The Role of Government as a Provider of Data for Artificial Intelligence

Phase 1 Full Report

May 2024

GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

This report was developed by Experts and Specialists involved in the Global Partnership on Artificial Intelligence's project on 'The Role of Government as a Provider of Data for Artificial Intelligence'. The report reflects the personal opinions of the GPAI Experts and External Experts involved and does not necessarily reflect the views of the Experts' organisations, GPAI, or GPAI Members. GPAI is a separate entity from the OECD and accordingly, the opinions expressed and arguments employed therein do not reflect the views of the OECD or its Members.

Acknowledgements

This report was developed in the context of the 'The Role of Government as a Provider of Data for Artificial Intelligence, with the steering of the Project Co-Leads and the guidance of the Project Advisory Group, supported by the GPAI Data Governance Expert Working Group. The GPAI Data Governance Expert Working Group agreed to declassify this report and make it publicly available.

Co-Leads:
Ching-Yi Liu*, National Taiwan University
Jhalak Kakkar*, Centre for Communication Governance

The report was written by: **Rachel Adams**‡, Research ICT Africa; **Fola Adeleke**‡, African Observatory for Responsible AI; **Jackie Akello**‡, Research ICT Africa, **Silvana Fumega**‡, Global Data Barometer; **Tai-Jan Huang**†, Information Law Center, Institute Iurisprudentiae, Academia Sinica, Taiwan AI Center of Excellence; **Claude K. Migisha**‡, Research ICT Africa; **Moremi Nkosi**‡, Research ICT Africa, **Gabriella Razzano**‡, OpenUp and **Chuan-Feng Wu**†, Information Law Center, Institute Iurisprudentiae, Academia Sinica, Taiwan AI Center of Excellence.

GPAI would like to acknowledge the tireless efforts of colleagues at the International Centre of Expertise in Montréal on Artificial Intelligence (CEIMIA) and GPAI's Data Governance Expert Working Group. We are grateful, in particular, for the support of **Stephanie King** and **Stefan Janusz** from CEIMIA, and for the dedication of the Working Group Co-Chairs **Shameek Kundu and Bertrand Monthubert,** and former Co-Chairs **Jeni Tennison*** and Maja Bogataj **Jančič*.** The report also benefitted from the detailed expert review of Working Group members, including in particular **Emmanuel Vincent** and **Anurag Agrawal**.

* Expert
** Observer
† Invited Specialist
‡ Contracted Parties by the CofEs to contribute to projects

Citation

GPAI 2024. The Role of Government as a Provider of Data for Artificial Intelligence: Phase 1 Full Report, May 2024, Global Partnership on AI.

Table of Contents

List of Figures

List of Tables

List of Acronyms

AfCFTA	Africa Continental Free Trade Area
AI	Artificial Intelligence
API	Application Programming Interface
ATD	Automated Decision-making
CART	Classification And Regression Trees
CEIMIA	International Centre of Expertise in Montreal on Artificial Intelligence
CIAT	International Centre for Tropical Agriculture
COVID-19	Coronavirus Disease 2019
CSO	Civil Society Organisations
DHS	Demographic and Health Survey
DPA	Data Protection Authority
GDB	Global Data Barometer
GDPR	General Data Protection Regulation
GPAI	Global Partnership on Artificial Intelligence
GPU	Graphics Processing Unit
HRA	Health Research Authority
HWD	Health and Welfare Data
HWDSC	Health and Welfare Data Science Centre
IBM	International Business Machines Corporation
IDEAM	National Institute of Hydrology, Meteorology and Environmental Studies
ISA	Information Sharing Agreement
IP	Internet Protocol
LMICs	Low and Middle Income Countries
MHPRA	Medicines and Healthcare Products Regulatory Agency
MLDS	Maryland Longitudinal Data System
MOHW	Ministry of Health and Welfare
MOU	Memorandum of Understanding
NASSCO	National Social Safety Net Coordinating Office
NASSP	National Social Safety Nets Project
NCC	Nigeria Communications Commission
NDPR	National Data Protection Regulation
NHIA	National Health Insurance Administration
NHIRD	National Health Information Research Database

NHS	National Health Service
NITDA	National Information Technology Development Agency
OECD	Organisation for Economic Cooperation and Development
TCP	Transmission Control Protocol
PDPA	Personal Data Protection Act
RIA	Research ICT Africa
RRR	Rapid Response Register
SDK	Software Development Kit
SFTP	Secure File Transfer Protocol
UK	United Kingdom
USA	United States of America

Executive Summary

Access to government data has the potential to be a catalyst for AI development. To achieve this, a multi-faceted approach that enhances data-sharing is necessary. Progressive models for data sharing are central to the promotion of responsible AI including equitable access and transparency in use of government data. As governments embrace the use of AI as the case studies in this report show, positive use cases are needed to further encourage the notion of AI for public good.

Various data sharing models are emerging to enable the supply of data by governments to AI developers. These models include contracts, open data, data stewardships, and public-private partnerships. Key issues related to these models for data sharing include the legal basis for the sharing of data by governments, compliance with data processing principles such as purpose limitation, and data minimisation in disclosure.

Four case studies are explored in this report to understand current models of data sharing by the government with the private sector. We assess the objectives of data sharing by governments and the mechanisms for data sharing with the aim of identifying benefits, risks and cross-cutting recommendations that can advance the goal of data sharing by governments for AI development.

