Plus Qu'une Simple Politique - Effets Quotidiens de la Gouvernance des Données sur le Scientifique des Données

Vukosi Marivate

Documents de travail DG-006

Apporter de la rigueur et des éléments de preuve à
l'élaboration des politiques économiques en Afrique

AFRICAN ECONOMIC RESEARCH CONSORTIUM
CONSORTIUM POUR LA RECHERCHE ÉCONOMIQUE EN AFRIQUE

Plus Qu'une Simple Politique - Effets Quotidiens de la Gouvernance des Données sur le Scientifique des Données

Par

Vukosi Marivate
Département des sciences informatiques,
Université de Pretoria,
Afrique du Sud

CREA Document de Travail DG-006
Consortium pour la Recherche Economique en Afrique, Nairobi
Février 2023

CETTE ÉTUDE DE RECHERCHE a été rendue possible grâce à une subvention du Consortium pour la Recherche Economique en Afrique. Toutefois, les conclusions, opinions et recommandations sont celles de l'auteur et ne reflètent pas nécessairement les points de vue du Consortium, de ses membres individuels ou du Secrétariat du CREA.

Publié par : Le Consortium pour la Recherche Economique en Afrique
B.P. 62882 - City Square
Nairobi 00200, Kenya

Table des matières

Liste des graphiques

Résumé

En peu de temps, le débat sur la gouvernance des données a pris du retard par rapport aux réalités des industries et des économies axées sur les données. Le flux et le commerce des données sont déterminés par les besoins des différentes parties prenantes et l'évolution des contextes mondiaux de nombreuses technologies qui sont considérées comme locales. Pour le scientifique des données, cela peut sembler être une période passionnante qui offre des possibilités et des opportunités infinies d'inventer le futur proche. L'écart entre la gouvernance des données sur le continent africain et la pratique des données constitue un défi qui doit être relevé au plus tôt. Cet article examine l'intersection de la pratique de la science des données et de la gouvernance des données, j'analyse une partie de la littérature récente pour identifier les domaines de préoccupation et d'intérêt. Enfin, j'examine comment les considérations non techniques sont importantes pour faire le lien entre la gouvernance des données et la pratique de la science des données, en empruntant à d'autres disciplines qui ont eu une avancée sur ces défis. Enfin, nous suggérons des mesures qui peuvent être prises par les praticiens pour réduire cet écart entre la gouvernance et la pratique.

Mots-clés : *Science des données, gouvernance des données, prise de décision en fonction des données, éthique des données.*

1. Introduction

L'essor continu de l'économie de l'information a entraîné une augmentation de l'utilisation des données pour construire et déployer de nombreux produits basés sur les données. Ces produits axés sur les données sont utilisés pour extraire des informations significatives à partir des informations brutes, qui sont ensuite utilisées pour relever des défis dans de nombreux domaines différents. Cette évolution a coïncidé avec l'émergence et le développement de la science des données en tant que domaine d'expertise unique pour la création de produits basés sur les données. La science des données se distingue de l'informatique (l'étude de la théorie et de la pratique du fonctionnement des ordinateurs) et englobe de nombreux domaines (abordés plus loin). Du point de vue des utilisateurs, les produits basés sur les données ont apporté de nombreux nouveaux services et conforts.

Dans le domaine de la santé, par exemple, des outils de données ont été rapidement déployés pour informer le public sur la pandémie de COVID-19 (Alamo et al., 2020 ; Shuja et al., 2021), des modèles de prévision de la pandémie (Evan et al., 2020) et des estimations de l'impact du COVID-19 (Bradshaw et al., 2021). Par ailleurs, certains des outils développés pour traiter les diagnostics/traitements n'ont pas connu le même succès. Un exemple de ces produits axés sur les données est constitué par les nombreux outils/algorithmes qui ont été développés ou déployés pour améliorer les scanners radiologiques (Roberts et al., 2021 ; Wynants et al., 2020). On pourrait être tenté de dire que ces déploiements ont été un échec total. Cependant, ces défis mettent en évidence certaines des lacunes des outils de données et des domaines d'amélioration. Plus important encore, ces défis soulignent la nécessité de gérer les données (et leurs produits) en tenant compte des facteurs humains et des impacts que les données peuvent avoir dans tous les domaines. Dans le cadre du thème COVID-19, la pandémie a également mis en lumière le manque d'infrastructures de données de base (Mbow et al., 2020), le manque de compétences en matière de données et/ou le manque de volonté politique dans de nombreux pays pour se concentrer sur l'amélioration des produits basés sur les données. Ces produits et outils fondés sur les données ont en fin de compte un impact sur la qualité des réponses à la pandémie. Ces exemples soulignent la nécessité d'une gouvernance des données qui adopte une vision affinée des données dans les différents pays et organisations pour illustrer ces différences et gouverner le domaine en conséquence. Je pense qu'il s'agit d'un défi sur lequel les pays africains doivent se concentrer alors qu'ils élaborent également leurs parcours de données (Abebe et al., 2021).

Je considère que le scientifique des données (ou l'équipe de science des données) est celui qui prend la plupart des décisions sur les outils de données qu'il développe ou crée. Cette vision simplifiée n'englobe pas tous les défis associés à ce qui se passe actuellement. Il serait préférable de se pencher sur les produits axés sur les données à travers le regard des systèmes sociotechniques. Les systèmes sociotechniques sont des systèmes qui ont des interactions entre les humains, les machines et l'environnement (Baxter et Sommerville, 2011). Même au sein de l'organisation, l'équipe de science des données ou le scientifique des données ne peut pas prendre de décisions sans une variété de parties prenantes différentes, en particulier les décisions qui ont un impact sur les humains et d'autres facteurs environnementaux. En tant que tel, le scientifique des données devrait être en mesure de comprendre les autres interdépendances des organisations et de la société pour mieux comprendre sa place, et que des structures de gouvernance devraient exister pour guider le développement de systèmes avec de telles interdépendances.

Dans ce travail, je vise à fournir une meilleure compréhension de la gouvernance / des facteurs humains dont les scientifiques de données et les organisations devraient être conscients. Pour relever ce défi, je répondrai à des questions de recherche fondamentales pour le domaine.

- *Question de recherche :* Quels sont les points saillants que les spécialistes des données devraient connaître en matière de gouvernance des données au sein des organisations ?

- *Sous-question de recherche :* Les politiques ou mécanismes actuels sur le continent africain fournissent-ils une vision cohérente qui peut être utilisée par les scientifiques des données pour naviguer et répondre de manière appropriée aux besoins de l'organisation/société ?

- *Sous-question de recherche :* Pouvons-nous apprendre de la communauté ICT4D pour mieux comprendre comment les interventions devraient aller au-delà du simple déploiement d'un outil ?

Il est important de mettre en contexte les raisons pour lesquelles nous devons répondre à ces questions. Nous sommes à une époque où les politiques sont en retard sur le déploiement des outils de données (ce point est abordé dans ce document). Cela signifie que les praticiens de la science des données et les décideurs politiques (dans les secteurs public et privé) ont des lacunes et des angles morts. Ces points morts ont des conséquences. L'élaboration de la politique de protection des données a fait couler beaucoup d'encre, tout comme la pratique et les limites de la science des données. Dans cet ouvrage, je relie les deux afin d'avoir une compréhension commune du fait que la prise de décision doit se faire ensemble.

