



**LE FÉMINISME  
DES DONNÉES**



**Data**

**Feminism**

**Catherine D'Ignazio and**

**Lauren F. Klein**

Cet article est une lecture synthétique et en français du livre de Catherine D'Ignazio et Lauren Klein, *Data Feminism*, publié en 2020 au MIT Press.

# POUR UN FÉMINISME DES DONNÉES

Le féminisme des données est à la fois un mouvement et un mode de pensée qui s'oppose à la manière dont le monde des datasciences considère les données, comme un matériel à exploiter. Il propose à la science des données et au mouvement pour l'ouverture des données qui le structure de se réinventer, de prendre en considération d'autres valeurs pour partager et exploiter les données. Le féminisme des données n'est pas qu'une critique, il est d'abord et avant tout un programme pour remettre le consentement au cœur de l'échange de données et construire une science des données qui émancipent les utilisateurs plutôt qu'elle ne les exploite. Retour sur un concept novateur et peu connu, une notion politique et manifeste qui propose rien de moins que d'inverser la façon dont on considère les données.

# UNE SCIENCE DES DONNÉES POUR L'ÉMANCIPATION !

En 2020, les chercheuses Catherine D'Ignazio et Lauren Klein publient *Data Feminism* (MIT Press, 2020, non traduit, disponible en publication ouverte), un livre-manifeste qui propose une vision politique des données : le féminisme des données.

Pour ces deux spécialistes en science des données, la discipline a besoin du féminisme pour se réinventer. Le féminisme des données est « *un mode de pensée sur la donnée et sa communication informés par l'expérience directe, par un engagement à l'action et par les idées associées à la pensée féministe intersectionnelle* », qui vise à déconstruire les systèmes de pouvoir, de privilège et d'oppression. A l'heure où les données façonnent le monde et notre relation au monde, il est temps de s'intéresser très précisément à celles-ci, soulignent D'Ignazio et Klein.

**En cela, le féminisme des données n'est pas seulement un livre, il se veut aussi un mouvement, une réflexion politique, un mode d'action sur le monde.**

C'est un livre sur qui a le pouvoir et qui ne l'a pas, un livre sur les conséquences de ces différentiels de pouvoir pour les mettre en évidence et les changer.

La science des données a besoin du féminisme pour mettre fin à la spirale d'injustices qu'elle

participe à bâtir, expliquent D'Ignazio et Klein. Pour les deux chercheuses, les corps manquent dans les données qu'on collecte. La science des données pense qu'elle peut produire le monde sans les convoquer. Que son objectivité et sa neutralité nécessitent d'appliquer une froide raison sur tout calcul. C'est tout le contraire dont nous avons besoin, expliquent les chercheuses : les problèmes structurels ne peuvent être exposés qu'en les regardant sous l'angle spécifique des personnes et des corps. Nous devons compter ce qui ne l'est pas, notamment parce que les données et le pouvoir s'alignent trop souvent. Ramener les corps des gens, des femmes, des personnes de couleurs, des marginalisés dans les discussions et les décisions, suppose bien sûr, comme le disait déjà Sasha Costanza-Chock dans son livre *Design Justice*, de savoir quelles données sont collectées, par qui et pourquoi.

L'histoire est ancienne, mais reste fameuse : quand Andrew Pole, data scientist chez Target, a mis au point un algorithme de prévision de la grossesse des clientes pour favoriser des campagnes marketing dédiées, il a montré que les corps peuvent être exploités sans que les personnes visées n'aient leur mot à dire sur cette exploitation. Target a utilisé son capital de données pour consolider son pouvoir sur ses clientes, sans leur consentement. Les décisions prises du point de vue des données sont trop souvent utilisées pour amplifier les inégalités et asymétries d'information que le contraire. En se basant sur une liste de produits que les femmes

enceintes achètent, Target était capable de calculer un score de prévisibilité de grossesse, que l'entreprise a utilisé sans tenir compte des impacts et conséquences sur ses clientes. C'est justement sur les conséquences des décisions qui sont prises sur les gens, par-devers eux, qu'une approche des données par l'émancipation est essentielle. Klein et D'Ignazio déploient un propos qui va à l'encontre de la manière dont on valorise les données aujourd'hui.

L'enjeu n'est pas leur transparence ou leur ouverture, qui profitent trop souvent à ceux-là mêmes qui bénéficient également de leur opacité, mais bien à la libération des personnes que les données calculent : libérer les gens plutôt que de seulement libérer les données !

# UN POINT DE VUE NEUTRE NE L'EST JAMAIS

L'exemple de Target souligne l'asymétrie profonde de la data science aujourd'hui : la collecte et l'analyse de données se déroulent sans avoir à rendre de compte aux personnes et groupes concernés et ce d'autant plus que ces groupes et personnes sont le moins en capacité de se défendre. En data science, les femmes et les groupes les plus marginalisés sont absents des équipes qui décident : et personne ne pense que c'est un problème !

A la suite de Donna Haraway, les chercheuses rappellent qu'un point de vue neutre ne l'est jamais. Si nous voulons parler d'objectivité en data science, alors nous devons porter attention à la perspective par défaut que défend cette prétendue objectivité. Klein et D'Ignazio défendent une co-libération plutôt qu'une responsabilité (qui consiste à libérer les gens plutôt qu'à libérer les données), la justice plus que l'éthique, l'équité plus que la loyauté, la réflexivité plus que la transparence, la compréhension de l'histoire, de la culture et du contexte plus que la compréhension des algorithmes.

# LE FÉMINISME DES DONNÉES REPOSE SUR 7 PRINCIPES QUI DIRIGE SON ACTION :

## - Examiner le pouvoir :

le féminisme des données commence par l'analyse du fonctionnement du pouvoir.

## - Défier le pouvoir :

Le data féminisme s'engage à remettre en question les structures de pouvoir inégales et à œuvrer pour la justice.

## - Valoriser l'émotion et l'incarnation.

Le féminisme des données nous apprend à valoriser de multiples formes de connaissances, y compris celles qui proviennent des personnes en tant que corps vivants et sensibles dans le monde.

## - Repenser le binarisme et les hiérarchies.

Le féminisme des données nous oblige à remettre en question le binaire du genre, ainsi que d'autres systèmes de comptage et de classification qui perpétuent l'oppression.

## - Adopter le pluralisme.

Le féminisme des données insiste sur le fait que la connaissance la plus complète provient de

la synthèse de perspectives multiples, la priorité étant donnée aux modes de connaissance locaux, autochtones et expérientiels.

## - Tenir compte du contexte.

Le féminisme des données affirme que les données ne sont pas neutres ou objectives. Elles sont le produit de relations sociales inégales, et ce contexte est essentiel pour mener une analyse précise et éthique.

## - Rendre le travail visible.

Le travail de la science des données, comme tout travail dans le monde, est le travail de nombreuses mains. Le féminisme des données rend ce travail visible afin qu'il soit reconnu et valorisé.

*Data Feminism* n'est pas seulement un propos sur les femmes ou le genre. C'est une réflexion sur l'émancipation par la science des données. La force du livre de Klein et D'Ignazio est de bâtir une réflexion politique sur les données et les calculs... de politiser la science des données sans la réduire à une injonction à l'ouverture ou à l'amélioration sans fin des données. Ce n'est pas un petit pas de côté !

Comme le constate la journaliste et activiste britannique Caroline Criado-Perez dans son livre, *Femmes invisibles* (First, 2020), les deux chercheuses pointent combien les données sont faillibles, combien elles masquent d'innombrables lacunes de connaissance. Elles reviennent par exemple sur les complications postnatales que connaissent les femmes, et notamment les femmes noires (qui, aux Etats-Unis, ont trois fois plus de risque de mourir d'une complication postnatale que les femmes blanches). Ces différences structurelles sont connues depuis longtemps des associations comme *Black Mamas*, mais moins au-delà. La presse s'y est certes intéressée, mais ni le détail des complications ni le nombre de femmes noires qui étaient concernées n'étaient détaillés. En 2018, *USA Today* montrait qu'il n'y avait pas de système national de suivi des complications subies pendant la grossesse et l'accouchement, ni de système de rapport pour s'assurer que les hôpitaux respectaient les normes de sécurité (alors que ces modalités de reporting existent pour d'autres types d'opérations). Ce que nous choisissons de mesurer montre avant tout ce à quoi nous portons de la valeur.

Ou pour le dire autrement, «*ce que nous tolérons indique ce que nous sommes vraiment*», comme le disait Caterina Fake. Or, bien souvent, les problèmes de pouvoir structurels, systémiques, sont invisibles tant que les gens qui y sont confrontés ne les mettent pas à jour. Pour les deux chercheuses, l'invisibilité est un produit de la science des données et de la «*matrice de domination*» (un concept que forgea la sociologue Patricia Hill Collins) qui s'imposent à tous et plus encore à ceux qui sont à l'intersection de genre, de race, d'identité, de capacité, d'origine... des systèmes d'oppression.

Il faut prendre conscience de l'éléphant qui se cache dans les data centers. Pas plus les données que ceux qui pratiquent cette science ne sont représentatifs. «*Les risques encourus lorsque des personnes issues de groupes dominants créent la plupart de nos produits de données – ne sont pas seulement que les ensembles de données soient biaisés ou non représentatifs, mais d'abord qu'ils ne soient jamais collectés du tout*». Comme le dénonce l'artiste Mimi Onuoha depuis son projet de Bibliothèque de jeux données manquants : qui montre un classeur rempli de dossiers de don-

nées vides. Le livre de Criado-Perez comme la fondation Data2X pointent très bien cette «*fracture systématique des données de genre*».

