# More than Just a Policy - Day to Day Effects of Data Governance on the Data Scientist

*Vukosi Marivate*

***Working Paper DG-006***

AFRICAN ECONOMIC RESEARCH CONSORTIUM
CONSORTIUM POUR LA RECHERCHE ÉCONOMIQUE EN AFRIQUE

***Bringing Rigour and Evidence to Economic Policy Making in Africa***

# More than Just a Policy – Day to Day Effects of Data Governance on the Data Scientist

By

Vukosi Marivate[1 2]
*Department of Computer Science,*
*University of Pretoria,*
*South Africa*

AERC Working Paper DG-006
African Economic Research Consortium, Nairobi
February 2023

**THIS RESEARCH STUDY** was supported by a grant from the African Economic Research Consortium. The findings, opinions and recommendations are those of the author, however, and do not necessarily reflect the views of the Consortium, its individual members or the AERC Secretariat.

Published by: The African Economic Research Consortium
P.O. Box 62882 - City Square
Nairobi 00200, Kenya

# Contents

# List of figures

# Abstract

Within a short space of time, the debate about Data Governance has fallen behind the realities of data-driven industries and economies. The flow and trade of data is driven by the needs of different stakeholders and evolution of global contexts of many technologies that are seen as local. To the Data Scientist, it may seem like an exciting time that has infinite possibility and opportunity to invent the near future. The gap between Data Governance on the African continent and data practice poses a challenge that must be dealt with sooner than later. This paper looks at the intersection of Data Science practice and Data Governance, I analyze some of the recent literature to identify areas of concern and focus. Ultimately, I look at how non-technical considerations are core in bridging data governance and data science practice, borrowing from other disciplines that had a head start with these challenges. Finally, we suggest steps that can be taken by practitioners to reduce this gap between governance and practice.



**Keywords:** *Data Science, Data Governance, Decision Making with Data, Data Ethics*

# 1. Introduction

The continued rise of the information economy meant an increase in the use of data to build and deploy many data-driven products. These data-driven products are used to extract meaningful insights from raw information, which is then used to address challenges across many different fields. This has coincided with the emergence and development of Data Science as a unique field of expertise, building data-driven products. Data Science is unique from Computer Science (the study of theory and practice of how computers work) and it encompasses many fields (discussed later). From the perspective of users, the data-driven products have brought many new services and conveniences.

In health, for example, there were rapid deployment of data tools to inform the public on the COVID-19 pandemic (Alamo et al., 2020; Shuja et al., 2021), pandemic prediction models (Evan et al., 2020) and estimations of impact of COVID-19 (Bradshaw et al., 2021). At the same time, some of the tools developed to deal with diagnostics/treatments were not as successful. An example of such data-driven products are the many tools/algorithms that were developed or deployed to improve radiology scans (Roberts et al., 2021; Wynants et al., 2020). One may be tempted to say that such deployments were a complete failure. However, these challenges highlight some of the shortcomings of data tools and areas of improvement. More importantly, these challenges outline the need to manage data (and its products) to consider the human factors and impacts data may have across all domains. Keeping with the COVID-19 topic, the pandemic also put a spotlight on the lack of basic data infrastructure (Mbow et al., 2020), lack of data skills and/or lack of political will in many countries to focus on the improvement of data-driven products. These data-driven products and tools ultimately impact on the quality of responses to the pandemic. The examples highlight the need for Data Governance that takes a refined view of data in different countries and organizations to illustrate these differences and govern the field accordingly. I believe this is a challenge that African countries need to focus on as they also develop their data journeys (Abebe et al., 2021).