Le reste du document est organisé comme suit. Tout d'abord, j'examine le domaine de la science des données et la façon dont la gouvernance des données s'inscrit dans

la pratique. L'étape suivante consiste à examiner la gouvernance des données sur le continent africain. Je vais planter le décor et identifier les lacunes qui recoupent ensuite les deux domaines de la science des données et de la gouvernance des données. Dans la section suivante, je discute de la façon dont ICT4D a peut-être déjà ouvert une voie qui nous permet d'apprendre à comprendre les interactions entre la science des données et la gouvernance des données. Les dernières sections traitent des différentes étapes du processus de la science des données et des propositions sur la meilleure façon dont les scientifiques des données peuvent naviguer les facteurs humains tels que la vie privée, la partialité et la sécurité. Enfin, je conclus et résume les points de vue et les preuves élaborés dans ce document.

2. Science des données et pratique

J'examine d'abord la pratique de la science des données et ses liens avec la gouvernance des données. À ce titre, je donne une vue d'ensemble de ce qu'est la science des données - une définition importante qui évolue encore mais qui est importante pour une compréhension commune entre le lecteur et l'auteur.

Qu'est-ce que la science des données ?

La science des données est une discipline qui a vu le jour en raison d'un certain nombre de facteurs. La science des données est un domaine qui utilise des techniques de modélisation scientifique (généralement issues d'un ensemble diversifié de disciplines scientifiques) pour extraire des modèles/informations/connaissances d'une grande variété de données (Dhar, 2013). L'essor de cette discipline a été rapide pour de nombreuses raisons. Les organisations (publiques et privées) se sont efforcées d'explorer les données qu'elles ont amassées au fil du temps et d'extraire des informations pour trouver des modèles et des tendances susceptibles de leur donner un avantage concurrentiel. On a assisté à une explosion du nombre de grandes organisations basées sur Internet et du contenu généré par Internet. Simplement, avec plus d'utilisateurs sur Internet, et plus de contenu sur Internet, l'économie de l'information a besoin de meilleures données et de meilleurs outils de données pour rentabiliser ces utilisateurs (Mandl et Kohane, 2016 ; Zhang et Barr, 2021), par exemple pour la publicité ou pour des services qui motivent les utilisateurs à rester dans les produits d'une entreprise (un jardin clos) (Best, 2014 ; McCown et Nelson, 2009 ; Skorup et Thierer, 2013). Du côté des organisations publiques, la science des données a signifié le travail d'analyse ou de collecte de données qui améliore les services fournis par les gouvernements ou de nouvelles formes de moyens de comprendre les citoyens (résultant parfois en une surveillance de masse/hyper surveillance. Il est très important de comprendre ces facteurs, notamment parce qu'ils sont liés à la "création de valeur à l'ère de l'information" (Nyamwena et Mondliwa, 2020). Ces facteurs nous obligent à comprendre les infrastructures de données fondamentales (physiques, virtuelles, humaines et autres) sous l'angle de la gouvernance, et plus précisément de la gouvernance des données. Commençons par décomposer le processus de la science des données.

Le processus de la science des données

Pour permettre au lecteur de mieux comprendre la science des données, j'utilise les cycles d'analyse des données pour donner un aperçu du processus typique de la science des données. On peut utiliser le CRoss Industry Standard Process for Data Mining (CRISP-DM) comme représentation du processus (Wirth et Hipp, 2000). Les étapes sont généralement les suivantes:

- Comprendre un problème commercial
- Comprendre les données nécessaires
- Collecter les données
- Préparer les données
- Effectuer une modélisation
- Évaluer la solution au problème
- Ajuster la compréhension et/ou le déploiement (voir Figure 1)

L'on remarque que tout ceci vise à résoudre un défi commercial. Nous pouvons facilement étendre cela à la résolution de tout défi sociétal/organisation/scientifique ; il n'est pas nécessaire qu'il s'agisse d'une entreprise. Ce processus est similaire aux épicycles d'analyse (Peng et Matsui, 2015) qui divisent les processus du problème et de l'analyse pour une solution au problème. Le premier essaie de séparer la formulation du problème de la modélisation. La formulation du problème nécessite de comprendre les données correctes à recueillir ou auxquelles il faut avoir accès. En fin de compte, avec tous ces éléments, nous devons comprendre les facteurs et dimensions humains qui apparaissent dans toutes les parties des cycles. Les interdépendances sont abordées plus loin dans ce document.

L'essor de la science des données a également coïncidé avec l'essor de l'apprentissage automatique (ML) et de l'intelligence artificielle - IA (West et Allen, 2018) et on s'attend généralement à ce que les scientifiques des données comprennent et puissent utiliser les concepts de ces domaines (Tang et Sae-Lim, 2016). L'apprentissage automatique est un domaine d'étude qui s'intéresse à la création d'outils qui apprennent des modèles analytiques à partir de données (Alpaydin, 2020) et constitue un sous-ensemble de l'intelligence artificielle. L'intelligence artificielle est un domaine d'étude qui s'intéresse à la création de machines qui imitent l'intelligence des humains, généralement définie comme la création d'un agent capable de percevoir son environnement et d'effectuer des actions pour maximiser une certaine utilité ou atteindre un ou plusieurs objectifs (Russel et Novig, 1995).

Figure1: Modèle de flux CRISP-DM

Source : Jensen (2012)

La plupart des chercheurs/praticiens de la science des données sont également des praticiens/chercheurs en intelligence artificielle et/ou en apprentissage automatique. Ainsi, à partir de maintenant, je ferai référence aux chercheurs/praticiens de la science des données même si je parle d'intelligence artificielle et/ou d'apprentissage automatique. De nombreux chercheurs ou praticiens de la science des données sont à l'aise avec les modèles susmentionnés de compréhension des données et d'analyse ultérieure. Pour que cela réussisse, la société et les organisations ont un besoin croissant de comprendre ce qui se passe pendant le développement et le déploiement d'un système ou d'un modèle dans le monde réel. La gouvernance, à plus d'un titre, entre en jeu. La collecte de données nécessite de prendre en compte les humains et la dynamique humaine (Bender et Friedman, 2018 ; Gebru et al., 2018 ; Seo Jo et Gebru, 2020). Le choix de la modélisation nécessite la prise en compte des personnes

et de leurs besoins (Mitchell et al., 2019), le déploiement nécessite en outre la prise en compte de la dimension humaine sous toutes ses formes (Raji et al., 2020a ; Raji et al., 2020b). À ce titre, la gouvernance des données peut être un outil utile pour le scientifique des données afin qu'il soit conscient de ces facteurs humains et des défis à relever lorsque les humains et les données (collecte, modélisation ou produits) interagissent (Buolamwini et Gebru, 2018 ; Hooker, 2021 ; Ledford, 2019 ; Mehrabi et al., 2021 ; Sujan et al., 2019).

Pourquoi avons-nous besoin de la gouvernance des données ?

