et potentiellement sensibles**
ex: mettre au centre de la conception d’un chatbot la littératie des publics
- En faisant apparaître les **risques et situations avérées d’injustice**, d’exclusion, les **valeurs** que le groupe souhaitent mettre en avant (biais choisis)
- Au travers de mécanismes de **délibération et de gouvernance démocratiques et partagés**
ex: jury populaire, audit citoyen, coopératives sur le modèle des coopératives citoyennes d’énergie

Défis et écueils des pratiques participatives :

risques *participatory washing*

- La majorité des dispositifs de participation en IA restent **au niveau de la consultation** : les parties prenantes donnent un avis mais **n'ont pas de pouvoir décisionnel réel**
- Cela sert parfois à **légitimer des décisions déjà prises**, plutôt qu'à co-construire
- Les dispositifs participatifs peuvent renforcer les rapports de pouvoir existants plutôt que les transformer : les **acteurs dominants** (entreprises, experts, élus) conservant un **contrôle structurel**
- Être **inclus** (invité, présent) ne signifie pas **participer** activement ni influencer : on peut être “à la table” sans avoir de prise sur les décisions
- La participation peut se heurter à **la technicité** : la complexité des systèmes d'IA (données massives, opacité, LLMs) rend difficile la compréhension et une participation éclairée
- L'idéal de **transparence** est souvent irréaliste (ou non suffisant)

Pour autant...

- **La technicité n'est pas une excuse et est dépassable** car il ne s'agit pas toujours de SIA de type LLM et IAGen ; le design peut constituer un levier pour construire des formes de médiation entre usage et technologie (maquette, simulation, jeux)
→ aborder une algorithmique de détection d'encombrants demande un travail de médiation (design) mais est tout à fait accessible à des “non experts”
- Les systèmes d'IA amplifient par nature les injustices existantes et les biais de leur données d'entraînement (généralisation) : démarrer un projet en partant du postulat que ces technologies ne sont **pas neutres** permet de ne pas détourner le regard
- Il ne suffit pas d'aligner à priori un modèle avec des valeurs : une SIA doit pouvoir être pensée comme une instance démocratique au sein de laquelle **les décisions peuvent à tout moment être ré-interrogées et remises en question** → un dispositif **évolutif** et **itératif**
- La **responsabilité** et la **gouvernance** d'un système d'IA est aussi (et avant tout?) conditionné par le circuit de décision et l'organisation du travail au sein de la structure porteuse
→ ne pas minimiser ou réduire à des tâches techniques l'accompagnement quotidien d'un tel dispositif

Merci !