In the UK in late 2015, the National Health Service (NHS) and DeepMind entered into a data-sharing agreement for DeepMind to develop an app to quicken the diagnosis of acute kidney injuries. Through an information sharing agreement, the NHS released the personal health information of 1.6 million patients to DeepMind. The parties did not follow data governance safeguards in place and the public objective of improving diagnosis faced a public backlash and sanctions from oversight institutions when the project became public knowledge. Some of the issues arising from the case study include the legal basis for NHS to share data with DeepMind, data minimization safeguards, mechanisms in place for data subjects to exercise their rights and transparency in AI design, testing and use. This case study shows the need for evolved understanding of data governance and oversight institutions with strong enforcement powers.[1]

In Taiwan, the government introduced an app, the Health Passbook, launched in 2014 but updated to take advantage of real name SIM registration in Taiwan in 2018, where users of the app can select which third party providers can access their personal health information to provide customised healthcare services. The government does not directly share data with third parties in this case but serves as a facilitator by verifying third parties in line with data governance principles before they become eligible to receive data directly from data subjects. This case study provides an alternative for governments who are risk-averse to play a different role in how data is shared for the development of AI.

In Nigeria in 2021, the government implemented a new social protection program during the height of the COVID-19 pandemic to make cash transfers to urban poor Nigerians who were affected by the devastating consequences of the pandemic. To identify the number of beneficiaries who could be eligible under the new programme, the Nigerian government contracted a company to develop a

[1] The NHS subsequently entered into further data-sharing agreements with Deepmind, for which both parties acted with considerably greater attention to patient consent and medical and data ethics – for example in the case of Moorfields Eye Hospital, and with University College London Hospital, among others.

solution that used algorithmic decision-making to identify eligible beneficiaries. The government transferred its social protection database to the company and transferred its data to telecommunications companies who further helped in the identification of other eligible beneficiaries. This case study raises several issues including data privacy, implications of automated decision-making and algorithmic transparency.

In Colombia, the government in partnership with an international NGO and farmers' collectives provided data for the development of an app developed since 2014 that guided farmers on key decision-making to improve their crop yields. The government entered a tripartite alliance where data was collectively supplied by all parties to develop a chatbot that assisted farmers in their decision-making. This case study is a very useful demonstration of governments sharing non-personal data for responsible AI development. This approach is a great first step for many governments that are reluctant to share any data to start with sharing of non-personal data in pilot programmes for testing their data-sharing mechanisms. This case study is also an example of the need for flexible policy design of various data governance frameworks depending on the kind of data involved.

The case studies in this report raised several issues and our recommendations address different themes including:

- building public trust in AI
- the need for data collaboration
- algorithmic decision-making and the need for human oversight.
- tackling digital inequalities
- data and AI justice
- regulatory certainty and efficient redress mechanisms
- robust public procurement process for AI development
- transparency and accountability and
- the role of AI in advancing a development agenda.

Through our case study analysis, we identified key enablers in government data sharing with a focus on the appropriate regulatory environment, policy landscape as well as infrastructure development. The key legal areas identified are in antitrust, cross border data flows, intellectual property, data protection and access to information.

We also identified approaches to mitigating barriers facing governments as providers of data for AI, by evaluating which technologies can facilitate data sharing; by addressing barriers formed by internal organisational culture; by assessing public attitudes towards the use of public sector data in developing AI; and by identifying fair financial models in the development of AI – taking into account public financing and benefits for data subjects.

Governments play a crucial role as regulators and facilitators of enabling environments for the AI industry to thrive. Many are adopting AI capabilities as they hold the potential to vastly improve government operations and help meet the needs of citizens in new ways ranging from healthcare delivery to agriculture and social protection, as this report shows. Accordingly, we have developed eight principles to guide governments on when and how to provide data for AI development. As governments leverage the power of AI to improve their operations, they must consider that advancement in AI comes with reliance on bulk data access by AI developers. How they make this data available responsibly is central for the future of equitable and just AI.

Four key principles were identified as crucial for the establishment of an enabling environment for government data sharing – in other words, **when** data should be shared:

1. Data sharing for public benefit
2. Accountability mechanisms in the AI systems being developed
3. Robust data subject participation in their data sharing
4. Data-sharing models that embrace data equity and data justice

Four additional principles were identified on **how** governments should share data:

5. Transparency
6. Safety and security
7. Data minimisation and proportionality
8. Human rights protection and privacy

1. Introduction

Governments are important collectors, collators, and producers of data. Governments hold data that is nationally or sub-nationally representative. They also hold data that provides insights into important social issues and dynamics, such as school attendance, social protection use, crime, and the functioning of healthcare systems. Government data is generated through the provision of government services such as civil registration (i.e. issuing of IDs, birth and death certificates, etc.), healthcare, education, registration of businesses, policing services, research carried out by governments, and national statistics exercises such as national censuses.[2] This kind of data can be an important foundation for developing AI tools that address social challenges and developmental priorities, including efficiency of government service provision, or gaps in access to education, healthcare, or sanitation, for example.

The provision of government data to AI developers must be undertaken responsibly, considering various foundational principles such as respect for human rights, privacy, consent, inclusivity and ethical use. This requires several measures that may be adapted in relation to different country contexts and needs. These include mechanisms to ensure that sensitive data, including personal data, is legally and safely shared:

- by adhering to data standards that support interoperability, ensuring data shared is in a format that is structured, discoverable, reusable and accessible;
- via public engagement, participation or awareness programmes to ensure public buy-in for the provision of government data to AI developers;
- through transparency mechanisms to enable accountability and build public trust;
- by undertaking an impact assessment or similar risk-mitigation measures to prevent against the risk of harm, particularly for underserved or vulnerable communities.