Pour combler ces lacunes, l'enjeu, bien sûr, consiste à développer et promouvoir un activisme des données, une forme de science des données participative pour collecter justement les données manquantes, à l'image du projet de surveillance communautaire AirBeat du quartier noir de Roxbury à Boston qui souhaitait collecter des données pour dénoncer la piètre qualité de l'air, ou de celui initié dès 1895 par la militante des droits civiques Ida B Wells sur les agressions et lynchages à l'encontre des noirs aux Etats-Unis. Les initiatives de «*contre-données*» sont nombreuses, à l'image du minutieux travail de Maria Salguero qui compile des informations sur les féminicides au Mexique depuis 2016. Ces initiatives comblent souvent des vides de données, des manques, des lacunes, des négligences statistiques... qui concernent trop souvent les corps minorisés qui ne détiennent pas le pouvoir. Pourtant, bien souvent, le problème n'est pas tant l'absence de données que l'inverse : les bases de données et les systèmes de données des institutions les plus puissantes sont construits sur et pour la surveillance excessive des groupes minoritaires, comme le montrait Virginia Eubanks dans son livre, *Automating Inequality* (St Martin Press, 2018). Ces formes de surveillance excessives reposent souvent sur deux hypothèses erronées : croire que plus de données est toujours mieux et que les données sont neutres.

Or, dans le cadre du modèle de prévision du risque de maltraitance des enfants en Pennsylvanie qu'étudiait Eubanks, la disproportion de parents pauvres dans la base de données concentre ses effets sur la pauvreté plus que sur la maltraitance.

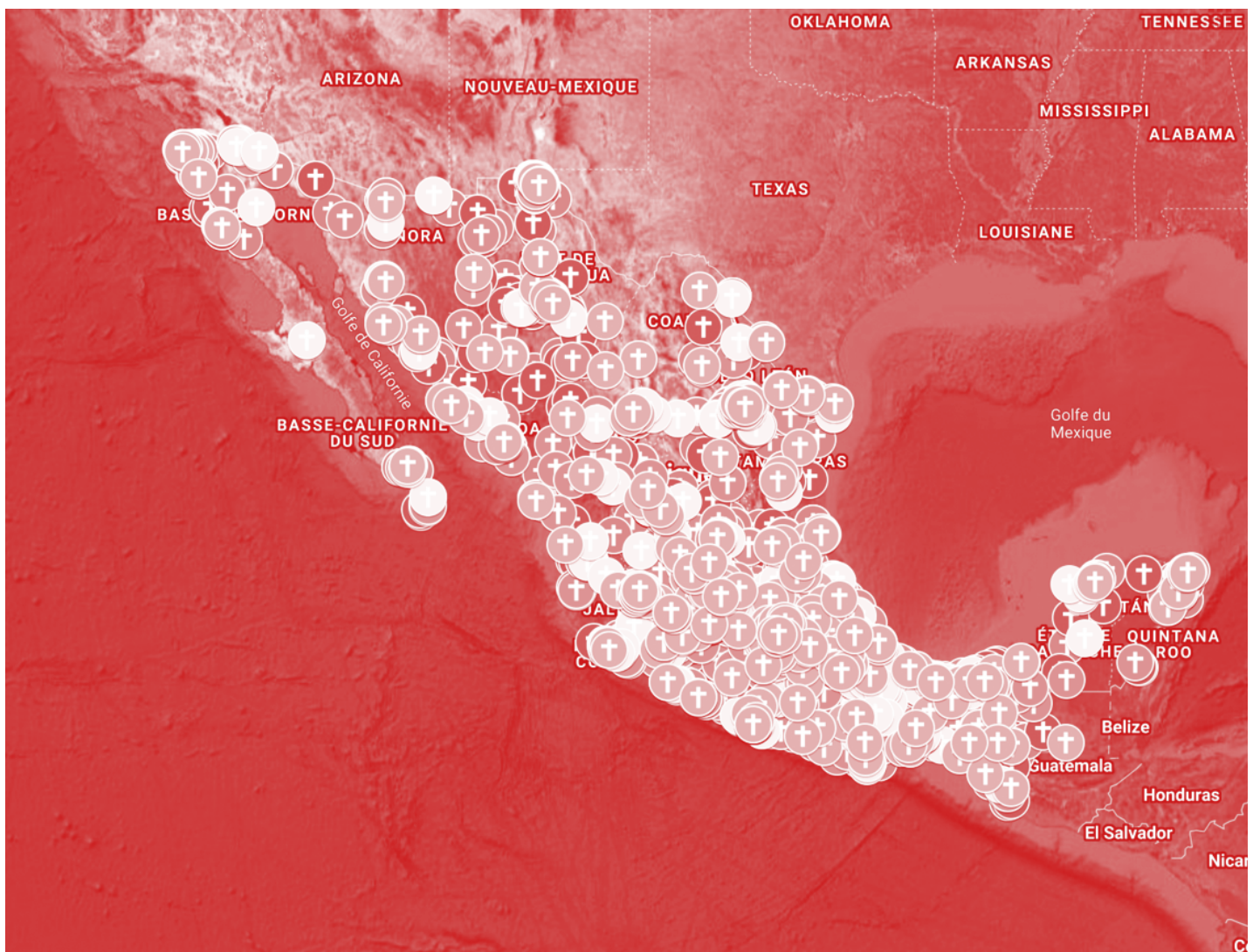
Pour Klein et D'Ignazio,  
«*les données ne sont jamais neutres ; elles sont toujours le résultat biaisé de conditions et d'inégalités sociales, historiques et économiques*».

La science des données masque souvent ses objectifs réels sous un principe d'efficacité qui ne tient pas compte de ses biais. Le système de prévision du risque de maltraitance visait avant tout à produire un mécanisme pour suppléer un personnel limité. La solution technique proposée ne visait pas tant à répondre à la maltraitance des enfants qu'à accélérer la capacité de traitement du donneur d'ordre. Comme le montrait l'exemple de Target, derrière la neutralité des déploiements techniques, les entreprises privilégient toujours leurs objectifs à ceux des personnes qu'elles sont censées servir.

Pour D'ignazio et Klein, la science des données est utilisée presque exclusivement au service du profit, de la surveillance et de l'efficacité. Cela s'explique certainement parce que les données restent coûteuses à produire et que seules les institutions puissantes peuvent les travailler. Les données servent principalement les objectifs des institutions et notamment à consolider leurs contrôles sur leurs clients. L'extractivisme de données engendre une

profonde asymétrie entre ceux qui calculent et ceux qui sont calculés. Pour renverser ce paradigme, il est nécessaire d'être attentif à l'impact des données et à qui les assemble. A l'image de l'application IRTH développée par Kimberly Seals Allers, une application à destination des mères de la communauté noire, pour à la fois documenter leurs problèmes spécifiques, les mettre en avant et leur trouver des solutions adaptées. Pour D'ignazio et Klein, comprendre et examiner qui détient le pouvoir sur les données ne sert pas seulement à comprendre, mais doit d'abord servir à contester ce pouvoir et à le changer, le renverser. Pour le dire autrement, c'est les gens et leurs corps qui doivent nous dire quelles données vont améliorer leur vie et comment. Plus encore, si la donnée ne fait qu'informer, que rendre plus efficace des formes d'extraction et de domination, alors elle rate son objectif principal : libérer celles et ceux qui sont l'objet des données et des calculs.

Carte des féminicides au Mexique depuis 2016



# CO-LIBÉRER : LIBÉRER LES GENS PLUS QUE LES DONNÉES

Le féminisme de données vise à remettre en question les structures de pouvoir. Trop souvent, « *les données sont déployées pour soutenir les intérêts des personnes et des institutions en position de pouvoir* ». Remédier à ce problème nécessite donc de compiler ses propres « *contre-données* », de cartographier l'oppression et cela ne peut se faire que depuis d'autres points de vue et donc par d'autres personnes. L'identité de ceux qui produisent les données, leur proximité avec le sujet, les conditions de la collaboration et la direction du projet sont essentielles. La science des données est essentielle pour contester le pouvoir, car l'examiner est essentiel. L'enjeu n'est pas tant l'équité que la co-libération, insistent les deux chercheuses. Le travail mené par Julia Angwin et *ProPublica* en 2016 sur les biais des systèmes d'évaluation de risque de récidive utilisés par le système judiciaire américain a montré combien les 137 questions qui président à ces évaluations encodent les inégalités structurelles de la société américaine. Ruha Benjamin dans *Race After Technology* (Polity, 2019) évoque d'ailleurs le concept de « *nouveau code Jim* » pour évoquer combien les systèmes techniques encodent le racisme de la société américaine. « *Les données sont toujours le produit de relations sociales inégales* ». Comme l'explique Ben Green dans son livre, *The Smart Enough City* (MIT Press, 2019), l'apprentissage automatique prédit plus le passé que l'avenir. Le problème est que les résultats de ces calculs sont activement promus comme objectifs ou neutres, alors qu'ils ne le sont pas. Et cela devient bien plus difficile à

mesure que les calculs sont barricadés dans des boîtes noires inscrutables, protégées derrière le droit des sociétés, et donc nécessitant de construire ses propres ensembles de données pour contourner leur inaccessibilité.