Du point de vue des gouvernements, dans le cadre du développement et de la croissance économique, ils veulent embrasser la « création de valeur à l'ère de l'information » (Nyamwena et Mondliwa, 2020). Pour ce faire, la collecte, l'utilisation et le flux des données doivent être gouvernés afin de pouvoir avoir un droit de regard sur cette création de valeur. En bref, la gouvernance des données doit toucher chaque partie du cycle de vie de la science des données, comme nous l'avons vu précédemment. La gouvernance des données prend également de l'importance en raison des pressions historiques en faveur de la numérisation des pays, notamment des pays africains. Les gouvernements craignent de rester à la traîne en matière de développement économique s'ils ne tirent pas parti des possibilités offertes par les données. Le défi se pose lorsque nous examinons la manière dont la gouvernance des données doit être façonnée pour les différents pays. Sans une gouvernance des données adéquate dans les pays, les opportunités pour les secteurs public et privé risquent de ne pas réaliser le plein potentiel de l'économie de l'information. Il s'agit d'un risque important car des produits qui ne correspondent pas aux valeurs des citoyens du pays peuvent être déployés et finalement causer des dommages. De tels exemples de manquement sont les protections inadéquates de la confidentialité (Metcalf et Crawford, 2016), les limitations sur ce à quoi les données peuvent être utilisées, la réglementation des produits basés sur les données qui pourraient être nuisibles (Metcalf et Crawford, 2016), les lignes directrices sur la souveraineté des données (Hummel et al., 2021), et comment des ensembles spécifiques de données devraient être traités comme des biens publics à partager à l'intérieur ou à l'extérieur d'un pays Borgesius et al., 2015). La bonne gouvernance des données ne concerne pas seulement l'étape de création des données, mais la façon dont la gouvernance imprègne l'ensemble du cycle de la science des données (Metcalf et Crawford, 2016). En outre, une bonne gouvernance des données nécessite la connaissance contextuelle des décideurs (dans le secteur public et dans le secteur privé) pour comprendre le cycle de la science des données (données, modélisation, algorithmes, etc.) (Keans et Roth, nd). Il est plus difficile pour les gardiens de réglementer l'industrie s'ils n'ont pas eux-mêmes une compréhension fondamentale de ce qui se passe généralement dans le cycle de la science des données. Il s'agit d'un point important à souligner, car les industries telles que la finance, par exemple, ont des régulateurs bien définis dans la

plupart des pays. Ces régulateurs financiers réglementent l'industrie pour atténuer la corruption et les dommages. Les conseils de régulation sont composés d'experts dans le domaine qui travaillent ensuite pour définir les meilleures pratiques, les limites et les sanctions en cas de violation des règlements. La difficulté de bon nombre de produits axés sur les données que nous voyons aujourd'hui réside dans le fait que beaucoup de décideurs qui déploient ces outils n'ont que peu d'expérience dans le domaine lui-même, et considèrent la plupart de ce qui se passe comme une boîte noire qui prend des données et produit des réponses "par magie". Cela souligne la nécessité d'une réglementation fondamentale de base qui pose les bonnes questions lors du développement de produits axés sur les données, mais qui ouvre également la voie à une compréhension commune du domaine qui devrait être comprise par tous (et pas seulement par les experts). Dans la section suivante, j'examine les parties importantes du cycle de la science des données et je souligne les facteurs humains et les questions qui devraient être posées par les scientifiques des données et comprises par les décideurs.

3. Les facteurs humains et le cycle de la science des données

Afin de promouvoir la compréhension conjointe de la science des données et de la gouvernance des données, j'aborde dans cette section les facteurs humains dans les phases d'acquisition, de modélisation et de présentation des données du cycle de la science des données.

Acquisition des données

L'une des étapes qui suscite des tensions dans le processus de science des données est le processus d'acquisition des données. Il peut s'agir d'un point mort (Mitchell et al., 2018 ; Zhang et al., 2018) qui peut faire ou défaire de nombreux projets. Imaginez que vous utilisiez un ensemble de données collectées dans les années 1950 sur les prêts financiers accordés par les banques. Maintenant, la construction d'un outil prédictif pour aider les décisions de prêt avec un tel ensemble de données sera pleine de biais de genre et de race dans de nombreux pays (Bond et Tait, 1997 ; Rice, 1996). En d'autres termes, le modèle apprendrait à discriminer. C'est encore un défi aujourd'hui (Fu et al., 2021). Même si les données sont considérées comme représentatives de la population étudiée, elles peuvent encoder des préjugés et des discriminations sociétales. La plupart du temps, lorsqu'ils interagissent avec des décideurs ou des clients, les personnes sans grande expérience ont tendance à négliger les défis liés à l'acquisition de données. Ces défis sont liés à des questions de gouvernance (Veale et Bins, 2017).

Processus et procédures

En acquérant des données, dans le cadre du processus de la science des données, on relie le problème abordé aux données qui seront nécessaires pour le résoudre. À un moment donné, il peut y avoir des données avant que les questions ne soient claires, alors qu'à d'autres moments, il y a une question à laquelle il faut répondre mais les données n'ont pas été cartographiées. Dans tous les cas, les données doivent être déplacées de l'endroit où elles se trouvent et mises en scène pour être traitées par l'équipe de science des données. Pour cela, il faut identifier la source de données pertinente, déterminer quel sous-ensemble d'informations est important et comment la transmission se fera. En effectuant ces étapes d'identification, nous devons tenir compte des facteurs humains.

Facteurs humains

Pour chacune des étapes du processus de la science des données, je me concentre sur trois facteurs humains. Pour l'acquisition des données, je me concentre sur : D'où viennent les données ? Pourquoi sont-elles/étaient-elles collectées ? Sur qui portent les données ? Il existe de nombreux autres facteurs, mais pour des raisons de concision et pour communiquer notre message, nous nous en tiendrons à trois facteurs par étape du cycle de la science des données.

D'où viennent les données ? Lorsque l'on identifie la source des données, il devient rapidement évident que l'on doit comprendre les structures des organisations internes ou externes qui contrôlent l'accès et l'utilisation des données. Dans un cas idéal, il existe une structure de gouvernance des données claire qui fournit également des informations sur la façon dont un scientifique des données peut demander des données, comment les données doivent être traitées et toute information sensible et saillante dont le scientifique doit être conscient (Abraham et al., 2019). Il y aura des questions qui sont liées à la sensibilité des données. Les données ont-elles été collectées de manière éthique ? Les données font-elles partie d'un dépôt de données ouvert ? Sous quelle licence les données sont-elles placées et quelles sont les attentes en matière d'utilisation ? Les données proviennent-elles d'une entité gouvernementale ? Quelles sont les attentes nationales en matière de données gouvernementales ouvertes ? Par exemple, dans une municipalité, on peut s'attendre à ce que les données agrégées sur l'utilisation de l'eau par quartier municipal soient ouvertes et disponibles (d'autant plus que de nombreuses zones dans certains pays sont confrontées à des pénuries d'eau), mais il se peut que certains fonctionnaires résistent à rendre ces données disponibles. Cela peut être dû au fait qu'il n'y a pas assez de ressources humaines pour créer et maintenir les données disponibles, que les données sont normalement disponibles moyennant des frais qui augmentent les recettes, qu'il peut y avoir des problèmes de transparence, etc.

Pourquoi les données sont-elles/étaient-elles collectées ? Il s'agit d'un facteur important, car il établit les attentes préalables quant à l'utilisation des données qui ont été ou sont collectées. Si nous imaginons que nous disposons de données sur les habitudes de transaction des usagers des bus dans une ville, l'utilisation initiale des données et les attentes étaient de gérer le système de transport. Si, à présent, les données sont utilisées pour comprendre le comportement des usagers du bus et leur proposer de la publicité, cette nouvelle utilisation peut ne pas être couverte par les termes de référence initiaux. Plus important encore, les usagers du bus peuvent ne pas être d'accord avec le changement d'utilisation de leurs données, et l'organisation a la responsabilité de traiter leurs informations avec soin et réflexion.