I look at the Data Scientist (or Data Science Team) as the ones who make most of the decisions on the data tools they develop or create. This simplified view does not encapsulate all the challenges associated with what is currently taking place. It would be better to look at data-driven products through the lens of socio-technical systems. Socio-technical systems are systems that have interactions between

humans, machines and the environment (Baxter and Sommerville, 2011). Even within the organization, the Data Science Team or Data Scientist cannot make decisions without a variety of different stakeholders, especially decisions that have an impact on humans and other environmental factors. As such, the Data Scientist should be able to understand the other inter-dependencies of organizations and society to better understand where they fit, and that governance structures should exist to guide the development of systems with such inter-dependencies.

In this work, I aim to provide a better understanding of the governance/human factors that Data Scientists and organizations should be aware of. To address this challenge, I will answer fundamental research questions for the domain.

- *Research Question:* What are the salient points that Data Scientists should be aware of when it comes to Data Governance within organizations?

- *Research Sub-Question:* Do the current policies or mechanisms on the African continent provide a coherent view that can be used by Data Scientists to navigate and respond appropriately to the needs of the organization/society?

- *Research Sub-Question:* Can we learn from the ICT4D community to better understand how interventions should take care of more than just deploying a tool?

It is important to contextualize why we need to answer these questions. We are at a time where policy is lagging deployment of data tools (this is discussed in this paper). This means that there are gaps and blind spots that both Data Science practitioners and policy makers (both in public and private sectors) have. These blind spots have consequences. There has been much written about the data protection policy making and much written about Data Science practice and limitations. In this work, I link the two to have a joint understanding that decision making has to be done together.

The rest of the document is organized as follows. First, I look at the field of Data Science and how Data Governance fits into practice. The next step is to look at Data Governance on the African continent. I will set the scene and identify gaps that then intersect both areas of Data Science and Data Governance. In the proceeding section, I discuss how ICT4D may have already blazed a path that allows us to learn from in understanding the interactions of Data Science and Data Governance. The latter sections deal with the different stages of the Data Science process and proposals on how best Data Scientists can navigate human factors such as privacy, bias and security. Lastly, I conclude and summarize the viewpoints and evidence elaborated on in this paper.

# 2. Data science and practice

I first look at the practice of Data Science and its connections to Data Governance. As such, I provide an overview of what Data Science is - an important definition that is still evolving but is important for joint understanding between the reader and the author.

## What is data science?

Data Science is a discipline that has arisen due to a number of factors. Data Science is a field that uses scientific modelling techniques (typically from a diverse set of scientific disciplines) to extract patterns/information/knowledge from a wide variety of data (Dhar, 2013). The rise in this discipline has been swift for many reasons. Organizations (public and private) have been working to explore the data that they have amassed over time and mine information for patterns and trends that may give them a competitive advantage. There has been an explosion in the number of large Internet-based organizations and Internet-generated content. Simply, with more users on the Internet, and more content on the Internet, the information economy needs better data and data tools to monetize these users (Mandl and Kohane, 2016; Zhang and Barr, 2021), for example for advertising or for services that motivate users staying within a company' products (a walled garden) (Best, 2014; McCown and Nelson, 2009; Skorup and Thierer, 2013). On the side of public organizations, Data Science has meant the work to analyze or collect data that improves on services provided by governments or new forms of ways to understand citizens (sometimes resulting in mass/hyper surveillance. It is very important to understand these factors, especially as they are connected to "value creation in the information age" (Nyamwena and Mondliwa, 2020). The factors necessitate that we understand the foundational data infrastructures (physical, virtual, human and otherwise) through the lens of governance, specifically Data Governance. Let us first break down the process of Data Science.

## The data science process

To provide the reader with better understanding of Data Science, I use the data analysis cycles to provide an insight into the typical Data Science process. One can use the CRoss Industry Standard Process for Data Mining (CRISP-DM) as a representation of the process (Wirth and Hipp, 2000). The steps are typically:

- Understand a business problem
- Understand the data required
- Collect data
- Prepare data
- Perform modelling
- Evaluate the solution to the problem
- Adjust understanding and/or deploy (see Figure 1)

One notes that all of this focuses on solving a business challenge. We can easily extend this to solving any societal/organization/scientific challenge; it does not need to be business. This process is similar to the Epicycles of Analysis (Peng and Matsui, 2015) that splits the processes of the problem and the analysis for a solution to the problem. The former tries to separate the problem formulation from the modelling. Problem formulation takes understanding the correct data to gather or get access to. Ultimately, with all of these, we need to understand the human factors and dimensions that arise in all parts of the cycles. The inter-dependencies are discussed later in this paper.