Pourtant, si l'analyse et la dénonciation de l'oppression des données peuvent obliger les institutions à rendre des comptes, l'efficacité de la production de contre-données ne suffit pas. Les données de Maria Salguero sur les féminicides au Mexique ont eu besoin d'être complétées par des commissions gouvernementales, des rapports d'ONG et des décisions de tribunaux pour produire des effets concrets. Les contre-données ne suffisent donc pas à produire du changement, d'autant plus que les données et les calculs sont toujours considérés comme moins partiels que les individus, plus objectifs par nature. Mais ce n'est le cas que parce que les perspectives de ceux qui les produisent passent pour la valeur par défaut.

**L'objectivité  
n'a rien d'immanent.**

L'éthique des données vise à produire de la responsabilité et de la transparence pour remédier à ce manque d'objectivité. Mais pour les chercheuses, les termes du débat posent problème. Il nous faut passer de concepts qui sécurisent (et entérinent le pouvoir) à des concepts qui le défient, expliquent-elles en nous invitant à modifier les concepts que nous utilisons dans les discussions sur l'utilisation

des données et des algorithmes. En passant de l'éthique à la justice, du biais à l'oppression, de la loyauté à l'équité, de la responsabilité à la co-libération, de la transparence à la réflexivité, de la compréhension des algorithmes à la compréhension de l'histoire, de la culture et du contexte... l'enjeu est de sortir de la recherche de solutions technologiques aux problèmes que pose la technologie.

Pour D'Ignazio et Klein, le tableau de changement de concepts qu'elles proposent ne signifie pas par exemple que l'éthique n'a pas sa place dans la science des données ou que la question des biais ou de la transparence ne devrait pas être abordée, mais que ces concepts ne suffisent pas pour rendre compte des causes profondes de l'oppression et limitent finalement l'éventail de réponses pour démanteler l'oppression ! C'est un moyen de remettre de la politique là où certains tentent de l'ôter pour mieux conserver le pouvoir.

Or, la capacité à dépolitiser les questions n'est que le privilège de ceux qui ont déjà le pouvoir. D'où la nécessité de parler de justice plutôt que d'éthique, d'équité plus d'égalité comme nous y invitait déjà Sasha Costanza-Chock, dans son livre *Design Justice* (MIT Press 2020). Le biais par exemple n'est pas un concept suffisamment fort pour ancrer les idées d'équité et de justice. En effet, pour l'éviter, on a tendance à vouloir enlever les humains de la boucle de décision, au motif que leurs préjugés personnels seraient la principale raison de la continuation des discriminations. Mais ce raisonnement suppose que le racisme par exemple serait plus le fait de mauvais acteurs individuels que de structures ou de systèmes. Alors qu'en fait, remplacer les travailleurs sociaux – qui sont souvent des femmes voire des femmes de couleur, qui font preuve d'empathie et de capacités d'écoute – par des systèmes automatisés qui appliquent un ensemble de critères rigides quels que soient les circonstances, ne nous prémunit pas des préjugés structurels, au contraire !

« Partir du principe que c'est l'oppression qui est le problème, et non les préjugés, conduit à

des décisions fondamentalement différentes sur ce sur quoi il faut travailler, avec qui il faut travailler, et quand il faut se lever et dire qu'un problème ne peut et ne doit pas être résolu par les données et la technologie. »

L'enjeu donc n'est pas tant d'ouvrir ou libérer les données, que de libérer les personnes ! La clé de cette co-libération nécessite un engagement et la croyance dans un bénéfice mutuel, tant pour les groupes dominants que pour les groupes minorisés. La libération des uns et des autres est liée. C'est ce qui motive par exemple le projet *Our Data Bodies* et son *Digital Defense Paybook*. La co-libération nécessite également de nouveaux outils et un autre état d'esprit.

« Partir du principe que l'oppression est le problème, que l'équité est la voie à suivre et que la co-libération est l'objectif souhaité conduit à des projets fondamentalement différents qui remettent en cause le pouvoir à sa source. »

Cela nécessite aussi de mesurer différemment la réussite d'un projet. La réussite ne repose plus alors sur l'efficacité d'une base de données, la précision d'un algorithme ou la taille de la base de données, mais plutôt sur le degré de confiance établi, le partage du pouvoir et des ressources, l'ampleur de l'apprentissage de ceux qui participent, la transformation des personnes et des organisations par le processus, la quantité d'inspiration nouvelle engendrée ! Des objectifs qui sont certes plus flous à mesurer, mais qui peuvent l'être tout de même !

Lorsque Gwendolyn Warren et les chercheurs du Detroit Geographic Expedition and Institute dans les années 70 ont recueilli des données sur les accidents de voiture impliquant les enfants noirs à Détroit, ils l'ont fait pour que les chercheurs rendent à la communauté les connaissances qu'ils exploitaient et permettent à la communauté de créer une stratégie pour combattre les discriminations qu'elle subissait. Pour Warren,

l'éducation était un mécanisme d'autonomisation et de transformation.

Pour y parvenir, il faut encore que les femmes comme les personnes de couleurs soient plus nombreuses dans les salles de classe de la science des données, alors que 80 % des professeurs en IA et que 66 % des professeurs d'informatique et de statistiques sont des hommes. Cela nécessite aussi de s'extraire de la croyance que tout cela ne serait qu'une question technique. Pour Laurie Rubel, à l'origine du projet Local Lotto, l'enjeu est d'enseigner autrement les mathématiques en tenant compte de l'équité. Laurie Rubel a proposé à des élèves du secondaire de répondre à la question : est-ce que la loterie est bonne pour votre quartier ? Les billets de loterie sont une production sociale. Ils sont bien plus achetés par les travailleurs à bas salaires que par les plus riches. Lors de ces exercices, les élèves devaient parcourir leur quartier pour dresser des cartographies, rencontrer des acheteurs... comprendre les enjeux. Ce travail leur a permis surtout de rétroagir sur leurs communautés, pour leur montrer combien la loterie leur était défavorable. Local Lotto utilise la science des données pour construire une approche concrète et sociale, plutôt qu'abstraite et technique et aider à permettre aux acteurs d'en prendre conscience pour y remédier. Le projet vise à adresser des problématiques d'inégalités de la vie quotidienne. Il valorise l'expérience vécue et le partage. « *La justice est un voyage* » qui nécessite de travailler aux problèmes que l'injustice cause.

**Examiner le pouvoir ne suffit pas, rappellent les deux chercheuses. Il faut aussi le remettre en question. Pour cela, il faut prendre l'oppression et l'inégalité comme une hypothèse de base pour créer des systèmes. Cela nécessite d'apprendre des communautés et de concevoir les systèmes avec elles.**

**The next few questions are about the family or caretakers that mainly raised you when growing up.**

31. Which of the following best describes who principally raised you?
- Both Natural Parents
  - Natural Mother Only
  - Natural Father Only
  - Relative(s)
  - Adoptive Parent(s)
  - Foster Parent(s)
  - Other arrangement
32. If you lived with both parents and they later separated, how old were you at the time?
- Less than 5  5 to 10  11 to 14  15 or older  Does Not Apply
33. Was your father (or father figure who principally raised you) ever arrested, that you know of?
- No  Yes
34. Was your mother (or mother figure who principally raised you) ever arrested, that you know of?
- No  Yes
35. Were your brothers or sisters ever arrested, that you know of?
- No  Yes
36. Was your wife/husband/partner ever arrested, that you know of?
- No  Yes
37. Did a parent or parent figure who raised you ever have a drug or alcohol problem?
- No  Yes
38. Was one of your parents (or parent figure who raised you) ever sent to jail or prison?
- No  Yes

# DATA FEMINISM READING GROUP

## DATA FEMINISM READING GROUP 01

### CATHERINE D'IGNAZIO

THE CURRENT UNDERSTANDING OF VISUALIZATION AND INTERACTIVE DESIGN IS A MALE PERSPECTIVE. DESIGNING VISUALIZATION AND INTERACTIVE DESIGN IS A MALE PERSPECTIVE.

### LAUREN F. KLEIN

THE IDEA OF VISUALIZING DATA IS A MALE PERSPECTIVE.

**INTERSECTING FORCES OF OPPRESSION**

**MATRIX OF DOMINANCE**

**WHY VISUALIZATION?**

**THE TRAVEL HAZARD**

**THE WORLD HAS STOPPED FOR PEOPLE LIKE ME**

**LAUREN F. KLEIN**

**LAUREN F. KLEIN**

**LAUREN F. KLEIN**

## DATA FEMINISM READING GROUP 02

### CATHERINE D'IGNAZIO

WHAT IS WHITENESS?

**ETHICS**

**DATA JUSTICE**

**LAUREN F. KLEIN**

**LAUREN F. KLEIN**

**LAUREN F. KLEIN**

## DATA FEMINISM READING GROUP 03

### CATHERINE D'IGNAZIO

IF WE DO VISUAL MINIMALISM IS IT REALLY NEUTRAL?

**ELEVATE EMOTION**

**LAUREN F. KLEIN**

**LAUREN F. KLEIN**

**LAUREN F. KLEIN**

## DATA FEMINISM READING GROUP 04

### CATHERINE D'IGNAZIO

WHO CONTROLS THE FLUX OF DATA COLLECTION AND ANALYSIS?