Sur qui portent les données ? En menant à bien le processus de constitution des données, il faut se demander si elles sont représentatives de la population qu'elles servent. Encore une fois, lorsque les données portent sur des personnes, nous devons comprendre qui elles représentent et si cette distribution est équitable, juste (Mitchell et al., 2018 ; Zhang et al., 2018). De plus, cette distribution de personnes correspond-

elle à celles pour lesquelles nous nous attendons à prendre des décisions dans le produit final basé sur les données ? Si ce n'est pas le cas, cela peut être un problème qui introduit une prise de décision biaisée. Par exemple, au cours de la dernière décennie, on a beaucoup insisté sur le biais des systèmes de reconnaissance faciale (Raji et al., 2020). Une partie de ce biais provient des données originales qui ont été utilisées pour les entraîner (Mitchell et al., 2018 ; Zhang et al., 2018). Une partie de ce biais provient des conceptions des systèmes et de la façon dont le succès est mesuré. J'en parlerai plus tard dans les sous-sections sur la modélisation et la présentation.

Il suffit de voir ce qui précède pour comprendre qu'il existe des facteurs humains importants qui ne peuvent pas être laissés à la discrétion du scientifique ou de l'organisation. Il est nécessaire de définir des attentes fondamentales en matière de traitement et de stockage des données, de sécurité, d'éthique et de tests réglementaires quant à l'utilisation des données.

Analyse des données et modélisation

Lors de l'étape d'analyse et de modélisation des données, le scientifique des données concentre son énergie sur l'utilisation des bonnes approches pour extraire des informations significatives des données. Ces choix influenceront le résultat et constitueront la base sur laquelle beaucoup choisiront de croire ou non les résultats. Même s'il s'agit d'approches informatiques, statistiques ou mathématiques établies, nous devons comprendre comment les choix influencent le produit final et les personnes.

Processus et procédures

Le scientifique des données prend les données qui ont été acquises à l'étape précédente. Il procède ensuite à leur nettoyage, en les transformant en une forme utilisable par les tâches de modélisation en aval, puis en les chargeant dans ses systèmes de modélisation. Le scientifique des données fera des choix sur les métriques à mesurer ou à optimiser. En fin de compte, ces mesures sont utilisées pour décider du succès de l'opération et permettent de savoir si de nouvelles données doivent être obtenues, si la question doit être reformulée ou si l'on peut passer à l'étape suivante du cycle de la science des données.

Facteurs humains

Pour les étapes d'analyse des données et de modélisation, je me concentre sur ces facteurs : Comment les choix de modélisation sont-ils faits ? Qui a les compétences pour modéliser ? Quels sont les modèles pour le cas d'utilisation employé ? Comment les choix de modélisation sont-ils faits ? Pendant un certain temps, il y avait une réplique populaire selon laquelle les gens sont biaisés et les machines sont

impartiales. Un certain nombre de travaux ont mis en évidence que les machines ne peuvent pas être impartiales car les données qu'elles utilisent pour apprendre peuvent être biaisées (Birhane et Cummins, 2019). Après cela, l'aiguille s'est déplacée vers le fait que les algorithmes ne peuvent pas être biaisés, seulement les données (Birhane et Cummins, 2019). Mais cela ne tient toujours pas compte des nombreux facteurs qui font que les choix de modélisation ont également un impact sur les résultats des modèles finaux (Jiang et al., 2020). Dans le domaine de l'apprentissage automatique, nous sommes fiers de travailler à la construction d'algorithmes de mieux en mieux généralisables, précis et efficaces, mais cela ne nous dispense pas de réfléchir à nos choix de modélisation (Birhane et al., 2021). Les travaux de Hooker et al. (2020) ont mis en évidence les biais des modèles réduits.

En outre, de plus en plus de modèles d'apprentissage automatique utilisent l'apprentissage par transfert (en se basant sur des modèles ou des ensembles de données antérieurs). Cela entraîne alors des biais en avant. C'est l'une des raisons pour lesquelles les scientifiques des données devraient s'efforcer de documenter leurs choix de modélisation (Mitchell et al., 2019). La modélisation peut sembler insignifiante au moment de la prise de décision, mais peut entraîner de grandes conséquences par la suite. Un exemple récent (Birhane et al., 2021) montre comment les modèles influencent la collecte d'ensembles de données massifs (pour lutter contre les biais) qui, lorsqu'on les examine au microscope, ne sont pas aussi représentatifs que les auteurs des ensembles de données le prétendaient. Cela met en évidence le manque de participation et les choix de conception inclusifs qui remettent également en question les compétences en matière de modélisation?

Qui a les compétences pour modéliser ? Le domaine des sciences des données, de l'intelligence artificielle et de l'apprentissage automatique est généralement biaisé en termes de démographie et de personnes qui finissent par construire les technologies sous-jacentes. On pourrait dire que cela ne s'applique pas au continent africain en ce qui concerne la composition raciale. Mais ce n'est pas le reflet exact de ce domaine. Pendant longtemps, dans les grandes entreprises technologiques du continent, les rôles techniques supérieurs étaient occupés par des hommes et des blancs (ce qui reflète les problèmes critiqués à propos de la Silicon Valley). Le manque de compétences en science des données sur le continent aggrave encore la situation. Sans ces compétences, la connexion entre les décideurs et ceux qui conçoivent les modèles est encore plus faible. Combien de décideurs ont une formation en données/calculs ? Un autre facteur est que les grandes entreprises de technologie qui font tourner la majeure partie de l'économie d'Internet ont tendance à n'avoir que des bureaux commerciaux sur le continent (Birhane, 2020). Leur objectif est de vendre leurs services (Birhane, 2020), d'extraire des données (Coleman, 2018) et de gérer les questions réglementaires - s'il y a une réglementation (Birhane, 2020 ; Coleman, 2018). Les bureaux ne construisent ni ne façonnent les technologies de base de ces entreprises. En tant que tel, si nous relions cette question à la précédente, nous voyons comment les choix de modélisation peuvent devenir une décision qui change la vie de ceux qui sont chargés des tâches en aval. Imaginez comment, dans les

entreprises, des systèmes d'embauche automatisés ont été déployés pour faciliter le processus d'embauche en utilisant l'IA pour sélectionner ou surveiller les candidats. Il a été démontré que ces systèmes sont discriminatoires (Sánchez-Monedero, 2020), mais quelles sont les chances que les décideurs et les équipes internes de science des données aient les compétences nécessaires pour évaluer leurs systèmes de reconnaissance du visage ou leurs services de filtrage de texte contre les préjugés ?

Quels sont les modèles pour le cas d'utilisation employé ? Les travaux récents dans le domaine de l'apprentissage automatique et de l'intelligence artificielle ont mis en évidence des modèles explicables dans la lutte contre les préjudices et la recherche d'une meilleure équité. Prenons, par exemple, l'augmentation des systèmes de surveillance et des systèmes de reconnaissance faciale au niveau international. La façon dont les modèles sont choisis et évalués pour de tels cas d'utilisation a une incidence sur l'impact final que ces systèmes auront sur la société. De nombreux travaux ont mis en évidence la façon dont les systèmes de reconnaissance faciale biaisés (Raji et al., 2020) peuvent entraîner un comportement discriminatoire de la part des forces de l'ordre. Cela peut finir par être une situation de vie ou de mort pour quelqu'un à l'extrémité de ces systèmes automatisés.

Un scientifique des données et un décideur doivent se demander quel est le coût d'une erreur de notre modèle. Cela devrait ensuite avoir un impact sur la manière dont le déploiement est effectué. En outre, il peut y avoir des restrictions réglementaires pour faire un choix ou un autre, en fonction des attentes de la société.