The rise of Data Science has also coincided with the rise of Machine Learning (ML) and Artificial Intelligence - AI (West and Allen, 2018) and typically it is expected that Data Scientists understand, and can use, concepts from these fields (Tang and Sae-Lim, 2016). Machine Learning is a field of study concerned with creating tools that learn analytical models from data (Alpaydin, 2020) and is a subset of Artificial Intelligence. Artificial Intelligence is a field of study concerned with creating machines that mimic the intelligence of humans, typically defined as creating an agent that can perceive its environment and perform actions to maximize some utility or achieve some goal(s) (Russel and Novig, 1995).

**Figure1: CRISP-DM flow model**


Source: Jensen (2012)

Many Data Science researchers/practitioners are also Artificial Intelligence and/or Machine Learning practitioners/researchers. As such, from here on, I will refer to Data Science researchers/practitioners even if I am talking about Artificial Intelligence and/or Machine Learning. Many Data Science researchers or practitioners are comfortable with the above models of understanding data and the subsequent analysis. For this to be successful, society and organizations have an overgrowing need to understand what happens during developing and deploying a system or model in the real world. Governance, in more ways than one, comes into play. The data collection needs considerations of humans and the human dynamic (Bender and Friedman, 2018; Gebru et al., 2018; Seo Jo and Gebru, 2020). The choice of modelling requires consideration of people and their needs (Mitchell et al., 2019), the deployment further requires the consideration of the human dimension in all its guises (Raji et

al., 2020a; Raji et al., 2020b). As such, Data Governance can be a useful tool for the Data Scientist to be aware of these human factors and the challenges when humans and data (collection, modelling or products) interact (Buolamwini and Gebru, 2018; Hooker, 2021; Ledford, 2019; Mehrabi et al., 2021; Sujan et al., 2019).

## Why do we need data governance?

From the perspective of governments, as part of economic development and growth, they want to embrace "value creation in the information age" (Nyamwena and Mondliwa, 2020). To do so, the collection, use and flow of data has to be governed to be able to have oversight over this value creation. In short, Data Governance must touch every part of the Data Science life cycle as discussed earlier. Data Governance also rises to prominence because of historical pushes for digitization of countries, especially that of African countries. Governments are concerned that if they do not capitalize on the data opportunity, they will be left behind on another economic development. The challenge arises when we look at ways Data Governance must be shaped for different countries. Without adequate Data Governance in countries, the opportunities for both public and private sectors are at risk of not realizing the full potential of the information economy. This is a big risk as products that may fall short of the values of the countries' citizens may be deployed and ultimately cause harm. Such examples of falling short are inadequate privacy protections (Metcalf and Crawford, 2016), limitations on what data can be used for, regulation of data-driven products that could be harmful (Metcalf and Crawford, 2016), guidelines on data sovereignty (Hummel et al., 2021), and how specific sets of data should be treated as public goods to be shared within or outside a country Borgesius et al., 2015). Good Data Governance is not only about the data creation stage, but about how governance permeates the full Data Science cycle (Metcalf and Crawford, 2016). Furthermore, good Data Governance requires the contextual knowledge of and from decision-makers (in both public and public sector) to understand the Data Science cycle (data, modelling, algorithms, etc (Keans and Roth, nd). It is harder for the gatekeepers to regulate industry if they themselves do not have a foundational understanding of what typically happens within the Data Science cycle. This is an important point to highlight because industries such as finance, for example, have well-defined regulators in most countries. These financial regulators regulate the industry to mitigate corruption and harm. Regulatory boards are made up of experts in the field who then work to set best practice, limitations and penalties for breaches of the regulations. The challenges with many of the data-driven products we see nowadays is that many of the decision-makers in deploying these tools have little experience with the field itself, and see most of what is going on as a black box that takes in data, and "magically" produces answers. This highlights the need for basic foundational regulation that asks the right questions when developing data-driven products but also sets the path for a joint understanding of the field that should be understood by all people (not just experts). In the next section, I look at important parts of the Data Science cycle and highlight the human factors and questions that should be asked by Data Scientists and be understood by decision-makers.