In addition, there is also a set of foundational requirements critical to the provision of government data to AI developers. These include the digitalisation of record-keeping, and protocols for how government-held data is collected, managed and stored, such that it can be useful for other parties and in future. Other measures relate to broader legal regimes which provide for access to information or transparency in government activities and AI, public procurement processes that support responsible innovation, and antitrust laws which regulate market risks, including big tech/data monopolies. Lastly, governments are responsible for regulating the responsible, fair and equitable access and use of data, for encouraging responsible innovation, and for leveraging strategic advantage through use and application of data, including for AI.

In June 2023, Research ICT Africa (RIA) was contracted to undertake a study for the Global Partnership on AI (GPAI) on the role of government as a provider of data for AI. This report sets out the key findings of the project, through the iterative approach to the identification, selection and review of four case studies, the extraction of key lessons regarding experiences of government sharing data with AI developers in three different sectors (i.e. health, climate change/agriculture and social protection) and contexts (Africa, Europe and Latin America). The methodological approach for the project included the following key phases:

[2]C. van Ooijen, B. Ubaldi and B. Welby (2019), "A data-driven public sector: Enabling the strategic use of data for productive, inclusive and trustworthy governance", *OECD Working Papers on Public Governance*, No. 33, OECD Publishing, Paris, https://doi.org/10.1787/09ab162c-en.

1) Deepening and refining the conceptual idea and understanding of what is meant by "government as a provider of data for AI";
2) Identification of four case studies, analysis and synthesis of key lessons in different sectors and regions;
3) In-depth analysis of the role of government as a provider of data for AI in the selected four case studies;
4) Identification and exploration of key themes associated with governments sharing data for AI;
5) Structured consultative sessions with AI experts to elicit their insights and guidance on the key considerations for government as a provider of data for AI; and
6) Conducting a crosscutting review of the principles that various governments (or regions) apply in relation to the sharing of data for AI, and using these to distil a succinct set of principles that can be applied more uniformly with regards to when and how governments share data for AI development.

It is important that governments create frameworks that enable them to share data for AI development in a responsible – that is, ethical and legal – manner. This project aims to support governments in making decisions about whether and how to share data they steward with AI developers, with the intention of increasing the availability, accessibility, and use of publicly held data for AI that is grounded in the principles of human rights, inclusion, diversity, innovation and economic growth. The project aims to assist governments to prioritise their efforts and reduce their concerns about the risks of sharing public data for AI.

Our objective from this stage of the research was to gain a deeper understanding of the global landscape concerning governments sharing data with AI developers. We were interested in the following elements:

- understanding more about how governments are currently sharing their data, and for what purpose;
- what risks are being acknowledged and what measures are being taken to mitigate them; is there evidence of harms having occurred as governments share data with AI developers?;
- what challenges and barriers exist that inhibit government data sharing; what legal frameworks currently regulate this?;
- the identification of key enablers as well as mitigation of barriers to access government data; and
- what principles governments can apply when sharing the data they steward with AI developers.

The research and compilation of the case studies were informed by GPAI and CEIMIA's previous work on the topic of Data for AI development. Particularly, the notions of responsible data stewardship, and data rights and justice, were integral to the analyses of the case studies throughout the report.

It is important to note that the concept of government as a provider of data for AI is not readily understood. We need to strengthen the conceptual framing of this phenomenon, and the narrative around why this is important, to advance AI solutions that support the realisation of developmental and public interest priorities. Additionally, there is a paucity of publicly available and accessible

details and information on existing/potential illustrative case studies related to governments sharing data with AI developers. It may be because of limited public trust in governments sharing data with AI developers, so where this occurs outside of formal public procurement processes for AI developers, there is limited publicly available information. This was the case in all regions we were examining. However, we can broadly determine that this was a particular concern in LMICs, except for a limited number of countries that are leading through regulatory sandbox initiatives. This is due to the objectives of these sandbox initiatives in developing best practice and implementation guidelines in pilot programmes that can be replicated by other countries which facilitates information sharing.

The investigation we conducted identified key findings across a number of different areas relevant to governments sharing data for AI development:

1. **Capacity building:** Governments currently have limited capacity to adequately assess, approve, share and review data that they hold with AI developers, including institutional oversight capacity. This is particularly so in LMIC contexts.

2. **Data-sharing models:** Government Open Data portals remain a critical venue through which to share government-held data fairly and transparently for AI development, but more needs to be done to ensure these platforms are populated, updated, used and accessible**.** In addition, new models for data sharing to balance the power between data controllers/processors and data subjects, including data commons, are emerging around the world. Further research is needed to better understand their risks and opportunities, and ensure they are responsibly regulated.

3. **Data Subject rights:** The empowerment of data subjects in new arrangements for data sharing between governments and AI developers stands out as an exemplar. However, they are not without costs and fundamental requirements, such as awareness and understanding about data rights from citizens.

4. **Public Trust:** Ensuring public trust in AI and government adoption of AI is a critical first step in the sustainability of fair partnerships between governments and AI developers.

This report is structured as follows:

Section 1 sets the scene for the report, the context and provides a snapshot of the main findings.

Section 2 discusses the concept of government as a provider of data for AI development.

Section 3 provides the rationale and objectives for the case study analysis, followed by close analysis of four case studies:

a) Google DeepMind – NHS Royal Free Foundation Trust (UK)
b) The Health Passbook (Taiwan)
c) The Rapid Response Register for Covid-19 Cash Transfers (Nigeria)
d) Aclimate (Colombia)

Section 4 synthesises the case study learnings and provides thematic recommendations.

Section 5 provides details on the key enablers for government provision of data for AI development purposes.

Section 6 identifies five key strategies/interventions that can be applied to overcome barriers to accessing government data for AI development.

Section 7 goes on to outline the recommended principles to guide governments in responsibly sharing their data with AI developers, followed by an overarching **conclusion**.