**RETHINK BINARIES AND HIERARCHIES**

**LAUREN F. KLEIN**

**LAUREN F. KLEIN**

**LAUREN F. KLEIN**

## DATA FEMINISM READING GROUP 05

### CATHERINE D'IGNAZIO

PROCESS REALLY MATTERS

**EMBRACE PLURALISM**

**LAUREN F. KLEIN**

**LAUREN F. KLEIN**

**LAUREN F. KLEIN**

## DATA FEMINISM READING GROUP 06

### CATHERINE D'IGNAZIO

BIG DATA

**CONSIDER CONTEXT**

**LAUREN F. KLEIN**

**LAUREN F. KLEIN**

**LAUREN F. KLEIN**

## DATA FEMINISM READING GROUP 07

### CATHERINE D'IGNAZIO

CONTEMPORARY TECHNOLOGY IS EXPLOITATION OF PEOPLE

**MAKE LABOR VISIBLE**

**LAUREN F. KLEIN**

**LAUREN F. KLEIN**

**LAUREN F. KLEIN**

DATA FEMINISM REQUIRES EXPANDING THE DEFINITION OF WHAT COUNTS AS DATA SCIENCE SO THAT IT INCLUDES MORE PEOPLE

IS ABOUT EQUITABLE WAYS OF USING DATA SCIENCE FOR SOCIAL JUSTICE!

AVOID THE IDEA THAT TECHNOLOGICAL SOLUTIONS ARE SUPERIOR. ASK INSTEAD: WHAT IS THE RIGHT TOOL FOR THE TASK?

WE NEED TO FIGHT AGAINST CERCERAL TECHNOLOGIES

WHAT DOES RESTITUTIVE JUSTICE LOOK LIKE FOR BLACK DATA?

THE INCLUSION OF UNDERSERVED AND MARGINALIZED COMMUNITIES PROMOTES DIVERSITY IN PRACTICE.

NAME THE WAYS WE HAVE BEEN ROBBED OF POSSIBILITIES BY THE COMPANIES THAT ONLY WANT PROFIT!

REPARATIONS FROM THE TECH INDUSTRY TO THE WORLD!

WE ARE CARRYING OUR WORK ON TERMS THAT ARE NOT OUR OWN MAKING

DOING THIS WORK ALSO MEANS REINFORCING THAT OPPRESSION ONLY COUNTS IF IT CAN BE COMPRESSED INTO DATA!

**CATHERINE D'IGNAZIO**

**LAUREN F. KLEIN**

**Meredith Brunsford**

**Sajha Constanza-Cook**

**De'Faithé James Day**

**Faizha Tomacha**

**De' Sajaya Hobbs**

**Mimi Onocha**

**Margaret W. Pearce**

**Tawana Petty**

# DE L'ÉTHIQUE DES DONNÉES À LA JUSTICE DES DONNÉES

## *Concepts qui sécurisent le pouvoir*

Parce qu'ils localisent la source des problèmes dans les individus ou les systèmes techniques

Ethique  
Biais  
Conformité  
Responsabilité  
Comprendre les algorithmes

## *Concepts qui défient le pouvoir*

Parce qu'ils reconnaissent les différentiels de pouvoir structurels et s'efforcent de les démanteler

Justice  
Oppression  
Équité  
Co-libération  
Comprendre l'histoire, la culture le contexte

Tableau des changements de concepts que propose le féminisme des données, alternatives aux concepts phares de l'éthique de la science des données, proposé par Klein et D'Ignazio.

« Moins de 5 % des artistes des sections art moderne du musée sont des femmes, mais 85 % des nus sont des femmes », infographie des Guerilla Girls, 1989.»



# PRIVILÉGIÉ L'INCARNATION ET L'ÉMOTION À LA FAUSSE NEUTRALITÉ

Contrairement à ce que l'on pense trop souvent, l'objectivité des données nécessite non pas de promouvoir une forme de neutralité immanente, mais au contraire de mieux les incarner, c'est-à-dire de mieux montrer et valoriser leur subjectivité intrinsèque. C'est ce que propose par exemple *Periscopic* dans sa formidable infographie sur le nombre de personnes tuées par arme à feu aux Etats-Unis. Contrairement à une présentation traditionnelle du phénomène, l'infographie de *Periscopic* valorise l'émotion en permettant d'accéder aux informations individuelles sur chaque personne tuée par arme à feu aux Etats-Unis et en totalisant le nombre moyen d'années volées à toutes ces vies !

Cette approche a généré certaines critiques. Alberto Cairo, auteur d'un livre sur la visualisation de données, *The Truthful Art* (New Riders, 2016, non traduit), estime que l'infographie de *Periscopic* instrumentalise une cause par l'émotion, alors que la visualisation doit rester neutre ! Dans *A Unified Theory of Information Design* (Routledge, 2013, non traduit), Nicole Amare et Alan Manning, expliquent également que la visualisation doit rester neutre. La sobriété est nécessaire pour que les lecteurs interprètent les résultats par eux-mêmes. Dans le domaine de la communication de données, l'ornement est souvent considéré comme suspect.

Comme le dit l'historien des sciences Theodore Porter, « *la quantification est une technologie de la distance* », c'est-à-dire qui met à distance sous principe d'objectivité. La neutralité

en est l'idéal. C'est oublier que la persuasion est partout, même dans le dépouillement de simples feuilles de calcul ou des plus simples graphiques. Haraway a pourtant été la première à souligner le lien entre l'apparente neutralité et l'objectivité des données. Pour elle, la visualisation de données encourage trop souvent à voir les phénomènes depuis nulle part – voire de loin ou d'en haut. Ce point de vue, cette « *ruse* », présente comme neutre une perspective partielle, qui est là encore, bien souvent, la perspective du groupe dominant. Pour le grand spécialiste de statistique et de visualisation de l'information, Edward Tufte, le concepteur de visualisation doit s'efforcer d'utiliser l'encre uniquement pour afficher les données... tout embellissement est suspect. Le minimalisme visuel est rationnel. Les éléments décoratifs sont tous suspects d'œuvrer à des formes de persuasion émotionnelle.

La recherche féministe a beaucoup critiqué cette opposition entre raison et émotion, qui vise plus à structurer et imposer des hiérarchies implicites qu'autre chose.

Pour D'Ignazio et Klein, le minimalisme visuel n'est pas si neutre qu'il le clame. Trop de théoriciens et de praticiens de la visualisation de données relèvent des disciplines techniques de l'ingénierie et de l'informatique, mais oublient que tout objet, aussi neutre soit-il présenté, est rhétorique. Dans un visuel, un même constat peut être présenté de manière très différente selon sa représentation ou le discours qui

l'accompagne, à l'image d'un même graphique factuel sur le chômage proposé par le *New York Times*, présentant les données selon un point de vue démocrate ou républicain. En fait, soulignent les chercheuses, la visualisation de données dépend toujours d'une interprétation. Les choix éditoriaux qui produisent une visualisation ont des effets de cadrage qui obscurcissent ou mettent en évidence certaines choses plutôt que d'autres. Les conventions autour du minimalisme favorisent une perception de visualisation comme étant neutre et objective. Au final, même les visualisations les plus simples, les plus factuelles, ne sont pas neutres. Par contre, elles sont très persuasives.

Dans une approche féministe des données, expurger toutes traces humaines pour produire une neutralité de façade n'est pas la bonne approche. L'enjeu, au contraire, consiste à affirmer la nature située des connaissances, à affirmer leur partialité. Pour D'Ignazio et Klein, il est nécessaire de rendre les données plus « *viscérales* » plutôt que neutres. L'émotion n'est pas aussi irrationnelle et illégitime que beaucoup la tiennent. Notre compréhension du monde est une combinaison d'expériences, d'intuitions, de raisons et d'émotions. En 2010, Kelly Dobson a fondé le groupe de recherche *Data Visceralization*, avec l'objectif de produire des données à ressentir, à expérimenter avec le corps, tant physiquement qu'émotionnellement. Cette viscéralisation n'était pas qu'une expérimentation créative. Pour ceux qui ont animé ce groupe de travail, l'enjeu était de rappeler que nous ne sommes pas des yeux attachés à un cerveau. La viscéralisation des données poursuit également un but d'accessibilité : les données ne peuvent être seulement visuelles. Klein et D'Ignazio évoquent ainsi le travail du Bureau de recherche créative pour le MoMA de New York qui s'est joué du catalogue des 123 951 oeuvres de la collection en faisant lire le nom des artistes masculins par des hommes et ceux des artistes féminins par des femmes pour souligner la nature hautement sexuée des collections... ou encore le travail du collectif *Guerrilla Girls* qui soulignait que pour entrer au *Metropolitan Museum of Art*, il fallait assurément être une femme nue arguant que les femmes nues sont le principal sujet des tableaux alors que les artistes femmes, elles, sont inexistantes dans les collections. Faire des données des expériences permet de casser la vue d'ensemble que les

visualisations proposent et permettent de faire l'expérience des enjeux peu à peu. L'enjeu est ici de ressentir la différence de genre plutôt que de seulement la voir, comme s'ils s'adressaient à une autre partie du cerveau.

Un autre enjeu également dans les représentations visuelles, consiste à mieux faire figurer l'incertitude, comme d'arriver à montrer la fourchette d'incertitude de sondages. Lors de l'élection de Trump en 2016, les simulations produites par les sondages étaient plus incertaines dans les intentions de vote que leurs représentations. Pour Klein et D'Ignazio, le fait que nombre de graphiques aient suggéré que Clinton allait gagner tient beaucoup au refus de montrer l'incertitude que les sondages exprimaient. Ajoutez-y les biais de nos propres raccourcis mentaux qui tendent à nous faire croire que l'avance d'une candidate sur l'autre tient d'une probabilité plus forte qu'elle gagne... Et vous comprenez mieux le risque à ne pas s'ouvrir à l'incertitude des données. Pour les deux chercheuses, ces exemples doivent nous rappeler que le contexte est roi en matière de visualisation.