# 3. Human factors and the data science cycle

To champion the joint understanding of Data Science and Data Governance, in this section I discuss the human factors in Data Acquisition, Modelling and Presentation phases of the Data Science cycle.

## Data acquisition

One of the steps that is fraught with tension in the Data Science process is the data acquisition process. This can be a blind spot (Mitchell et al., 2018; Zhang et al., 2018) that can make or break many projects. Imagine using a dataset collected in the 1950s on financial lending by banks. Now, building a predictive tool to assist in lending decisions with such a dataset will be full of gender and racial biases in many countries (Bond and Tait, 1997; Rice, 1996). Put simply, the model would learn to discriminate. This is still a challenge today (Fu et al., 2021). Even if the data is taken as representative of the population being studied, it may encode societal bias and discrimination. Most times, when interacting with decision-makers or clients, those without much experience tend to overlook the challenges in the acquisition of data. These challenges relate to governance issues (Veale and Bins, 2017).

### *Processes and procedures*

In acquiring data, as part of the Data Science process, one connects the problem being approached with the data that will be needed to solve the problem. At some point, there may be data before the questions are clear, while at other times there is a question to be answered but the data has not been mapped out. In all instances, data must move from where it rests and staged for processing by the Data Science team. This requires identification of the relevant data source, identification of which subset of the information is important and how the transmission will occur. In carrying out these identification steps, we have to look at the human factors.

### *Human factors*

For each of the proceeding steps of the Data Science process, I focus on three human factors. For the Data acquisition I focus on: Where does the data come from? Why is/

was it being collected? Who is the data about? There are many more factors, but for conciseness and to communicate our message, the message will remain with three factors per step of the Data Science cycle.

Where does the data come from? When identifying the source of data, it quickly becomes clear that one has to understand the structures of the organizations internally or externally that control access and use of the data. In an ideal case, there is a clear Data Governance structure that also provides information on how a Data Scientist can request data, how the data should be handled and any sensitive and salient information that the scientist should be aware of (Abraham et al., 2019). There will be questions that are related to the sensitiveness of the data. Was the data collected in an ethical manner? Is the data part of an open data repository? What licensing is the data under and expectations of use? Is the data from a governmental entity? what are the national expectations on Open Government data? For example, in a municipality, one may expect that aggregated water use data by municipal ward should be open and available (especially as many areas in some countries face water shortages), but there may be some resistance by some officials in making this data available. It may be that there is not enough human resource to create and keep the data available, the data may normally be available for a fee that adds to revenue; there may be issues of transparency, etc.

Why is/was it being collected? This is an important factor as it establishes prior expectations on what the data that was collected or is being collected was used for. If we imagine that we have data about the transaction habits of bus riders in a city, the original use of the data and expectation was to manage the transportation system. If now the data will be used to understand behaviour to deliver advertising to bus riders, this new use may not be covered by original terms of reference. More importantly, bus riders may not agree with the change of the use of their data, and there is a responsibility the organization has with them to treat their information with care and thought.