Defining the Concept

Established mechanisms for governments to provide data

The provision of government-held data for AI refers to the ways in which governments share data they hold and develop with third-party AI developers, whether these are for-profit companies or not-for-profit organisations. Governments may share data openly through open data platforms, where data is made available for anyone to use in a manner that permits re-use and redistribution. Governments may also share data directly with AI developers (including through processes such as public procurement) by ensuring due diligence prerequisites are met. Governments may also share data through data trusts or other data-sharing intermediaries. These are managed by third parties incumbent with fiduciary responsibility to responsibly manage the data they steward, for example for a set of AI-related projects established under a particular set of rules or parameters. Governments may also facilitate data sharing with third-party AI developers directly, for example under contractual arrangements. This expands the role of government, not only as a direct provider of data, but also as a provider of access to data.

Broadening the notion of "provider"

The authors explored the contours of the conceptual framework of government as a provider of data, finding that this concept was readily mixed up with discussion around how data generated and held by governments is used in AI development, whether or not the government had intentionally shared their data with the AI developers or not, and regardless of whether the AI system the data was being used to develop was to assist in the delivery of a government or public service. Indeed, this is a very common occurrence as considerable amounts of data is used to train AI, which can very often include data that originated from state-sources. However, these instances told us little about the value of government data for developing robust, public interest, AI solutions – or about the need for responsible sharing of data. It was decided that it would be useful to clearly set out the different roles of government as a "provider" of data for AI development, and different responsibilities of a data provider therein.

Contextualising the need for responsible provision

In looking for examples of how government data had been shared with or used by AI developers, with a particular emphasis on including parts of the world that were underrepresented in AI governance debates to date, we came across several negative examples that demonstrated how dominant AI companies were extracting data from low-resourced African governments. One

notable example here is the 2018 Zimbabwean case,[3] whereby the facial recognition company, CloudWalk, entered a deal with the Zimbabwean government in advance of the national elections. The aim was to provide the government with a secure digital bio-ID system to be used to strengthen the upcoming elections, however, CloudWalk walked away with the national database of Zimbabwean faces – a vital resource in designing facial recognition technologies that can accurately identify African faces. At the time, Zimbabwe's laws governing personal data and cross-border data transfers were non-existent, providing no protection to the data rights of Zimbabwean citizens.

Transparency challenge: lack of available information

Another barrier we faced in defining the concept of government as a provider of data for AI, and collating evidence to better understand it in practice, was the lack of publicly available data about public data transfers. This was true wherever in the world we were looking for information. A case in point here is the DeepMind example (*outlined further below*) in the UK where the Royal NHS Free Trust shared patient data with Google Health for use in the Streams app. This information only became public due to the investigation commissioned by The Information Commissioner's Office after complaints were submitted that private patient data was being shared without proper approvals and public notification/awareness. Foundational documentation to the data sharing process was difficult to access; for example, the extension agreement for the second phase of work (issued while an investigation was ongoing) is still not publicly available.

[3] MISA Zimbabwe (2018). Digest: Facial recognition Technology and its possible impacts on privacy rights. https://zimbabwe.misa.org/2018/05/29/digest-facial-recognition-technology-privacy-rights/

2. Government as a Provider of Data

Governments have the potential and ability to drive AI development by leading the way by making available the main resource for AI development – data – and working with other actors to pursue the development of responsible AI further and deliver services and products together. Data collection, use, and disclosure have social and economic implications, with direct implications for its regulation.[4] The complexity by which data derives value within economies is an additional factor within governance options.[5] Data governance is not just about protection, but also access and interoperability, security and trust. The emergence of domestic AI solutions can nevertheless mean reliance on global, oligopolistic systems and platforms.[6] It is highly likely that AI will begin to expand these dependencies, except if governments step in and act as a provider of data for machine learning. In light of these considerations, there are four primary ways through which government owned data could be provided to AI developers. These are described below.

Open Data

Open data is the most well-known process for the government to share data and several governments have led initiatives on open data for many years (see for example the US[7] and UK[8] governments Open Data Portals). Open data is used to promote equal access to government data at limited cost to the users. There are principles governing open data such as non-discriminatory access and reusability of the data. Given open data means unrestricted public access, personal data cannot be shared via this mechanism.

In GPAI's 2022 report on Data Justice, they argued for embedding fair data-sharing practices in open data by stating that "calls for open data access can sometimes risk opening up opportunities for existing commercial interests to appropriate and exploit data assets, alongside raising privacy and data sovereignty concerns. Advancing data justice calls for the establishment of robust regimes of social licence and public consent, so communities can equitably access and benefit from their data. This includes by ensuring the provision of public data infrastructure which allows people not only to port or own their personal data, but to gain access to and beneficially use public data resources."[9]

To enable democratised access to data, the FAIR Principles, calling for findability, accessibility, interoperability and reusability, have been embraced.[10] In addition, these principles have been supplemented by the introduction of the CROP principles (Contracts, Rights in data, Open data,

[4] S. Viljoen (2021). *A Relational Theory of Data Governance*. The Yale Law Journal.
https://www.yalelawjournal.org/feature/a-relational-theory-of-data-governance
[5] G. Razanno (2021). *Data Localisation in South Africa.* Mandela Institute Policy Brief 06.
https://www.wits.ac.za/media/wits-university/faculties-and-schools/commerce-law-and-management/research-entities/mandela-institute/documents/research-publications/800482%20PB6%20Missteps%20in%20valuing%20data_REV%20Dec2021.pdf
[6] G. Razanno (2021). *AI4D - Digital and Biometric Identity Systems.* Research ICT Africa.
https://researchictafrica.net/publication/policy-paper-ai4d-digital-and-biometric-identity-systems/
[7] DATA.GOV. The Home of the US Government's Open Data
https://data.gov/
[8] Data.gov.uk. Find open data.
https://www.data.gov.uk/
[9] GPAI (2022). Data Justice: A Primer on Data and Social Justice, Report, November 2022, Global Partnership on AI.
https://gpai.ai/projects/data-governance/primer-on-data-and-social-justice.pdf
[10] FAIR - Findability, Accessibility, Interoperability, Reusability.