Infographie de Perisopic sur le nombre de personnes tuées par arme à feu aux Etats-Unis afin de rendre le message plus percutant en montrant le nombre d'années de vies volées.

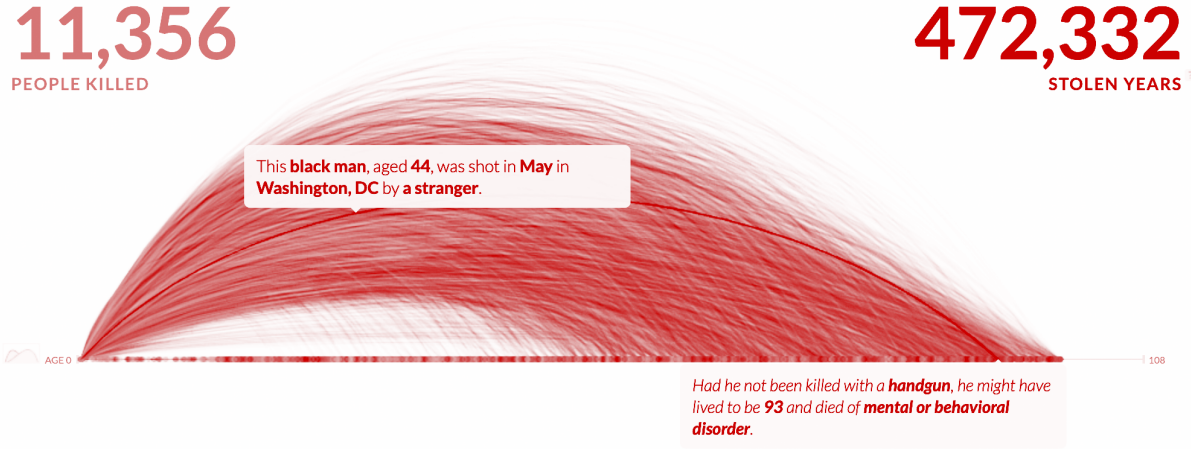
## U.S. GUN KILLINGS IN 2018

# 11,356

PEOPLE KILLED

# 472,332

STOLEN YEARS



This **black man**, aged **44**, was shot in **May** in **Washington, DC** by a **stranger**.

Had he not been killed with a **handgun**, he might have lived to be **93** and died of **mental or behavioral disorder**.

# INTERROGER LES CLASSEMENTS ET LES CATÉGORIES

Le féminisme des données oblige à sortir des dichotomies et donc à repenser les catégories et hiérarchies. De nombreux problèmes d'inégalité structurelle sont des problèmes d'échelles. Trop souvent, seul ce qui est compté compte, et ce qui n'est pas compté ne compte pas, rappellent les chercheuses. Ici, il est nécessaire également de nous défier de l'affichage de grands principes. Ainsi, quand Facebook a élargi les modalités pour indiquer son genre, cette nouveauté a été saluée comme le signe d'une grande ouverture. Pourtant, comme l'ont montré les travaux de Rena Bivens, Facebook a continué de résoudre le genre des utilisateurs de manière binaire pour ses clients payants. Mêmes constats en ce qui concerne la politique des « vrais noms » de Facebook qui est justifiée pour des raisons de sécurité, mais qui met activement en danger les utilisateurs les plus marginaux de la plateforme, comme l'ont montré Oliver Haimson et Anne Lauren Hoffmann, par exemple ceux qui portent des noms Amérindiens où les personnes qui cherchent à masquer leurs identités parce qu'elle peut leur porter préjudice.

Les féministes, comme les militants antiracistes et homosexuels, ont beaucoup réfléchi aux problèmes de classification, notamment parce que cela les affectait directement.

La solution pourtant ne consiste pas à refuser le classement, mais à nous interroger sur la manière dont les systèmes de classification sont construits, sur la manière dont ils encodent les valeurs et jugements, sur comment ils sont construits et dans quel but, comme l'expliquent Geoffrey Bowker et Susan Leigh Star dans *Sorting Things Out* (MIT Press, 2000, *Arranger les choses*, EHESS, 2023). Le problème consiste toujours à savoir s'il vaut mieux être compté ou pas. C'est le paradoxe de l'exposition ou de la classification. Pour un sans papier par exemple, être dénombré peut être problématique selon ce qui est fait de ce dénombrement. Du point de vue du féminisme de données, la question est toujours de savoir si ce sont les catégories qui sont inadaptées ou le système de classification.

A quoi sert-il par exemple que les scanners corporels des aéroports distinguent les hommes des femmes et s'affolent dès qu'une personne ne correspond pas à cette normativité réductrice ?

L'Association médicale américaine qualifie désormais le genre de spectre, plutôt que sur un mode binaire. L'organisme de santé publique britannique a lui développé un questionnaire inclusif et sensible qui permet d'ouvrir la question des catégories de genre en proposant des

modalités de non divulgation. En fait, soulignent les chercheuses, selon les circonstances, l'institution qui collecte, et le but de ces collectes, les décisions éthiques peuvent varier. Pour le dire plus simplement, l'éthique ne peut être bornée : chaque comptage nécessite de reposer les questions de contexte et de pouvoir : faut-il compter selon le genre ? Quand et comment ? Même constat du côté de la couleur de peau... Mais la réponse à ces questions n'est possible qu'en répondant à la question à qui et à quoi servent ces comptages et plus encore en assurant que leur utilité reste circonscrite à l'objectif formulé – ce qui est plus difficile à assurer, notamment dans le monde des données, où un comptage, par nature, peut-être réutilisé à d'autres fins. Enfin, soulignent les chercheuses, la question du consentement à ces différents types de comptages est essentielle, pour les groupes minoritaires et notamment du fait des préjudices potentiels qu'il y a à distinguer les gens selon leurs identités fait peser sur eux. Respecter les gens qui sont derrière les données peut sembler tenir de l'évidence. C'est pourtant si rarement le cas qu'il faut bien le rappeler (signalons d'ailleurs sur ce sujet, le travail stimulant du *Manifeste pour des technologies pleinement consenties*, qui propose des principes pour des technologies de « *plein consentement* » : sans pression, réversibles, informées, enthousiastes et spécifiques) !

Historiquement, compter a toujours été utilisé pour dominer, discipliner et exclure. C'est pour cela qu'il est nécessaire d'interroger notre infrastructure de classification. Les chiffres ne sont pas toujours des outils d'oppression, mais ils doivent être travaillés pour rééquilibrer la répartition inégale du pouvoir.

# NOTRE BESOIN DE DIVERSITÉ EST IMPOSSIBLE À RASSASIER

Pour le féminisme de données, la connaissance la plus complète provient de la synthèse de perspectives multiples, et notamment les plus locales, les plus expérientielles... Sur le site Anti Eviction Map qui lutte contre les expulsions à San Francisco on trouve une centaine de cartes différentes. Beaucoup ont été produites en collaboration avec des organisations différentes afin de documenter le phénomène de l'expulsion dans sa plus complète diversité. L'une des cartes par exemple, montre que la concentration des expulsions recoupe celle des arrêts des Bus qui viennent chercher les employés des grandes entreprises de la tech. Toutes les cartes ne sont pas aussi lisibles, mais l'objectif reste de documenter ce qu'il se passe de multiples façons. Plutôt que de raconter une seule histoire « vraie » ou consensuelle.

**L'enjeu est aussi de reconnaître les versions multiples voire contradictoires de la réalité...**

... explique l'architecte Katie Lloyd Thomas, fondatrice du collectif féministe *Taking Place*. Pour le féminisme de données, l'enjeu est d'adopter le pluralisme, de valoriser les voix, les points de vue, de la collecte à la communication...

Pour se faire, il faut se défaire de la façon dont les méthodes de la science des données ont tendance à supprimer les voix, au profit de la clarté, de la propriété ou du contrôle.