Who is the data about? In carrying through the process to build up the data, one must think if it is representative of the population it is serving. Again, when the data is about people, we need to understand who the data represents and if this distribution is equitable, fair (Mitchell et al., 2018; Zhang et al., 2018) Further, does this distribution of people match those we expect to make decisions about in the end data-driven product? If not, this may be a problem that introduces biased decision-making. For example, in the recent decade, much has been highlighted about the bias in facial recognition systems (Raji et al., 2020). Some of this bias comes from the original data that was used to train them (Mitchell et al., 2018; Zhang et al., 2018). Some of this bias comes from the designs of the systems and how success is measured. I will discuss more on this later in the modelling and the presentation sub-sections.

One can see just from looking at the above that there are important human factors that cannot just be left to the Data Scientist or organization to make decisions about. There needs to be foundational expectations on data handling, data storage, security, ethics and regulatory tests on what the data would be used for.

## Data analysis and modelling

In the Data Analysis and Modelling step, the Data Scientist focuses energy on using the correct approaches to extract meaningful information from the data. These choices will influence the result and be the foundation on which many will choose to believe the results or not. Even though these may be established computational, statistical or mathematical approaches, we still need to understand how choices impact the end product and people.

### *Processes and procedures*

The Data Scientist takes the data that has been acquired in the prior step. They then work to clean it, transforming it into a form that can be used by downstream modelling tasks and then loading it into their modelling systems. The Data Scientist will make choices on metrics to be measured or optimized. Ultimately, these metrics are used to decide on success and are then used to know if new data should be sourced, the question should be re-framed or can one move to the next step of the Data Science cycle.

### *Human factors*

For the Data Analysis and Modelling stages, I focus on these factors: How are the modelling choices made? Who has the skills to model? What are the models for the use-case being used? How are the modelling choices made? For a period, there was a popular retort that people are biased and machines are unbiased. A number of works have highlighted that machines cannot be unbiased as the data that they use to learn from may be biased (Birhane and Cummins, 2019).. After this, the needle moved to that algorithms cannot be biased, only the data (Birhane and Cummins, 2019). But this still ignores many factors that modelling choices also impact the results of the final models (Jiang et al., 2020). In Machine Learning, we pride ourselves in working to build better and better generalize-able, accurate and efficient algorithms, but this does not absolve us about thinking about our modelling choices (Birhane et al., 2021). Work by Hooker et al. (2020) highlighted the biases in compressed models.

Further, more and more Machine Learning models use transfer learning (building on prior models or datasets). This then carries forward biases. This is one of the reasons Data Scientists should work to document their modelling choices (Mitchell et al., 2019). Modelling may seem insignificant at the time of decision-making but may lead to big consequences later. A recent example (Birhane et al., 2021) is how models influence the collection of massive datasets (to fight against bias) that when looked at under a microscope, are not as representative as the dataset authors claimed. This highlights the lack of participation and inclusive design choices that also call into question who has the modelling skills?

Who has the skills to model? ML/AI/Data Science is a field that is typically skewed in terms of demographics and who ends up building the underlying technologies. One may argue that this does not apply on the African continent when it comes to racial make-up. But that is not a true reflection of the field. For a long period, in major technology companies on the continent, the senior technical roles were skewed between male and white (mirroring the challenges that have been criticized about Silicon Valley). Further making this worse is the lack of Data Science skills on the continent. Without these skills, we further have less connection between decision-makers and those who design models. How many of the decision-makers have a data/computational background? Another factor is that the major tech companies that drive most of the Internet economy tend to only have business offices on the continent (Birhane, 2020). Their aim is to sell their services (Birhane, 2020), extract data (Coleman, 2018 and handle regulatory issues - if there is regulation (Birhane, 2020; Coleman, 2018). The offices do not build or shape the core technologies at these companies. As such, if we connect this question to the prior one, we see how modelling choices can become a life changing decision for those on the downstream tasks. Imagine how in organizations, automated hiring systems, were deployed to assist in the hiring process by using AI to screen or monitor candidates. These systems have been shown to be discriminatory (Sánchez-Monedero, 2020), but what are the odds that the decision-makers and internal Data Science teams had the skills to be able to evaluate their facial recognition systems or text screening services against bias?