Présentation et déploiement de produits basés sur des données

L'étape finale de nombreux projets de science des données est la présentation des résultats aux décideurs et/ou le déploiement des produits basés sur les données.

Processus et procédures .

À cette étape, le scientifique des données s'efforcera de présenter un rapport sur les résultats de la modélisation pour répondre aux questions initiales. À partir de là, des décisions peuvent être prises sur la base de ces rapports. Les rapports peuvent être des visualisations, des simulations ou des produits basés sur des données avec des mesures qui montrent leur efficacité. Des décisions seront prises sur ce qu'il convient de montrer et à qui les produits basés sur des données seront destinés. Ceux-ci ont des facteurs humains.

Facteurs humains

Pour les étapes de Présentation et de Déploiement des produits basés sur les données, je me concentre sur ces facteurs : Quelles décisions sont prises avec les modèles ? Quels choix sont faits quant à ce qui doit être montré ? Comment les modèles seront-ils mis à jour ?

Quelles décisions sont prises à l'aide des modèles ? Le test ultime de l'utilité d'un modèle pour le décideur est lorsqu'il est déployé pour être utilisé ou présenté pour la prise de décision. Il s'agit d'un moment du cycle de vie de la science des données qui nécessite une compréhension approfondie des parties précédentes du cycle, faute de quoi de mauvaises décisions pourraient être prises. Lorsqu'il examine le produit de données ou les prédictions d'un modèle, l'utilisateur doit comprendre comment le modèle fonctionne, comment il a été construit et quelles sont ses limites. La sous-question ici pourrait être : comment les gens interprètent-ils les résultats/prédictions du produit de données ? Pour cela, il ne suffit pas d'afficher un résultat, il faut aussi travailler avec des praticiens de l'interaction homme-machine pour concevoir le modèle de manière équitable et transparente et atténuer les préjugés ou la discrimination (Holstein, 2019 ; Seng Ah Lee et Singh, 2021).

Quels choix sont faits dans ce qui doit être montré ? Comme dans le domaine statistique, nous pouvons aussi mentir avec les produits basés sur les données. La pandémie de COVID-19 a donné lieu à de nombreux exemples où les décideurs se sont efforcés de déformer les données, de fausser les prédictions des modèles et même de censurer les données des chercheurs et des praticiens pour qu'elles correspondent à l'opinion du décideur (Abazi, 2020 ; Zhang et Barr, 2021). On peut considérer qu'il s'agit là d'un exemple public extrême, mais cela se produit de nombreuses façons. L'une d'elles peut consister à tester les dommages au moment de l'exécution.

Comment les modèles seront-ils mis à jour ? Lors du déploiement de produits basés sur des données, les modèles internes doivent être mis à jour régulièrement. Le monde n'a pas cessé de changer lorsque le modèle a été formé et déployé. Les modèles vont donc commencer à dériver. Cette dérive peut également provenir de la façon dont les utilisateurs réagissent à ce que fait le modèle. L'organisation de l'équipe de science des données a-t-elle des procédures sur l'entretien des modèles dans le produit piloté par les données, et comment testons-nous le dérapage avant que le système ait une erreur élevée dans ses résultats (prédictifs, prescriptifs, diagnostiques, etc.) ?

Dans cette section, j'ai examiné comment la science des données et la gouvernance des données se croisent. Dans la dernière partie de la section, j'ai choisi trois sections des cycles de la science des données pour pouvoir analyser les facteurs humains. En identifiant ces facteurs humains, nous pouvons mieux comprendre comment la gouvernance des données fait partie intégrante du cycle complet, car les décisions prises par le scientifique auront un impact sur les utilisateurs et les humains en général. Dans la section suivante, j'aborde la gouvernance des données sur le continent africain.

4. La gouvernance des données et le continent africain

Avec les appels aux pays africains pour qu'ils profitent des avancées actuelles des économies basées sur les données, il y a eu quelques initiatives vers des stratégies et des politiques de gouvernance par les gouvernements qui couvrent les données. L'Union africaine a publié « La stratégie de transformation numérique pour l'Afrique 2020-2030 » (Union africaine, 2020). Cette stratégie doit être comprise dans le contexte des défis plus larges et plus localisés de la gouvernance des données et de la numérisation dans différents pays africains.

En ce qui concerne la protection de la confidentialité, le règlement général européen sur la protection des données (RGPD) (Commission européenne, nd) a eu un effet et un impact très larges sur l'économie de l'internet, car de nombreuses entreprises qui traitaient les données des citoyens européens ont dû se conformer aux règles établies par l'UE. Sur le continent africain, comme le montre l'étude de Davis (2021), des efforts sont déployés pour renforcer les politiques de protection des données, même si seulement 52% des pays africains disposent d'une telle législation.

La Convention de l'Union africaine sur la cyber sécurité et la protection des données personnelles (connue sous le nom de Convention de Malabo) (Union africaine, 2014) a été adoptée par les États membres de l'UA en 2014. Elle vise à assurer la protection des cyber-infrastructures, la protection des informations personnelles, la cyber sécurité et les bases nécessaires pour permettre une économie de l'information à travers le continent africain. Bien que ratifiée en 2014, seuls 8 pays avaient ratifié la convention au 18 juin 2020 (Commission européenne, nd). La convention touche à de nombreux aspects qui peuvent former une base unifiée pour que les pays africains puissent bénéficier de l'économie de l'information. Sans ratification, nous avons la réalité que les organisations et les praticiens n'ont pas une vue unifiée sur la façon de déployer les outils de données et, pour certains pays, la réalité est bien pire avec des protections très laxistes ou inexistantes (Davis, 2021).

En Afrique du Sud, la loi sur la protection des informations personnelles (POPIA) (Gouvernement d'Afrique du Sud, nd), qui a mis de nombreuses années à être promulguée, a également lancé une discussion dans le public sur l'acquisition des données, la protection des informations personnelles et l'utilisation des données pour des actions en aval (surtout lorsqu'elles ne sont pas destinées à l'objectif initial de la collecte des données). Malgré cela, la gouvernance des données ne se limite pas à la protection des informations personnelles ; les données interagissent avec

de nombreux autres facteurs humains et organisationnels. J'espère que la section précédente a montré clairement que la gouvernance des données ne doit pas se limiter aux données utilisées.

Mais, comme nous l'avons vu précédemment, de nombreux facteurs humains doivent être pris en compte à toutes les étapes du cycle de la science des données. Pour gérer efficacement l'ensemble du processus, les pays doivent avoir une compréhension claire des étapes et des responsabilités des gouvernements envers les scientifiques spécialisés dans les données, ainsi que des responsabilités des scientifiques spécialisés dans les données envers le public.

Le continent africain a fait de grands progrès dans le secteur des TIC et le renforcement des compétences locales et la défense des entreprises locales (Ponelis et Holmner, 2015). Malgré cela, il y a toujours une domination des géants de la Big Tech (Microsoft, IBM, Google, Facebook, etc) sur le continent physiquement ou avec des services qui traversent les frontières. Même si nous n'avons pas de définition commune de la pénurie de compétences en matière de données, le travail de Sey et Mudongo (2021) met en évidence le manque de compréhension du besoin de compétences en IA et la nécessité de déployer des efforts pour développer ces compétences sur le continent, ce qui doit relier les secteurs public et privé. Ces observations sont importantes car elles mettent en contexte le fait que peu de grandes entreprises technologiques ont peu ou pas de recherche et développement sur le continent. Les compétences en gouvernance de l'IA sont recommandées dans le cadre du développement des compétences en IA sur le continent (Sey et Mudongo, 2021), ce qui fait écho au message de ce document sur le lien plus large entre la science des données et la gouvernance des données..