Trop souvent, l'un des principaux aspects du travail du data scientist consiste à nettoyer les données pour mieux les structurer, un travail d'entretien à la fois patient et minutieux, et également un travail d'organisation glorifiant visant à dompter le chaos des chiffres. Le problème, c'est que ces processus contribuent à faire disparaître des perspectives et à en imposer d'autres. Nettoyer les données consiste aussi à les contrôler. « *Cette croyance selon laquelle les données doivent toujours être propres et contrôlées a des racines historiques douteuses* ». Trop souvent, le nettoyage est « *une astuce pour réduire la diversité* », comme l'ont expliqué Katie Rawson et Trevor Munoz dans leur article « *Contre le nettoyage* ». Le désordre des données est riche d'informations sur les circonstances de leur collecte. Trop souvent, leur nettoyage permet qu'elles soient plus accessibles à d'autres qu'à ceux qui les ont collectés. En fait, expliquent Klein et D'Ignazio, ce nettoyage des données élargit surtout leur

disponibilité et notamment leur disponibilité à un plus large niveau, les coupant du contexte de leur production. Ce nettoyage bénéficie également aux data scientists, à ces génies solitaires, souvent blancs et masculins, dont la maîtrise individuelle et l'expertise technique semblent sans bornes. L'un de ces génies est Matthew Desmond, directeur du Laboratoire de l'expulsion à l'université de Princeton et auteur d'*Avis d'expulsion, enquête sur l'exploitation de la pauvreté urbaine* (Lux éditeur, 2019). La difficulté pour le laboratoire a longtemps consisté à trouver sur quelles données baser ses recherches. Faut-il privilégier des données plus propres et à grandes échelles, faciles à traiter et à acheter où des données plus précises, à l'échelle locale, comme celles produites par Anti Eviction Map mais qui nécessitent d'établir une relation de confiance afin qu'elles ne soient pas utilisées n'importe comment ? Comme nombre d'acteurs dans le domaine des données, le laboratoire a donné la priorité à la rapidité sur la précision. A l'inverse, si l'Anti Eviction Map a privilégié la diversité, leur décision a aussi contribué, de manière tout à fait intentionnelle, à renforcer les capacités techniques et relationnelles des acteurs locaux. L'enjeu était bien de construire une solidarité entre participants afin qu'ils s'entraident dans la lutte contre les expulsions, qu'ils apprennent les uns des autres. La multiplicité permet d'obtenir une image plus complète d'un problème et en renforce les collectifs. Encore faut-il parvenir à reconnaître la partialité de sa compréhension. Cela suppose, bien souvent, de donner accès aux données utilisées et de décrire les méthodes employées. Mais également de compléter ces informations en précisant qui a travaillé sur ces données, quels étaient les points de tensions...

« *La prise en compte de la valeur des perspectives multiples ne doit pas se limiter à la transparence et à la réflexivité. Il s'agit également d'inviter activement et délibérément d'autres points de vue dans le processus d'analyse et de narration des données – plus précisément, ceux des personnes les plus marginalisées dans un contexte donné.* » Pour l'Anti Eviction Map cela signifie centrer ses données sur les voix et les expériences de ceux qui ont été expulsés, à l'image de la proposition du réseau *Design Justice* qui souhaite centrer son travail sur ceux les plus touchés par les données. Pour Klein et D'Ignazio, l'enjeu est bien ici de concevoir des projets de co-libéra-

tion plutôt que des « *données pour le bien* » (ce qu'on appelle le mouvement « *Data for good* », un réseau de réseaux international qui promeut l'usage de la donnée pour l'intérêt général). Si le mouvement *Data for good* décrit des projets de science des données socialement engagés, que signifie faire le bien ? De quel bien parlons-nous ? Au profit de qui ? Pour la spécialiste en machine learning chez Google Brain et fondatrice de Delta Analytics, Sara Hooker, ce concept manque de précision, notamment parce que l'engagement ne suffit pas à qualifier l'objectif – l'ingénieur Ben Green faisait une critique assez proche, soulignant les limites à renforcer le bien social sans interroger l'engagement. A l'inverse, la co-libération repose sur la dénonciation et la résolution des relations de pouvoir asymétriques. La co-libération présuppose une lutte qui qualifie l'enjeu... et augmente l'engagement d'objectifs, que ce soit le transfert de connaissance ou la création d'une infrastructure sociale dédiée, que ce soit pour la prise de conscience ou pour armer la lutte sociale. Elle implique des échanges à double sens pour renforcer les capacités techniques de la communauté et un renforcement de la solidarité, par exemple en allouant des ressources à l'infrastructure communautaire et pas seulement une aide technique.

« **Dans le modèle de co-libération, les projets de science des données deviennent des projets de science communautaire. Ils se déroulent simultanément dans la base de données et dans l'espace public. »** L'enjeu est de construire une compréhension partagée autour d'une question et de mobiliser autour de cette compréhension pour en décliner des actions. Les données sont un moyen de générer une prise de conscience et de l'activer.

En cela, la co-libération devient une « *technologie de rassemblement* », où « *les informations sont échangées, la cohésion sociale est renforcée et les actions futures sont co-conspirées* », une

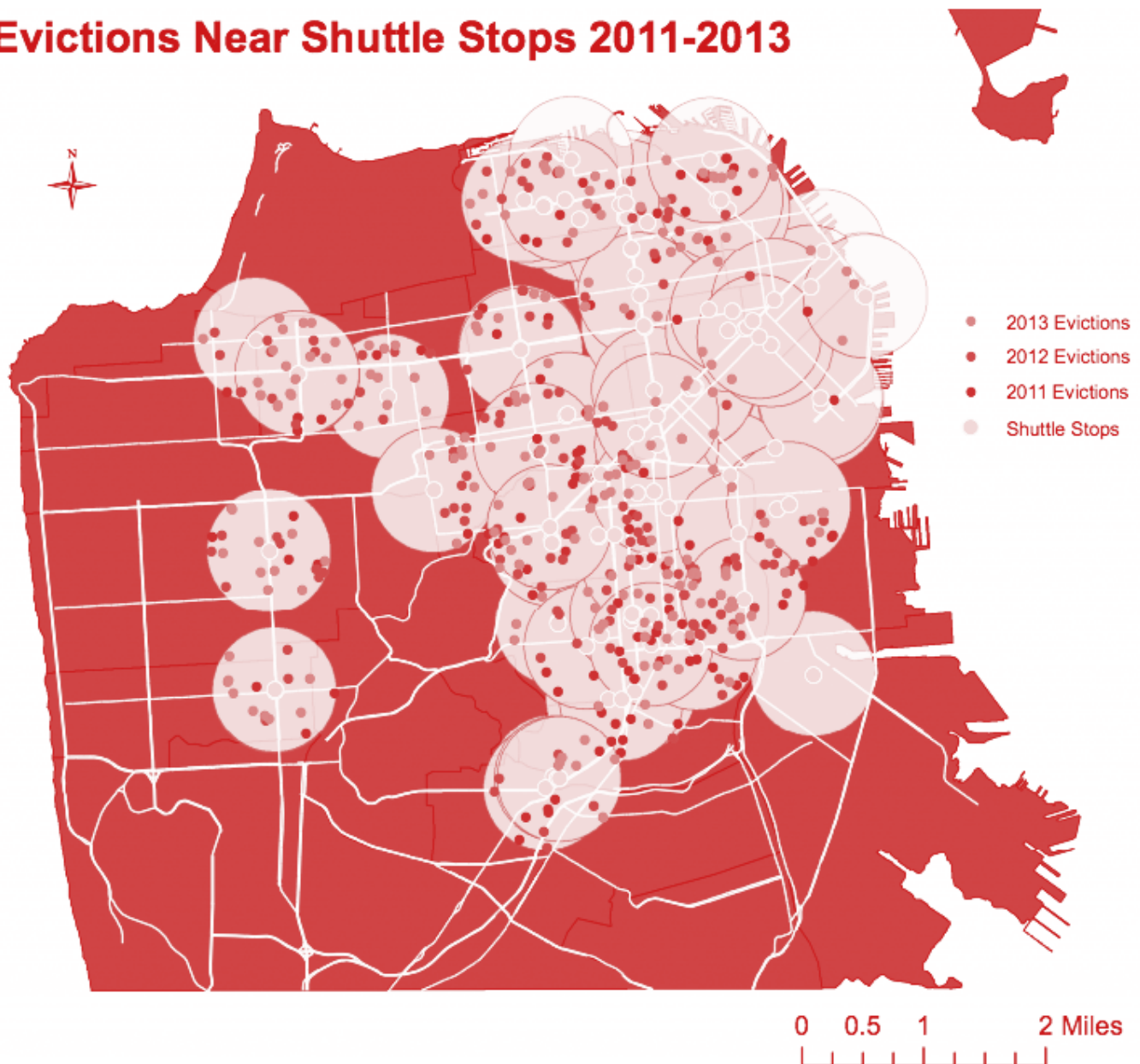
forme de « *feu de camp* » pour réactiver la participation politique, expliquent Kein et D'Ignazio – une forme de nouvelle cartographie d'alliances.

L'enjeu n'est pas pourtant de rester petit ou local. Les données peuvent aussi permettre d'atteindre d'autres échelles de co-libération, expliquent les chercheuses en évoquant le projet *d'Atlas mondial de la justice environnementale*, initié par Leah Temper et Joan Martinez-Alier qui collecte des informations sur les conflits écologiques depuis les luttes mêmes de ces acteurs. Certes, l'approche est plus pluraliste, participa-

tive et itérative que les approches extractivistes et quantitatives du Big data traditionnel, mais en fin de compte, même si cela prend plus de temps, les données, les relations et les capacités des communautés produites sont d'une bien meilleure qualité.

La carte des expulsions à San Francisco de 2011 à 2013 les comparant à l'emplacement des arrêts des navettes qui conduisent vers les grands sièges sociaux des entreprises de la tech.

## Evictions Near Shuttle Stops 2011-2013



# LES CHIFFRES NE PARLENT PAS D'EUX-MÊMES

Le féminisme des données affirme que les données ne sont ni neutres ni objectives. Elles sont le produit des inégalités et, sans contexte, elles ne disent rien : ne sont ni précises ni éthiques.

En 2014, le célèbre blog FiveThirtyEight a publié un article montrant que l'enlèvement de jeunes femmes au Nigeria était un problème qui empirait. Mais l'article a dû rapidement être rétracté. Le site avait utilisé une base de données mondiale spécialisée sur les conflits, le GDELT. Le problème est que cette base collecte des articles de presse en démultipliant les faits depuis les sources. FiveThirtyEight avait compté chaque article comme un événement alors que beaucoup d'articles de presse relayés décrivaient le même événement. L'histoire raconte la difficulté à agréger des données dont on ne sait pas grand-chose, à notre tendance à surestimer l'exhaustivité sur l'exactitude.