What are the models for the use-case being used?: Recent work in the ML/AI field has brought into focus explainable models in the fight against harm and pursuit for better fairness. Let us take, for example, the increase in surveillance systems and facial recognition systems internationally. How the models are chosen and evaluated for such use-cases affect the ultimate impact these systems will have on society. Much work has highlighted how biased facial recognition systems (Raji et al., 2020) can lead to discriminatory behaviour by law enforcement. This may end up being a life or death situation for someone at the end of these automated systems.

A Data Scientist and decision-maker needs to ask themselves, what is the cost of an error of our model? This should then impact how the deployment is done. Further, there may be regulatory restrictions in making one choice or another, depending on the societal expectations.

## Presentation and deployment of data-driven products

The final step in many Data Science projects is presenting results to decision-makers and/or the deployment of the data-driven products.

### *Processes and procedures.*

In this step, the Data Scientist would work to present a report on findings of the modelling to answer the original questions. From here, decisions may be made on these reports. Reports may be visualization, simulations or data-driven products with

metrics that show their efficacy. Decisions on what to show and who the data-driven products will be aimed at will be made. These have human factors.

## *Human factors*

For the Presentation and Deployment of data-driven products stages, I focus on these factors: What decisions are being made with the models? What choices are being made in what to be shown? How will the models be kept updated?

What decisions are being made with the models? The ultimate test for the usefulness of a model for the decision-maker is when it is deployed for use or presented for decision-making. This is a spot in the Data Science life cycle that requires careful understanding of the prior parts of the cycle, or wrong decisions could be made. When looking at the data product or predictions of a model, the user must understand how the model works, how it was built and what limitations it has. The sub-question here could be: how do people interpret the results/predictions from the data product? This requires more than just displaying a result but also working with human computer interaction practitioners to design the model in such a way that is fair, transparent and mitigates bias or discrimination (Holstein, 2019; Seng Ah Lee and Singh, 2021).

What choices are being made in what to be shown? As in the statistical domain, we can also lie with data-driven products. The COVID-19 pandemic had many examples where decision-makers worked to distort data, distort model predictions and even censor data researchers and practitioners to fit with a view that the decision-maker held (Abazi, 2020; Zhang and Barr, 2021). This may be taken as an extreme public example, but this does happen in many ways. One may be testing for harm at run-time.

How will the models be kept updated? When deploying data-driven products, the internal models must be kept updated. The world did not stop changing when the model was trained and deployed. As such, the models will start exhibiting drift. This drift may also come from how users respond to what the model does. Does the organization of Data Science team have procedures on the maintenance of the models in the data-driven product, and how do we test for drift before the system has high error in its results (predictive, prescriptive, diagnostic, etc)?

In this section, I have discussed how Data Science and Data Governance intersect. In the latter part of the section, I chose three sections of the Data Science cycles to be able to analyze for human factors. By identifying these human factors, we can better understand how Data Governance is an integral part of the full cycle as decisions being made by the scientist will impact users and humans in general. In the next section, I then discuss Data Governance on the African continent.

# 4. Data governance and the African continent

With calls for African countries to jump on to the current advances of data driven economies, there has been some movements towards strategies and governance policies by governments that cover data. The African Union released "The Digital Transformation Strategy for Africa 2020-2030" (Africa Union, 2020). This strategy should be understood in the context of the wider and more localized Data Governance and digitization challenges in different African countries.

When it comes to privacy, the European general data protection regulation (GDPR) (European Commission, nd) has had wide ranging effect and impact on the internet economy as many companies who processed European citizen data had to abide by the rules set out by the EU. Around the African continent, as shown by the research by Davis (2021), there are efforts to strengthen data protection policies, even with only about 52% of African countries having such legislation.