Le continent risque de n'être qu'une source de données (Birhane, 2020) pour créer des services qui sont ensuite utilisés par les citoyens sans aucun développement local de ces services. Cela a été récemment mis en évidence par le fait que seulement 13 % de l'équipe chargée de lutter contre les abus sur les plateformes en ligne de Facebook travaillent sur des contenus non américains, même si 90 % des utilisateurs de Facebook se trouvent en dehors des États-Unis (Purnell et al., 2021). Ce point est important car les fausses informations diffusées sur Facebook en dehors des États-Unis ont un effet sur de nombreux pays, mais ne peuvent être combattues par Facebook lui-même. En outre, les gouvernements doivent être en mesure de gouverner l'espace numérique et de veiller à ce que les citoyens puissent bénéficier des biens publics numériques (Gillwald et van der Spuy, 2019).

Un autre défi est l'utilisation de certains des produits basés sur les données pour la surveillance par les gouvernements et le secteur privé sur le continent (Mudongo, 2021). Comme nous l'avons déjà souligné, les systèmes sont moins susceptibles d'être développés localement et peuvent contenir des préjugés et conduire à la discrimination. Cela illustre une autre lacune de la gouvernance (qu'elle soit planifiée ou non), car les décideurs doivent être en mesure d'évaluer les risques et les préjudices que ces systèmes peuvent causer à la population (Mudongo, 2021).

5. Étude de cas : Apprendre de notre passé récent, entrée en scène de ICT4D

La science des données et l'intelligence artificielle ont été saluées comme la solution miracle à de nombreux problèmes ; les données elles-mêmes sont considérées comme le nouveau gisement de pétrole que les nations et les organisations doivent exploiter (Hirsch, 2013). Mais un défi que les organisations et les nations devraient être en mesure de repérer refait surface. Avec l'essor des TIC et des efforts de numérisation, de nombreux problèmes ont été pointés du doigt où les TIC pourraient être la solution (Curtis, 2019). Lancé dans les pratiques de développement, ICT4D a été une force pour les deux dernières décennies ou plus (Walsham, 2017).

Je soutiens que nous avons maintenant assez de temps pour constater que certaines des lacunes de la vision de nombreux problèmes comme nécessitant les TIC comme solution, en particulier de la part de praticiens qui viendraient de l'extérieur, s'installeraient, se déploieraient puis partiraient, sont très similaires à ce qui se passe actuellement dans le monde de la science des données et doit être modifié (Shilton et al., 2021). Il peut y avoir des différences, dont la principale est la familiarité avec ce que sont les TIC et moins avec ce que sont la science des données, l'intelligence artificielle ou l'apprentissage automatique (Osoba et Welser, 2017). Fondamentalement, les chercheurs et les praticiens de la science des données sont simplement considérés comme des magiciens à qui vous lancez un problème et des données, et une solution arrive de l'autre côté. Nous le voyons avec l'avènement des stratégies 4IR pour les nations africaines qui sont dirigées par des institutions publiques qui n'ont pas les compétences ou les connaissances pour vraiment s'engager dans le sujet qu'ils présentent comme une solution à de nombreux problèmes auxquels ils sont confrontés (McBride et al., 2018 ; Moorosi et al., 2017).

Dans le domaine des ICT4D, un débat historique a porté sur l'efficacité d'avoir des chercheurs et des praticiens qui ne sont pas des locaux et qui proposent des " solutions " utilisant les TIC pour de nombreux problèmes de développement (Andrade et Urquhart, 2012). Au fil du temps, cette question est devenue un domaine d'étude au sein même du secteur. Il est devenu très évident que le développement et la conception des systèmes doivent être participatifs (Andrade et Urquhart, 2012 ; Tongia et Subramanian, 2006 ; Toyama, 2015) et ne pas se limiter au défi technique. Ce travail difficile a nécessité du temps et de nombreux échecs. En revanche, dans le domaine de la science des données et de l'intelligence artificielle, de nombreux efforts ont été déployés pour comprendre l'équité, l'éthique et les effets à long terme

des interventions techniques. C'est un changement bienvenu dans l'histoire des ICT4D, mais nous sommes encore en retard dans la compréhension de la nécessité d'une conception et d'une gouvernance participatives qui guident le domaine (Singh et Flyverbom, 2016). Nous disposons de grands organismes internationaux tels que l'Union internationale des télécommunications, auxquels de nombreux États appartiennent, qui ont façonné les politiques de TIC dans toutes les régions.

Dans le domaine de l'intelligence artificielle, on peut dire que le débat sur l'équité et le préjudice a été très ouvert en raison des menaces d'impact à grande échelle sur les personnes. Mais cela ne signifie pas que les débats résolvent les problèmes. Dans la plupart des débats et des discussions, ce sont surtout les chercheurs et non les décideurs et les responsables politiques qui font le travail pour documenter les dommages et faire des recommandations pour les atténuer (Whittaker et al., 2018). Les décideurs doivent aussi venir à la table pour également façonner le débat en apportant la contribution du gouvernement. Nous devons tirer les leçons d'autres domaines tout en comprenant le caractère unique de l'adoption des produits basés sur les données avant même de penser à leur impact.

6. Conclusion

Dans cet ouvrage, j'ai utilisé une étude de la littérature relative à la science des données et à la gouvernance des données pour mettre en évidence les liens qui existent entre ces deux domaines. Laisser les décisions de conception au seul scientifique des données, c'est ignorer les nombreux facteurs humains que comportent les produits basés sur les données. En tant que telle, la gouvernance des données est un élément clé pour pouvoir créer et déployer des produits qui contribuent aux économies en développement du continent tout en limitant les dommages. Pour cela, les pays africains doivent avoir une idée des besoins en matière de gouvernance et des compétences nécessaires à l'adoption de politiques efficaces. L'étude de cas présentée sur les ICT4D nous permet d'apprendre d'une discipline connexe qui est active depuis deux décennies et qui a dû relever des défis similaires pour déployer des interventions dans le Sud.

Recommandations

- Les gouvernements africains doivent travailler ensemble pour mettre en œuvre une politique de gouvernance des données. La réalité flagrante selon laquelle seuls 8 pays (à ce jour) ont ratifié la Convention de l'Union africaine sur la cyber sécurité et les données personnelles laisse beaucoup à désirer.

- Les industries publiques et privées doivent s'engager avec les scientifiques des données pour mieux comprendre les domaines de préoccupation soulignés dans ce document au-delà de la confidentialité des données. La plupart des politiques sur le continent se concentrent sur les protections de la confidentialité et sur certaines prises de décision automatisées, mais de nombreuses autres décisions prises dans le processus de développement des outils de données ont un impact sur le résultat.

- Pour le scientifique spécialiste des données, il doit être évident que la politique et le développement d'outils de données vont de pair. Même si les politiques nationales, régionales ou continentales n'ont pas rattrapé leur retard, il existe un phénomène croissant au sein de notre pratique qui vise à développer les meilleures pratiques et à mettre en évidence les défis en matière d'éthique, d'équité et d'atténuation des abus.

Remarques

1. Adresse de l'auteur : Vukosi Marivate, vukosi.marivate@cs.up.ac.za, Département des sciences informatiques, Université de Pretoria, Pretoria, Afrique du Sud.