Certains chercheurs parlent de « Big Dick Data » pour qualifier des projets qui fétichisent la taille, exagèrent leurs capacités scientifiques ou techniques et ignorent le contexte.

Nous sommes de plus en plus cernés par des portails de données qui permettent de télé-

charger ou d'échanger de vastes ensembles de données, dans un modèle d'accès illimité à l'information que nous considérons trop souvent comme un bien inhérent.

L'un des principes centraux de la pensée féministe est que toute connaissance est située, nécessitant de relier les données au contexte de leur production pour mieux comprendre leurs limites fonctionnelles, les obligations éthiques associées et la manière dont le pouvoir a contribué à leur élaboration.

Trop souvent, pourtant, rappellent les chercheuses, les jeux de données sont fournis sans contexte ni métadonnées, c'est-à-dire sans informations sur leur provenance et sans dictionnaires pour les décrire. Klein et D'Ignazio donnent l'exemple simple et courant d'un jeu de données sur les dépenses de la ville de Sao Paulo au Brésil. Le jeu de données n'est pas difficile à comprendre en tant que tel, mais le processus qui l'a produit est bien moins clair. Comment la ville gère-t-elle ses appels d'offres ? Comment décide-t-elle de leur attribution ? Les offres publiées correspondent-elles à toutes les offres ou seulement à celles qui ont passé un contrat avec la ville ? Qu'est-ce qui explique les variations de numérotation ?... Or sans réponses, il semble difficile d'explorer, analyser

ou traiter ces données. Comme l'explique Christine Borgman dans son livre *Big Data, Little Data, No Data* (MIT Press, 2015, non traduit) : une infrastructure de connaissances est « *une écologie de personnes, de pratiques, de technologies, d'institutions, d'objets matériels et de relations* ».

Klein et D'Ignazio rappellent rapidement l'histoire du mouvement pour l'ouverture des données publiques, né au milieu des années 2000 d'une conjonction d'acteurs (gouvernements, associations, individus et militants) pour favoriser l'accès aux données, ces nouveaux types de documents publics. Sous prétexte de promouvoir le développement économique par la création de services les utilisant, d'accélérer le partage de connaissance et la science, et d'améliorer la transparence démocratique, nombre d'États, de villes et régions ont peu à peu ouvert des portails de données. Mais, comme elles le rappellent encore, le financement limité a conduit à prioriser l'ouverture de certaines données sur d'autres et surtout a rarement permis de documenter les jeux de données. Pour le spécialiste de l'ouverture des données britannique, Tim Davies, on s'est parfois retrouvé face à des « *décharges de données brutes* », incapables d'assurer l'engagement ou la responsabilité. Nombre de jeux de données demeurent peu utilisés, attendant que « *les utilisateurs entreprennent le travail intensif de déchiffrer les arcanes bureaucratiques qui obscurcissent leur signification* ». On parle aussi de données zombies pour désigner des ensembles de données publiés sans objectifs ou cas d'utilisation. Qu'importe ! Finalement, comme l'affirmait Chris Anderson de *Wired* en 2008, toute théorie devenait inutile, les données parleraient d'elles-mêmes ! Mais la corrélation sans contexte est clairement insuffisante, comme le montrait Safya Umoja Noble dans son livre *Algorithms of Oppression* (NYU Press, 2018, non traduit), en démontrant que les résultats de Google Search ne sont pas seulement racistes et sexistes à l'image de nos sociétés, mais entretiennent le sexisme et le racisme en renvoyant vers ces contenus sous prétexte de les classer en fonction des liens et de leur audience, encourageant finalement leur audience dans une boucle de circulation perpétuelle.

Depuis 1990, suite au meurtre de Jeanne Clery, une loi visant à améliorer la collecte de données relatives aux crimes commis sur les campus

américains a été votée, imposant de mettre à la disposition du public ces statistiques. En 2016, trois étudiants en journalisme de données ont téléchargé ces données collectées par le Clery Center pour comprendre ce qu'il en était de la culture du viol sur les campus. Ils se sont rendu compte qu'un petit établissement rural du Massachusetts semblait connaître bien plus d'agressions sexuelles que la prestigieuse université de Boston. En enquêtant, ils se sont rendu compte que certains établissements publiaient leurs chiffres et que d'autres avaient les ressources institutionnelles nécessaires pour les minimiser, notamment en accompagnant les victimes pour minimiser les plaintes. Les chiffres ne parlent pas d'eux-mêmes à nouveau, les données sont bien souvent « *cuites et recuites* » que brutes, c'est-à-dire toujours obtenues, toujours travaillées et retravaillées, sensibles à nombre de facteurs externes qui président aux multiples étapes de leur production.

**Bien souvent, les données tiennent plus d'indicateurs culturels qu'elles ne décrivent une réalité.**

Pour les chercheuses, nous avons besoin de bien plus de théorie et de bien plus de contexte pour saisir les enjeux de pouvoir des données. Les chercheuses prennent l'exemple des travaux qu'a mené Desmond Patton du SafeLab afin d'utiliser l'IA pour comprendre le rapport à la violence des jeunes de Chicago. Mais pour mieux comprendre un jeu de données de tweets, il a utilisé certains de ces jeunes pour mieux classer les données. Il a ainsi remarqué que des tweets utilisant des termes agressifs, comme le terme « *tuer* » ne visait pas tant à menacer d'autres personnes qu'à faire souvent référence à des paroles de chansons et donc à partager des symboles culturels. Pour les deux chercheuses, cet exemple souligne combien...

**il ne peut y avoir de sciences des données sans le nécessaire travail social qui en éclaire le sens et l'action.**

Communiquer en contexte nécessite non pas de chercher à être neutre, mais à regarder ce que les données produisent effectivement, ex-

pliquent les chercheuses en montrant deux graphiques représentant exactement les mêmes données, le plus exact pouvant être taxé de partialité alors que le second, semblant plus neutre se révèle bien plus manipulateur et plus ouvert à une mauvaise interprétation.

**La justice nécessite aussi de reconnaître et nommer les formes d'oppression que ce soit le racisme, le sexisme, l'homophobie... pas de les neutraliser, expliquent-elles.**

Rendre les méthodes plus robustes et conserver le contexte des données, tel est l'enjeu de Cuidando do Meu Bairro (Prendre soin de mon quartier), une initiative à destination des habitants de Sao Paulo qui vise justement à ajouter du contexte et des éclaircissements aux données locales. Heather Krause, fondatrice de Datasist et We all count, un projet pour développer l'équité en data science, a développé nombre d'outils pour favoriser l'équité et notamment le concept de biographie des données, consistant à demander à leurs producteurs de répondre à des questions de base : « *D'où vient le jeu de données ? Qui l'a collecté ? Quand ? Comment ? Pourquoi ?* ». C'est exactement l'enjeu du « *datasheets for datasets* » développé par Timnit Gebru pour Microsoft. Les guides d'utilisation de données sont une piste essentielle, insistent Klein et D'Ignazio ! Les deux chercheuses signalent ainsi l'exemple de Woman Stats initié par le géographe Chad Emmett et la professeure de sciences politiques Valerie Hudson, une vaste base de données sur les violences faites aux femmes qui distingue des variables de pratiques et de droit selon les conditions juridiques des pays d'où proviennent les données. Une façon de souligner l'incomplétude de leurs données. Ainsi les statistiques sur le viol soulignent que les modalités d'enregistrements sont peu comparables d'un pays à l'autre.

Reste à savoir qui doit éclaircir les données. Bien souvent, les journalistes et les ONG font un bon travail de nettoyage et de contextualisation, pour autant qu'ils bénéficient de financements durables, d'expertises et de normes. Le problème, trop souvent, c'est que gérer les données nécessite de l'investissement au risque de

garder durablement des ressources de qualité médiocres, voire dangereuses.

**Sans argent, la production de contexte risque de rester l'exception plus que la norme.**

# RENDRE LE TRAVAIL VISIBLE

L'ultime principe du féminisme de données consiste à rendre le travail visible. Une étude sur Github, cette plateforme collaborative de travail sur le code source, montrait que les utilisatrices avaient moins de chances de voir leurs contributions acceptées si elles s'identifiaient comme femmes. Pourtant, en permettant de voir les contributions à chaque projet, Github a participé à changer le regard sur la réalité du travail du code, montrant combien il relève d'abord et avant tout d'une collaboration patiente. Mais contrairement à Github, la production de données et de visualisations s'inscrit dans de longues chaînes de collaborations qui tendent à écraser les contributions plutôt qu'à les mettre en avant. La formidable cartographie des navires qui naviguent à travers le monde est le résultat d'un long travail d'innombrables contributeurs. Comme l'explique Miriam Posner, les chaînes d'approvisionnements elles-mêmes, ont trop souvent intérêt à écraser et transformer les contributions. Et on constate la même chose dans le monde de la production de données.