The African Union Convention on Cyber Security and Personal Data Protection (known as the Malabo Convention) (Africa Union, 2014) was adopted by AU member States in 2014. It sets out to provide protection for cyber infrastructure, protection of personal information, cyber security and the necessary foundations to enable an information economy across the African continent. Even though ratified in 2014, only 8 countries had ratified the convention by 18th June 2020 (European Commission, nd). The convention touches on many aspects that can form a unified foundation for African countries to benefit from the information economy. Without ratification, we have the reality that organizations and practitioners do not have a unified view on how to deploy data tools and, for some countries, the reality is much worse with very lax or non-existent protections (Davis, 2021).

In South Africa, the Protection of Personal Information Act (POPIA) (Government of South Africa, nd), which has taken many years to get enacted, has also begun a discussion in the public on data acquisition, protection of personal information and the use of data for downstream tasks (especially when it is not for the original purpose of data collection). Even so, Data Governance is not only the protection of personal information; there are many more human and organizational factors that data interacts with. I hope the preceding section has made it clear that Data Governance should cover more than just the data being used.

But, as earlier discussed, there are many human factors that should be taken into consideration in all the stages of the Data Science cycle. To effectively govern

the full process, countries need to have a clear understanding of the stages and the responsibilities of governments towards the Data Scientists and the responsibilities of the Data Scientists towards the public.

The African continent has made big strides in the ICT sector and building local skills and championing local companies (Ponelis and Holmner, 2015). Even so, there is still a dominance of the Big Tech Giants (Microsoft, IBM, Google, Facebook, etc) on the continent physically or with services that cross borders. Even though we do not have an agreed definition of the data skills gap, the work by Sey and Mudongo (2021) highlights how there is lack of understanding of the need for AI skills and that we need to have efforts to build these skills on the continent, and this must connect public and private sectors. These insights are important as they place in context how few of the Big Tech firms have few or any research and development in the continent. AI governance skills are recommended as part of the development of AI skills on the continent (Sey and Mudongo, 2021), echoing the message in this paper on the broader Data Science and Data Governance nexus.

The continent risks being just a source of data (Birhane, 2020) to build services that then are used by citizens without any local development of these services. This has been recently brought to bear with how Facebook only has 13% of its abuse team (which fights abuse on their online platforms) working on non-US content, even though 90% of Facebook users are outside the US (Purnell et al., 2021). This is important as misinformation on Facebook outside the US has effect on many countries, but cannot be battled by Facebook itself. Further on, governments have to be able to govern the digital space and ensure that citizens get to benefit from the digital public goods (Gillwald and van der Spuy, 2019).

Another challenge is the use of some of the data-driven products for surveillance by both governments and private sector on the continent (Mudongo, 2021). As already highlighted, the systems are less likely to be developed locally and may encode biases and lead to discrimination. This illustrates another governance gap (whether planned or unplanned) as decision-makers must be able to evaluate the risks and harms such systems may pose to the population (Mudongo, 2021).

# 5. Case Study: Learning from our recent past, enter ICT4D

Data Science and Artificial Intelligence have been hailed as a silver bullet to many problems; data itself is referred to as the new oil to be exploited by nations and organizations (Hirsch, 2013). But a challenge that organizations and nations should be able to spot rears its head again. With the rise of ICT and digitization efforts, many problems were pointed to where ICT could be the solution (Curtis, 2019). Throw in development practices, ICT4D has been a force for the last two or more decades (Walsham, 2017).

I argue that we now have had enough time that some of the shortcomings of seeing many problems as requiring ICT as the solution, especially from practitioners who would come from outside, drop in, deploy and then leave is very much akin to what is happening in the Data Science world currently and needs change (Shilton et al., 2021). There may be differences, chief among them, familiarity with what ICT is and less familiar with what Data Science, Artificial Intelligence or Machine Learning are (Osoba and Welser, 2017). Basically, Data Science researchers and practitioners are just seen as magicians you throw a problem and data at, and a solution arrives on the other side. We see this with the advent of touting of 4IR strategies for African nations that are driven by public institutions that do not have the skills or knowledge to really engage with the subject they are touting as a solution to many of the problems they face (McBride et al., 2018; Moorosi et al., 2017).