Public Interest) focusing on contractual agreements, portability and similar rights in data, as well as open data for the public interest.[11] As important as these principles are, they do not address concerns in the Global South around data ownership and data justice for original owners of data – people who may lose control and subsequent value in the use of their data in the development of AI. Another set of principles, known as the CARE principles (Community benefit, Rights in data, Responsibility, Ethical re-use) have been put forward by indigenous rights groups focusing on data sharing and use for community, data subject rights, and responsible and ethical re-use of data.[12]

An emerging model to address data justice is the notion of data solidarity. The objective of data solidarity is strengthening collective control and ownership of data.[13] It consists of three main pillars focusing on facilitating good data use, preventing and mitigating harm and returning profits to the public domain. In this model, governments can serve as independent trustees to ensure compliance with the pillars of this model.

Data Stewardship

While the government as a provider of data can have immense benefits for the development of AI, it can also create risks for governments, and data subjects, as the case studies discussed later in this report demonstrate. Data stewardship is emerging as an option for governments who are risk-averse to facilitating direct access to data from data subjects to third parties. This option helps navigate issues such as data subject consent and the protection of data subject rights. In GPAI's Framework Paper on Data Governance, it was argued that "data stewardship has emerged as a responsible, rights-preserving and participatory concept of data intermediaries, with the goal of providing more agency, transparency and protection to data subjects, negotiating with data requesters and seeking ways in which data can be of benefit to society."[14]

Two important data-sharing vehicles under data stewardship are data trusts and data institutions. Data trusts are "a form of data stewardship that supports data producers to pool their data (or data rights) with the aim of collectively negotiating terms of use with potential data users, through the oversight by independent trustees, with fiduciary duties, and within a framework of technical, legal and policy interventions that facilitate data use and provide strong safeguards against mis-use."[15] Data trusts can be important for building trust among publics in government data sharing, enabling data subject participation in the use of their data and limiting the liability exposure of the government as a provider of data. However, data trusts have not gained enough traction as a mechanism for data sharing by governments. Consequently, data institutions are emerging as key players in facilitating access to data.

[11] CROP - Contracts, Rights in data, Open data, Public Interest
[12] CARE - Community benefit, Rights in data, Responsibility, Ethical re-use
[13] B. Prainsack, S. El-Sayed, N. Forgó, Ł. Szoszkiewicz, P. Baumer (2022) *Data solidarity: a blueprint for governing health futures* Volume 4, Issue 11
[14] GPAI (2022). Data Governance Working Group – A Framework Paper for GPAI's Work on Data Governance 2.0, Report, November 2022, Global Partnership on AI, Paris. https://gpai.ai/projects/data-governance/Data%20Governance%20-%20A%20Framework%20Paper%20for%20GPAI%E2%80%99s%20Work%20on%20Data%20Governance%202.0%20.pdf
[15] OECD (2019), Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies, OECD Publishing, Paris. https://www.oecd.org/sti/enhancing-access-to-and-sharing-of-data-276aaca8-en.htm

Data institutions are organisations that steward data use towards public interest use. Governments can act as data institutions and steward data in various ways including protecting sensitive data and granting access under restricted conditions, through data analysis for other providers of data for public use, creating open datasets and acting as a gatekeeper for data held by other organisations.[16] Under this model, governments can also "develop and maintain identifiers, standards and other infrastructure for a sector or field, such as by registering identifiers or publishing open standards."[17]

For data institutions to function effectively, it requires a progressive recognition of data subject rights such as the right to data portability. Data portability provides restricted access through which the government can provide data "in a commonly used, machine-readable structured format, either to the customer or to a third party chosen by the customer."[18] Examples of initiatives using data portability are the US government's MyData series, and the UK government's Midata and subsequent Open Banking data portability initiatives.

However, while data portability will enable individual sharing of personal data with AI developers facilitated by governments, it will not enable bulk access to data and alternative data-sharing models such as public-private partnerships are needed for bulk data sharing.

Public-Private Partnerships (PPPs)

Some governments have used public-private partnerships to share their data with specific sectors to harmonise services offered to data subjects. For example, the OECD has cited the example of the Norwegian tax authority and its partnership with the financial sector to implement automatic exchange of loan-application data.[19] In the PPP, "instead of having to repeatedly ask users for information they have already provided to public administration, such data can be re-used (based on users' consent) during loan applications. The data is stored by Altinn, a digital infrastructure that links data from public agencies, municipalities and registers of more than 4 million inhabitants and 1 million enterprises in Norway."[20]

This approach raises several issues such as the legal basis for the further processing of the data submitted to government by third parties, unequal access to such data by different parties within the sector, privacy of data subjects, and the impact on competition if exclusive access is granted to selected parties.

Contracts

While governments often offer their data for free, contractual arrangements with third parties are used to prescribe the scope of obligations a third party must comply with when accessing government data. These contracts come in the form of data-sharing, or data-processing

[16] B. Snaith & J. Massey. (2021), 'Data Institutions for Climate Action', https://theodi.org/article/data-institutions-for-climateaction/
[17] Ibid.
[18] OECD (2019), Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies, OECD Publishing, Paris. https://www.oecd.org/sti/enhancing-access-to-and-sharing-of-data-276aaca8-en.htm
[19] Ibid.
[20] Ibid.

agreements, and cover various areas: description of the product or services; data collection, access and control; nature, purpose and duration of data processing; types of data and categories of data subjects; obligations and rights of the controller and processor; data location and who has access to it; data retention, ownership, and security measures; limits of disclosure to other parties; or processing beyond the initial purpose of data sharing.