La dénonciation des formes de travail invisibles est ancienne et s'est cristallisée longtemps sur le travail domestique non rémunéré des femmes. Elle s'est considérablement élargie depuis, notamment à celui des utilisateurs des plateformes. Reste que la science des données prospère sur cette invisibilisation en mobilisant largement des plateformes de crowdsourcing que ce soit pour étiqueter des films, des documents, corriger des fautes... comme l'ont montré nombre d'études sur la sous-classe mondiale de la modération, allant de *Ghost Work*

(Mariner Books, 2019, non traduit) de Mary Gray et Siddharth Suri, à *Derrière les écrans* (La découverte, 2020) de Sarah T. Roberts en passant par *En attendant les robots* (Seuil, 2019) d'Antonio Casilli. Derrière les hiérarchies actuelles du travail des données se répètent d'anciennes hiérarchies technologiques fondées sur le sexe, la classe sociale et la race, explique Lilly Irani, comme celles de la première génération d'informaticiennes, ces *Figures de l'ombre* (Harper Collins France, 2017) qu'évoquait Margot Lee Shetterly. En 2008, d'ailleurs, Irani avait lancé le Turkopticon pour permettre aux utilisateurs du Mechanical Turk d'Amazon de signaler les conditions de travail que les commanditaires de la plateforme leur imposaient. Mais après 10 ans de service, le service, entièrement bénévole, a fermé sous l'épuisement. L'exploitation d'une main d'œuvre précaire trouve son origine bien sûr dans la longue histoire coloniale et esclavagiste, qui ne cesse de prolonger ce même schéma d'exploitation.

Pour répondre à ces invisibilisations, nous devons améliorer les études sur la production de données, estiment Klein et D'Ignazio, comme le propose *Anatomy of an AI System* de Kate Crawford et Vladlan Joler. Montrer le travail invisible est un moyen de résister au storytelling de l'innovation dominante.

Documenter ce travail permet également de montrer la quantité de travail physique nécessaire à la numérisation, comme l'a pointé la représentation visuelle du catalogage des livres pour la librairie du Congrès imaginé par Ben Sch-

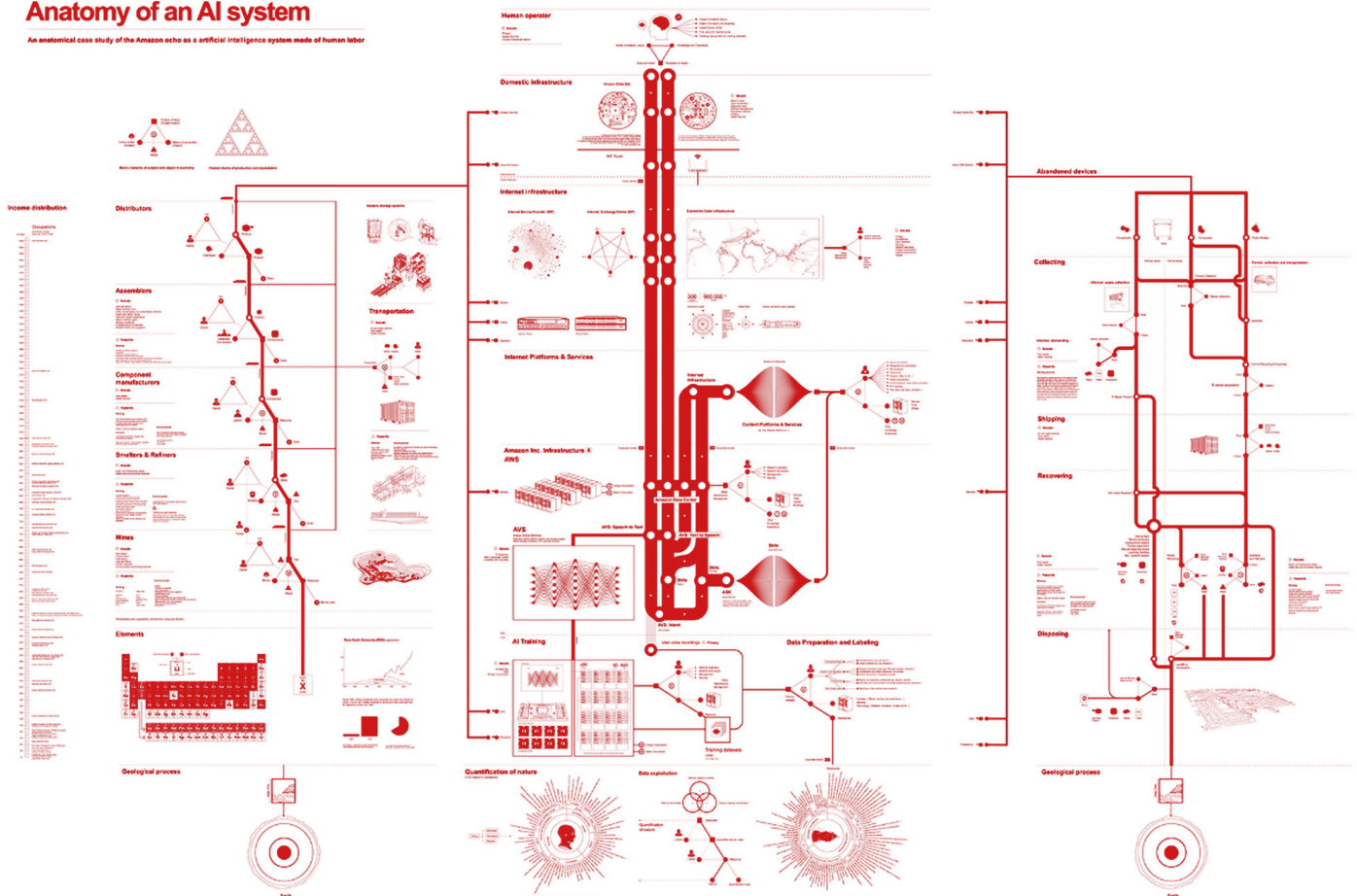
midt. Enfin, bien sûr, c'est un moyen de valoriser également le travail de soin et le travail émotionnel à l'image du projet *Atlas of Caregiving*, visant à documenter le travail nécessaire autour de malades atteints de maladies chroniques afin d'en souligner sa complexité et sa pluralité.

Les déséquilibres du pouvoir sont partout dans les données... Ce n'est qu'en montrant toute la fragilité de chaque donnée que nous pourrions déifier et relever leurs conséquences.

Anatomie d'un système d'IA de Kate Crawford et Vladlan Joler qui tente de cartographier les ressources, les données et le travail humain nécessaire à la conception d'Echo, l'enceinte vocale et connectée d'Amazon.

## Anatomy of an AI system

An anatomical case study of the Amazon echo as a artificial intelligence system made of human labor



# RÉINVENTER NOTRE RAPPORT AUX DONNÉES ET AUX CALCULS

La force du livre de Klein et D'Ignazio consiste à nous confronter à nos propres impasses. Elles nous invitent à mettre à jour nos concepts, à renverser nos catégories de pensées et, pour y parvenir, à dépasser, renforcer et politiser plus avant nos exigences. Peut-être est-il plus que jamais nécessaire, comme elles nous y invitent, de modifier notre cadre de réflexion, de changer les concepts avec lesquels nous légitimons les évolutions des calculs...

Les limites de l'ouverture et de la transparence dans les systèmes techniques sont dénoncées depuis longtemps. Peut-être que pour les dépasser, faut-il remiser et réinventer les concepts qu'on utilise, en tout cas les mettre à jour ! En nous invitant à passer de l'éthique à la justice, du biais à l'oppression, de la loyauté à l'équité, de la responsabilité à la co-libération, de la transparence à la réflexivité, de la compréhension des algorithmes à la compréhension de l'histoire, de la culture et du contexte... l'enjeu que Klein et D'Ignazio formulent nous invite à sortir de la recherche de solutions technologiques aux problèmes que pose la technologie. Ce n'est pas un petit pas de côté. En nous invitant notamment à dépasser les limites de l'ouverture des données ou du gouvernement ouvert, pour revenir au sens et à l'impact des

données elles-mêmes, les deux chercheuses nous montrent une voie d'appropriation et de libération pour résister à l'impact des calculs et des bases de données.

Dans la conclusion de leur ouvrage, Klein et D'Ignazio nous invitent, à l'image des contestations qu'ont connu l'industrie de la technologie, notamment la grève chez Google de 2018 (le Google Walkout for real change), à « occuper en masse les infrastructures numériques », à ralentir les chaînes numériques, à canaliser les solidarités numériques vers les espaces physiques... De la Tech Workers Coalition aux plateformes comme Coworker.org ou aux organisations comme Tech Solidarity en passant par l'essor des coopératives technologiques aux manifestes comme ceux de Design Action Collective et bien sûr aux communautés de pratiques elles aussi revendicatrices comme le réseau Design Justice, Data for Black Lives ou l'Algorithmic Justice League... sont autant de formes de mobilisation qui visent à faire des données un outil de changement social profond plutôt qu'une arme d'oppression au service du pouvoir financier.

La science des données doit désormais dépasser ses présupposés premiers, à savoir ses méthodes uniquement quantitatives reposant sur la seule puissance des données, la force brute des traitements automatisés, sa fausse neutralité, son ouverture sans enjeu...

Le féminisme des données ne vise pas à produire de la technologie pour elle-même, mais au contraire, à en libérer les acteurs. Utiliser les données pour mettre fin à l'oppression est un objectif bien plus stimulant que les utiliser pour la perpétuer.

Cet article est une synthèse du livre *Data Feminism* de Catherine D'Ignazio et Lauren Klein, réalisée par Hubert Guillaud.

Livret édité par l'association Vecteur et le média Danslesalgorithmes.net

Mise en page et conception graphique, Mathurine Guillaud.

**HUBERT  
GUILLAUD**

**DANS  
LES  
ALGORITHMES  
.NET**