2. Je tiens à remercier les organisations de base de IA/ML/DS à travers le continent africain et la diaspora qui ont travaillé à façonner notre participation et à façonner les technologies en question. Je tiens également à remercier le groupe de recherche en science des données pour l'impact social de l'Université de Pretoria qui m'a permis d'explorer ces sujets avec eux. Enfin, je tiens à remercier l'ABSA qui soutient financièrement le président des sciences des données de l'UP ABSA..

Références

Abazi, V. 2020. “Truth distancing? Whistleblowing as remedy to censorship during COVID-19”. *European Journal of Risk Regulation*, 11, 2: 375–381.

Abebe, R., Kehinde Aruleba, Abeba Birhane, Sara Kingsley, George Obaido, Sekou L. Remy and Swathi Sadagopan. 2021. Narratives and counternarratives on data sharing in Africa. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 329–341.

Abraham, R., Johannes Schneider and Jan Vom Brocke. 2019. “Data governance: A conceptual framework, structured review, and research agenda”. *International Journal of Information Management*, 49: 424–438.

African Union. 2014. *African Union Convention on Cyber Security and Personal Data Protection.* Addis Ababa: African Union.

African Union. 2020. *The Digital Transformation Strategy for Africa (2020-2030)*. Addis Ababa: African Union.

Alamo, T., Daniel G. Reina, Martina Mammarella and Alberto Abella. 2020. “COVID-19: Open-data resources for monitoring, modeling, and forecasting the epidemic”. *Electronics*, 9, 5: 827.

Alpaydin, E. 2020. *Introduction to machine learning*. Massachusetts: MIT Press.

Andrade, A.D. and Urquhart, C. 2012. “Unveiling the modernity bias: A critical examination of the politics of ICT4D”. *Information Technology for Development,* 18, 4: 281–292.

Baxter, G. and Sommerville, I. 2011. “Socio-technical systems: From design methods to systems engineering”. *Interacting with Computers,* 23, 1: 4–17.

Bender, E.M. and Friedman, B. 2018. “Data statements for natural language processing: Toward mitigating system bias and enabling better science”. *Transactions of the Association for Computational Linguistics*, 6: 587–604.

Best, M.L. 2014. “The internet that Facebook built”. *Communication ACM*, 57, 12: 21–23.

Birhane, A. and Cummins, F. 2019. “Algorithmic injustices: Towards a relational ethics”. *arXiv preprint arXiv,* 1912.07376.

Birhane, A. Kalluri, P., Card, D. Agnew, W. Dotan, R. and Bao, M. 2021. “The values encoded in machine learning research”. *arXiv preprint arXiv*: 2106.15590.

Birhane, A., Prabhu, V.U. and Kahembwe, E. 2021. “Multimodal datasets: misogyny, pornography, and malignant stereotypes”. *arXiv preprint arXiv:* 2110.01963.

Birhane, A. 2020. “*Algorithmic colonization of Africa*”. SCRIPTed 17: 389.

Bond, P. and Tait, A. 1997. “The failure of housing policy in post-apartheid South Africa”. *Urban Forum*, Vol. 8., Springer, 19–41.

Borgesius, F.Z., Gray, J. and van Eechoud, M. 2015. "Open data, privacy, and fair information principles: Towards a balancing framework". *Berkeley Technology Law Journal,* 30, 3: 2073–2131.

Buolamwini, J. and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency. PMLR, 77–91.

Bradshaw, D., Dorrington, R.E., Laubscher, R. Moultrie, T.A. and Groenewald, P. 2021. "Tracking mortality in near to real time provides essential information about the impact of the COVID-19 pandemic in South Africa in 2020". *South African Medical Journal*, 111, 8: 732–740.

Coleman, D. 2018. "Digital colonialism: The 21st century scramble for Africa through the extraction and control of user data and the limitations of data protection laws". Michigan Journal of Race and Law, 24: 417.

Curtis, S. 2019. "Digital transformation—the silver bullet to public service improvement?" *Public Money and Management*, 39, 5: 322–324.

Dhar, V. 2013. "Data science and prediction". *Communication ACM,* 56, 12: 64–73.

Davis, T. 2021. Data protection in Africa: A look at OGP member progress. Technical Report. Alt Advisory.

European Commission. nd. 2018 reform of EU data protection rules. European Commission. https://ec.europa.eu/commission/sites/betapolitical/files/data-protection-factsheet-changes_en.pdf.

Fu, R., Yan Huang, Y. and Vir Singh, P. 2021. "Crowds, lending, machine, and bias". *Information Systems Research*, 32, 1: 72–92.

Gebru,T., Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III and Kate Crawford. 2018. "Datasheets for datasets". *arXiv preprint arXiv:*1803.09010.

Gillwald, A. and Spuy, Anri van der. 2019. "The governance of global digital public goods: Not just a crisis for Africa". *GigaNet, Berlin.*

Government of South Africa. nd. Protection of Personal Information Act 4 of 2013. Government of South Africa. https://www.gov.za/documents/protection-personal-information-act.

Hirsch, D. 2013. "The glass house effect: Big data, the new oil, and the power of analogy". Maine Law Review., 66: 373.

Holstein, K., Jennifer Wortman Vaughan, J.W., Daumé III, H., Dudik, M. and Wallach, H. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the 2019 CHI conference on human factors in computing systems. 1–16.

Hooker, S., Nyalleng Moorosi, Gregory Clark, Samy Bengio and Emily Denton. 2020. "Characterizing bias in compressed models". *arXiv preprint arXiv,* 2010.03058.

Hooker, S. 2021. "Moving beyond algorithmic bias is a data problem". *Patterns*, 2, 4: 100241.

Hummel, P. Matthias Braun, Max Tretter, and Peter Dabrock. 2021. "Data sovereignty: A review". *Big Data and Society*, 8, 1: 2053951720982012.

Jiang, Z., Chiyuan Zhang, Kunal Talwar and Michael C. Mozer. 2020. "Characterizing structural regularities of labeled data in over-parameterized models". *arXiv preprint arXiv*, 2002.03206.

Jensen, K. 2012. CRISP-DM process diagram. https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png.

Kearns, M. and Roth, A. nd. Ethical algorithm design should guide technology regulation. https://www.brookings.edu/research/ethical-algorithm-design-should-guide-technology-regulation/.

Ledford, H. 2019. "Millions of black people affected by racial bias in health-care algorithms". *Nature*, 574: 7780, 608-610.

Mandl K.D. and Kohane, I.S. 2016. "Time for a patient-driven health information economy?" *New England Journal of Medicine*, 374: 3: 205–208.

Mbow, M., Lell, B., Simon P. Jochems, Badara Cisse, Souleymane Mboup, Benjamin G. Dewals, Assan Jaye, Alioune Dieye and Maria Yazdanbakhsh. 2020. "COVID-19 in Africa: Dampening the storm?" Science, 369, 6504: 624–626.

McBride, V., Ramasamy Venugopal, Munira Hoosain, Tawanda Chingozha and Kevin Govender. 2018. "The potential of astronomy for socio-economic development in Africa". *Nature Astronomy*, 2, 7: 511–514.

McCown, F. and Nelson, M.L. 2009. "What happens when Facebook is gone? In Proceedings of the 9th ACM/IEEE-CS joint conference on 609 Digital libraries, 251–254.

McCague, C. and Beer, L. 2021. "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans". *Nature Machine Intelligence*, 3, 3: 199–217.

Mehrabi, N., Fred Morstatter, Nripsuta Saxena, Kristina Lerman and Aram Galstyan. 2021. "A survey on bias and fairness in machine learning". *ACM Computing Surveys* (CSUR), 54, 6: 1–35.