In ICT4D, a historical debate was on the efficacy of having researchers and practitioners who were not locals come up with "solutions" using ICT to many development issues (Andrade and Urquhart, 2012). Over time, this has become an area of study within the field itself. It became very apparent on how the development and design of systems should be participatory (Andrade and Urquhart, 2012; Tongia and Subramanian, 2006; Toyama, 2015) and consider more than just the technical challenge. This tough took time and many failures. In contrast, within Data Science and Artificial Intelligence field, a lot of work has been put into understanding fairness, ethics and the longer-term effects of the technical interventions. This is a welcome change to the ICT4D history, but we still are lagging in the understanding of the need for participatory design and governance that guides the field (Singh and Flyverbom, 2016). We have large international bodies such as the International Telecommunications Union, that many States belong to, that have shaped ICT policies across regions.

In Artificial Intelligence, one can say the debate on fairness and harm has been very much open due to the threats of wide scale impact on people. But this does not mean that debates solve the problems. In most of the debate and discussion, it is mostly researchers and not decision and policy-makers who are doing work to document harm and make recommendations to mitigate it (Whittaker et al., 2018). Policy makers need too come to the table to also shape the debate by providing input from government. We need to draw from lessons of other fields while at the same time understanding the uniqueness of the take up of data-driven products before we even think about their impact.

# 6. Conclusion

In this paper, I used a survey of literature around Data Science and Data Governance to bring to the fore the connections within this nexus. Leaving decisions of design to only the Data Scientist ignores the many human factors that data-driven products have. As such, Data Governance is key to being able to create and deploy products that add to the developing economies on the continent while mitigating harm. This requires that African countries have an appreciation of the needs of governance and skills to enable effective policy. The case study presented on ICT4D allows us to learn from a related discipline that has been active for two decades and has had similar challenges in deploying interventions in the Global South.

## Recommendations

- There is need for African governments to work together to practically implement Data Governance policy. The glaring reality that only 8 countries (as of this writing) have ratified the African Union Convention on CyberSecurity and Personal Data leaves much to be desired.

- Both public and private industries must engage with data scientists to get a better understanding of the areas of concern highlighted in this paper beyond data privacy. Most policy on the continent focuses on privacy protections and some automated decision-making, but there are many other decisions made in the process of developing data tools that impact the outcome.

- For the Data Scientist, it must be a reality that policy and development of data tools go hand in hand. Even if national, regional or continental policies have not caught up, there is growing movement within our practice that works to develop best practice and also highlight challenges in ethics, fairness and mitigating abuse.

# Notes

1. Author's address: Vukosi Marivate, vukosi.marivate@cs.up.ac.za, Department of Computer Science, University of Pretoria, Pretoria, South Africa. 

2. I wish to acknowledge the AI/ML/DS Grassroots organizations across the African continent and the diaspora that have worked to shape how we participate and shape the technologies in question. I would also like to thank the Data Science for Social Impact Research Group at the University of Pretoria that has allowed me to explore these topics with them. Finally, I would like to acknowledge ABSA who sponsor the UP ABSA Data Science Chair.

# References


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

To strengthen local capacity for conducting independent, rigorous inquiry into the problems facing the management of economies in sub-Saharan Africa.

The mission rests on two basic premises: that development is more likely to occur where there is sustained sound management of the economy, and that such management is more likely to happen where there is an active, well-informed group of locally based professional economists to conduct policy-relevant research.

www.aercafrica.org

## Learn More

www.facebook.com/aercafrica

www.instagram.com/aercafrica_official/

twitter.com/aercafrica

www.linkedin.com/school/aercafrica/

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