All four models of data sharing have their benefits and risks. Open data, which may facilitate better public access to data, still attracts concerns around data access and justice in LMICs; these may be addressed by the adoption of FAIR principles and the model of data solidarity. Data stewardship has the potential to grow in usage with governments as facilitators of access to data – particularly using data institutions. Public Private Partnerships provide a useful model allowing collaboration between government and the private sector through cross-licensing agreements, incentivising partnerships with research funding or the provision of the relevant infrastructure for data sharing. However, these partnerships can also lead to unequal access to data and grant unfair competitive advantage to private sector partners. Contractual arrangements between governments and private sectors come with risks as well which can be potentially mitigated if data sharing agreements embrace an assessment of the public value of data use by AI developers.[21] This assessment can enable governments to extend obligations such as open AI to developers creating AI solutions based on using government data.

3. Case Study Analysis

This section provides illustrative case studies from around the world on access to government-held data for the development of AI systems. The case studies provide existing examples of how government-held data has been shared with AI developers for the development of AI systems in key sectors such as climate, agriculture, and social protection. Another case study examines the government's role in facilitating access to data by AI developers in healthcare. These case studies are intended to bring to the forefront the practicality of government data sharing by detailing current practices of government data sharing around the world. They also demonstrate the unavoidable legal and technical challenges that emerge from the role of governments in data sharing, together with the legal benchmarks that prevent harms and ensure legal compliance.

This section covers case studies from the UK, Taiwan, Colombia, and Nigeria which provide practical examples of how government data has been used for AI development and how these governments have navigated issues around legal compliance i.e., data protection, public procurement, automated decision-making, public-private partnerships, technology, and public attitudes.

[21] S. El-Sayed & B. Prainsack (2022) “PLUTO/PublicVal - Public Valut Tool” https://digitize-transformation.at/news-und-events/detailansicht/news/plutopubval-public-value-tool/

Name of the case study	Summary of the case study	Location	Sector	Data-sharing model
National Health Services	A contractual partnership between The UK's NHS and Google's AI firm DeepMind to share data for digital solutions development.	United Kingdom	Health	Data-Sharing Agreement
The Health Passbook	Consented health data transfer to third-party app developers via a government-run "Health Passbook."	Taiwan	Health	Data Stewardship
The Rapid Response Register (RRR) for cash transfers in Nigeria	AI-powered data generation for social protection programming targeting vulnerable populations affected by the pandemic.	Nigeria	Social protection, benefits data	Contract
The Aclimate Agricultural Data Platform	A data commons initiative to build AI tools to provide farmers with actionable information.	Colombia	Agriculture	Data stewardship

Table 1: Summary of the case studies

Each case study provides a context on the objectives and approach of each government in data sharing or facilitating access to data by third party AI developers. This is followed by an analysis of the relevant legal frameworks enabling data sharing in those countries and government attitudes towards digital innovation. We then undertake a detailed analysis on the mechanisms used for data sharing in these use cases and the potential benefits and risks in these approaches from a data governance perspective. We identify a detailed set of findings from each case study and develop recommendations that could inform future data governance approaches.

All four countries studied in this report have a data protection law which provides a framework of principles and obligations to guide data sharing by the government. In the UK, the Data Protection Act of 1998 was applicable during the events of the case study and our analysis will show why a strengthened data protection law with an oversight body with strong enforcement powers is crucial in maintaining responsible data governance practices. In Taiwan, our analysis will show that the application of a general data protection law alongside other sectoral laws also play a crucial role for the government to facilitate access to data by third parties. In Colombia, while a data protection law was in place, a contractual framework to guide a tripartite alliance for data sharing was more important to establish a data commons model of data sharing. In Nigeria, the absence of a data protection law during the time period when the use case occurred shows why alternative legal regimes such as a data protection regulation which was applicable at the time is not an effective mechanism to safeguard data subject rights and protect the privacy of data subjects.

3.1. National Health Service Case Study

Location: United Kingdom (UK)
Players: UK National Health Services (NHS) and Google's DeepMind

Context

The UK's National Health Services[22] (NHS) Royal Free London NHS Foundation, which is the UK's national healthcare service provider, contracted Google's AI firm DeepMind[23] to develop a technological software that would be used for the detection and treatment of kidney diseases. The two parties entered into an Information Sharing Agreement in 2015, for the development of an app—'Streams'—that would be used for the detection of acute kidney injuries across NHS hospitals.[24]

The Agreement resulted in the transfer of personal identifiable information of 1.6 million patients across three NHS hospitals. In early 2016 when announcing its collaboration with NHS on the development of the Streams app, DeepMind did not indicate the vast amount of patient data that the NHS had given it access to.[25] An investigation conducted by the *New Scientist* revealed the vast amount of patient data that DeepMind had been given by the NHS.[26] This caused a public uproar among the UK citizenry and raised critical concerns on privacy and data protection.[27]