Metcalf, J. and Crawford, K. 2016. "Where are human subjects in big data research? The emerging ethics divide". *Big Data and Society*, 3, 1: 2053951716650211.

Mitchell, M. Wu, S., Zaldivar, A., Barnes, P. Vasserman, L. Hutchinson, B., Spitzer, E. Raji, I.D. and Gebru, T. 2019. "Model cards for model reporting". In Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A. and Lum, K. 2018. "Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions". *arXiv preprint arXiv:* 1811.07867.

Moorosi, N. Mamello Thinyane, and Vukosi Marivate. 2017. "A critical and systemic consideration of data for sustainable development in Africa". In International Conference on Social Implications of Computers in Developing Countries. Springer, 232–241.

Mudongo,). 2021. Africa's expansion of AI surveillance-regional gaps and key trends. Policy Brief 2021, No. 3 . Research ICT Africa

Nyamwena, J. and Mondliwa, P. 2020. Policy Brief 3: Data governance matters: Lessons for South Africa. https://www.competition.org.za/ccred-blog-digital-industrial-policy/2020/7/28/data-governance-matters-lessons-for-south-africa.

Osakwe, S. and Adeniran, A.P. 2021. Strengthening data governance in Africa.

Osoba, O.A. and Welser, W. 2017. An intelligence in our image: The risks of bias and errors in artificial intelligence. Rand Corporation.

Peng, R.D. and Matsui, E. 2015. *The art of data science: A guide for anyone who works with data.* Skybrude Consulting, LLC.

Ponelis, S.R. and Holmner, M.A. 2015. ICT in Africa: Building a better life for all.

Purnell, N., Scheck, J. and Horwitz, J. 2021. Facebook employees flag drug cartels and human traffickers. The company's response is 597 weak, documents show. https://www.wsj.com/articles/facebook-drug-cartels-human-traffickers-response-is-weak-documents-11631812953.

Raji, I.D., Gebru, T., Mitchell, M. Buolamwini, J. Lee, J. and Denton, E. 2020a. Saving face: Investigating the ethical concerns of facial recognition auditing. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 145-151.

Raji, I.D., Smart, A. White, R.N., Mitchell, M. Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. 2020b. "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing". In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44.

Ray, E. L Nutcha Wattanachit, Jarad Niemi, Abdul Hannan Kanji, Katie House, Estee Y. Cramer, Johannes Bracher, Andrew Zheng, Teresa K Yamana and Xinyue Xiong. 2020. "Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the US". *MedRXiv*.

Rice, W.E. 1996. "Race, gender, redlining, and the discriminatory access to loans, credit, and insurance: An historical and empirical analysis of consumers who sued lenders and insurers in federal and state courts, 1950-1995". *San Diego Law Review,* 33: 583.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., ... & Schönlieb, C. B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3(3), 199-217.

Russell, S.J. and Norvig, P. 1995. Artificial intelligence: A modern approach. Pearson Education, Inc..

Sánchez-Monedero, J., Dencik, L. and Edwards, L. 2020. What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 458-468.

Seng Ah Lee, M. and Singh, J. 2021.Risk identification questionnaire for detecting unintended bias in the machine learning development lifecycle. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 704-714.

Seo Jo, E. and Gebru, T. 2020. "Lessons from archives: Strategies for collecting socio-cultural data in machine learning". In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 306–316.

Sey, A. and Mudongo, O. 2021. Case studies on AI skills capacity building and AI in workforce development in Africa.

Skorup, B. and Thierer, A. 2013. "Uncreative destruction: The misguided war on vertical integration in the information economy". Fed. Comm. LJ 65:, 157.

Shilton, K., Finn, M. and DuPont, Q. 2021. "Shaping ethical computing cultures". *Communication ACM*, 64, 11: 26-29.

Shuja, J., Alanazi, E., Alasmary, W. and Alashaikh, A. 2021. "COVID-19 open source data sets: A comprehensive survey". *Applied Intelligence*, 51, 3: 1296-1325.

Singh, J.P. and Flyverbom, M. 2016. "Representing participation in ICT4D projects". *Telecommunications Policy*, 40, 7: 692-703.

Sujan, M., Furniss, D., Grundy, K., Grundy, H., Nelson, D., Elliott, M., White, S., Habli, I. and Reynolds, N. 2019. "Human factors challenges for the safe use of artificial intelligence in patient care". *British Medical Journal of Health and Care Informatics*, 26: 1.

Tang, R. and Sae-Lim, W. 2016. "Data science programmes in US higher education: An exploratory content analysis of programme description curriculum structure, and course focus". *Education for Information*, 32, 3: 269-290.

Tongia, R. and Eswaran Subrahmanian, E. 2006. Information and Communications Technology for Development (ICT4D) - A design challenge? In 2006 International Conference on Information and Communication Technologies and Development. IEEE, 243–255.

Toyama, K. 2015. "Geek heresy: Rescuing social change from the cult of technology". *Public Affairs.*

Veale, M. and Binns, R. 2017. "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data". *Big Data and Society*, 4: 2, 2053951717743530.

Walsham, G. 2017. "ICT4D research: Reflections on history and future agenda". *Information Technology for Development*, 23, 1: 18-41.

West, D. and Allen, J. 2018. How artificial intelligence is transforming the world. Technical Report. Washington DC: Brookings Institution.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S.R., Richardson, R., Schultz, J. and Schwartz, O. 2018. *AI now report 2018*. New York: AI Now Institute at New York University.

Wirth, R. and Hipp, J. 2000. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Vol. 1. Springer-Verlag London, UK.

Wynants, L., Calster, B.V., Gary S. Collins, Richard D. Riley, Georg Heinze, Ewoud Schuit, Marc M.J. Bonten, Darren L. Dahly, Johanna A. Damen, and Thomas P.A. Debray. 2020. "Prediction models for diagnosis and prognosis of COVID-19". *Systematic Review and Critical Appraisal*, 369.

Zhang, J. and Barr, M. 2021. "Harmoniously denied: COVID-19 and the latent effects of censorship". *Surveillance and Society,* 19, 3: 389-402.

Zhang, Y. 2017. "The information economy". *Non-Equilibrium Social Science and Policy.* Springer, Cham, 149-158.

Zhang, B.H., Lemoine, B. and Mitchell, M. 2018. "Mitigating unwanted biases with adversarial learning". In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 335-340.

Mission

Renforcer les capacités des chercheurs locaux pour qu'ils soient en mesure de mener des recherches indépendantes et rigoureuses sur les problèmes auxquels est confrontée la gestion des économies d'Afrique subsaharienne. Cette mission repose sur deux prémisses fondamentales.

Le développement est plus susceptible de se produire quand il y a une gestion saine et soutenue de l'économie.

Une telle gestion est plus susceptible de se réaliser lorsqu'il existe une équipe active d'économistes experts basés sur place pour mener des recherches pertinentes pour les politiques.

www.aercafrica.org/fr

Pour en savoir plus :

www.facebook.com/aercafrica

www.instagram.com/aercafrica_official/

twitter.com/aercafrica

www.linkedin.com/school/aercafrica/

Contactez-nous :
Consortium pour la Recherche Économique en Afrique
African Economic Research Consortium
Consortium pour la Recherche Économique en Afrique
Middle East Bank Towers,
3rd Floor, Jakaya Kikwete Road
Nairobi 00200, Kenya
Tel: +254 (0) 20 273 4150
communications@aercafrica.org