This pushed a move from the UK's Information Commissioner's Office (ICO) which commenced investigations into the Information Sharing Agreement between DeepMind and NHS. The ICO ruled that the data-sharing agreement between the two entities failed to comply with UKs Data Protection Act, 1998.[28] This Ruling was based on the fact that the Information Sharing Agreement (ISA) failed to comply with key principles on data protection: Principle One on fairness and lawfulness; Principle Three on adequacy, relevance, and minimization; Principle Six on protecting data subjects rights when processing personal data; and Principle Seven on ensuring data contracts have the appropriate technical and organisational mechanisms in place.[29] It also based the ruling on the fact that patients were not informed of the sharing of their personal data with DeepMind, nor was their consent sought.[30] The ICO added that patients were not provided with an option to opt out, and that there was lack of transparency around the data sharing Agreement.[31]

---

[22] NHS Royal Free London, NHS Foundation Trust
https://www.royalfree.nhs.uk/about-us/
[23] Google DeepMind
https://www.deepmind.com/about
[24] Sarah Boseley and Paul Lewis (2016). *Smartcare: How Google DeepMind is Working with NHS Hospitals.* The Guardian.
https://www.theguardian.com/technology/2016/feb/24/smartphone-apps-google-deepmind-nhs-hospitals
[25] Google DeepMind (2016). *We are very happy to announce the launch of DeepMind Health*
https://www.deepmind.com/blog/we-are-very-excited-to-announce-the-launch-of-deepmind-health
[26] Hal Hodson (2016). *Revealed: Google AI has access to huge haul of NHS patient data.* New Scientist.
https://www.newscientist.com/article/2086454-revealed-google-ai-has-access-to-huge-haul-of-nhs-patient-data/
[27] Natasha Lomas (2016). *Concerns raised over broad scope of DeepMind-NHS data sharing deal.* TechCrunch.
https://techcrunch.com/2016/05/04/concerns-raised-over-broad-scope-of-deepmind-nhs-health-data-sharing-deal/
[28] UK Information Commissioner's Office Ruling.
https://ico.org.uk/media/action-weve-taken/undertakings/2014353/undertaking-cover-letter-revised-04072017-to-first-person.pdf
[29] Ibid.
[30] Timothy Revell (2017). *Google DeepMind's NHS data deal failed to comply with the law*. New Scientist.
https://www.newscientist.com/article/2139395-google-deepminds-nhs-data-deal-failed-to-comply-with-law/
[31] Ibid.

The ICO found there was no legal basis for the use of patient data for the testing of the app by DeepMind. However, it issued no fine to Google and DeepMind but urged them to take action to comply with the UK Data Protection Act, 1998, including conducting a Data Protection Impact Assessment (DPIA) and a legal audit on the Streams app.[32]

Governance Setting

The UK left the EU on 31st January 2020.[33] The UK's exit from the EU means the EU General Data Protection Regulation (GDPR) no longer applies to the UK's data governance regime. Data Protection in the country is instead regulated by the Data Protection Act, 2018[34] which brough key provisions of the EU GDPR into UK law. The Data Protection Act, 2018, repealed the previous Data Protection Act, 1998, which was in force before the EU GDPR. For the purposes of this case study, it is important to note that the repealed Data Protection Act, 1998,[35] of the UK was the applicable law during the data-sharing agreement between the NHS Trust and Google.

Health data in the UK is governed by a comprehensive data protection and information governance framework that provides key guidelines on protection and sharing of health data. At the helm of this framework is the country's Data Protection Act, 2018, which governs processing of personal data by the UK Government, businesses, and organisations.[36] The implementation and enforcement of this Act falls under the purview of the Information Commissioner's Office (ICO)[37] which was established under the Act with this mandate. Following the Data Protection Act in the governance of health data are the Caldicott Principles which are formulated to govern the use and sharing of health data "*within health and social care".*[38] Health data in the UK is further governed by the Health and Social Care Act, 2012,[39] which sets out the legal basis of processing patient identifiable information. The Health Services (Control of Patients Information) Regulations 2002 further "*sets out the circumstances in which confidential patient information may be processed for medical purposes".*[40]

There are three key regulatory bodies in the UK which are relevant to the scope of the ISA and their approval was not sought. These are the UK's ICO,[41] the Medicines and Healthcare Products

[32] UK Information Commissioner's Office Ruling.
https://ico.org.uk/media/action-weve-taken/undertakings/2014353/undertaking-cover-letter-revised-04072017-to-first-person.pdf
[33] Nigel Walker (2021). *Brexit timeline: events leading to the UK's exit from the European Union.* House of Commons Library.
https://researchbriefings.files.parliament.uk/documents/CBP-7960/CBP-7960.pdf
[34] Data Protection Act, 2018
https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted
[35] Data Protection Act,1998
https://www.legislation.gov.uk/ukpga/1998/29/enacted
[36] Data Protection Act
https://www.gov.uk/data-protection#:~:text=The%20Data%20Protection%20Act%202018,Data%20Protection%20Regulation%20(GDPR).
[37] UK Information Commissioner's Office
https://ico.org.uk/
[38] National Data Guardian. The Eight Caldicott Principles
https://assets.publishing.service.gov.uk/media/5fcf9b92d3bf7f5d0bb8bb13/Eight_Caldicott_Principles_08.12.20.pdf
[39] NHS Digital. Appendix 2.4: Section 251 (S.251)
https://digital.nhs.uk/services/national-data-opt-out/operational-policy-guidance-document/appendix-2-definitions#a2-4-section-251-s-251-
[40] The Health Service (Control of Patient Information) Regulations 2002
https://www.legislation.gov.uk/uksi/2002/1438/schedule/made
[41] Data Protection Act, 2018

Regulatory Agency,[42] responsible for regulating medical devices such as the Streams app in the UK, and the Health Research Authority.[43] The latter organisation reviews and approves medical research to ensure transparency and maintenance of ethical standards, and provides guidance in the processing of patients' personal information in instances where there is no practicality in obtaining consent, *"for research and non-research projects".*[44] Approval from these regulatory bodies would have ensured the protection of patients' data protection rights.

Analysis

Patient data under the Agreement was transmitted in *"live batch data streams"* using technologies such as TCP/IP encrypted channel and SFTP secure transfer.[45] The data in question included patient medical records for the past five years.[46] The data was stored by a third party contracted by DeepMind's parent company, Google.[47]

The NHS agreement raises several data governance issues that highlight the need for both effective legal frameworks, strong oversight institutions and collaborative approaches towards implementation and enforcement of the law.

Consent

The NHS Trust, which acted as a data controller when entering into the Information Sharing Agreement with DeepMind, did not seek the consent of patients whose data had been processed by the Trust. The patients were not aware that their data had been transmitted to DeepMind for the development of the app, contradicting the Caldicott Guidelines on informing patients on how their data is used.[48]

In responding to this investigation and the concerns raised on patient consent, the NHS stated that it relied on 'implied consent'[49] as a healthcare provider given that it was providing 'direct care' to patients.[50] DeepMind on the other hand also argued that it was exempted from seeking patient consent as it was providing 'direct care' to NHS patients.[51] Both parties stated that the fact that they were sharing personal data for providing 'direct care' to NHS patients gave them a legal basis for processing patient personal identifiable information. While the UK has a comprehensive legal

https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

[42] Medicines & Healthcare products Regulatory Agency
https://www.gov.uk/government/organisations/medicines-and-healthcare-products-regulatory-agency

[43] Health Research Authority
https://www.hra.nhs.uk/about-us/what-we-do/

[44] Data Protection Act, 2018
https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

[45] Natasha Lomas (2016). *DeepMind Health Inks New Deal with UK's NHS to Deploy Streams App in Early 2017*. Tech Crunch.
https://techcrunch.com/2016/11/21/deepmind-health-inks-new-deal-with-uks-nhs-to-deploy-streams-app-in-early-2017/

[46] Ibid.

[47] Ibid.

[48] National Data Guardian. The Eight Caldicott Principles
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/942217/Eight_Caldicott_Principles_08.12.20.pdf

[49] Jane Wakefield (2017). *Google DeepMind patient app legality questioned*. BBC News.
https://www.bbc.com/news/technology-39934316

[50] Subhajit Basu (2016). *Should the NHS share patient data with Google's DeepMind?*. WIRED.
https://www.wired.co.uk/article/nhs-deepmind-google-data-sharing

[51] Amy Dickens (2021). *The right to health implications of data-driven health research partnerships.*
https://repository.essex.ac.uk/31194/1/PhD-%20FINAL%20VERSION%20(w.%20corrections).pdf

framework for the processing of health data, which provides sufficient grounds for 'implied consent' to be used as a legal basis for processing patient personal data, the original objective for sharing data with DeepMind was not to provide direct care to patients. This nullified the legal basis relied on by the parties.

Data protection

The ISA between the parties allowed the transmission of excessive identifiable patient information, exceeding what would have been ordinarily required for the development of such an app. The data included information of patients who were not active at the NHS, and of those who were not receiving treatment for acute kidney injury – the intended purpose of the app. The second concern on data protection is that DeepMind was not restricted in the further processing of the data shared with it for other purposes in contravention of the Data Protection Act.

The NHS Trust only ran checks with the Information Governance Toolkit which is a "*self-assessment"* tool designed by NHS digital to enable entities such as DeepMind in this case to cross-check their "*technical infrastructure"* to ensure that it is secure.[52] The Toolkit is designed for the purpose of enabling organisations to ensure that their "*computer systems"* have the technical capability of processing data from the NHS but it does not address mechanisms for data protection when data is transferred between the parties.[53]

Transparency

The whole data-sharing arrangement between DeepMind and NHS lacked transparency, and the public only knew of the data-sharing agreement due to a third party investigation several months after data had already been shared with DeepMind.

Potential Benefits and Risks

Benefits

The key potential benefit of the NHS project with DeepMind was the accelerated diagnosis of acute kidney injuries by healthcare workers within the NHS. The app was designed to provide medical professionals with instant access to patient medical data for diagnosis and treatment. It would improve efficiency by helping healthcare workers avoid having to manually go through various NHS systems, and would assist healthcare workers to diagnose acute kidney injuries, and provide patient-medical alerts directing doctors and nurses to patients in seconds – thus saving patients from deteriorating and losing their lives in hospitals.

Risks

Consent

A key risk with the data-sharing agreement between the NHS and DeepMind was the fact that the NHS had shared patient data with DeepMind without informing data subjects about the project and obtaining their consent, in contravention of the UK Data Protection Act.

[52] Julia Powles and Hal Hodson (2017). *Google DeepMind and Healthcare in an Age of Algorithms.* National Library of Medicine. https://pubmed.ncbi.nlm.nih.gov/29308344/
[53] Ibid.

Oversight review and approval

None of the regulatory bodies with oversight powers over this project were informed and their guidance was not sought before data was shared with DeepMind.

Data use limitations

The agreement between NHS and DeepMind did not restrict DeepMind from further processing the sensitive personal health information shared with it for its own purposes.

Data minimisation

The NHS could not provide any satisfactory justification for why DeepMind was granted access to vast amounts of data including those of inactive patients who were not under NHS care.

Conclusion

The aim of sharing data by the NHS with DeepMind is an excellent example of the immense potential for AI to do good. However, the NHS's non-compliance with applicable legal frameworks turned a worthwhile initiative into an illustration of how data governance is central to the responsible use of AI. The NHS initiative tested the boundaries on use of technology in health diagnosis and the role of data subjects in granting consent for use of their personal data for AI development in the public interest. While the NHS used its information governance toolkit to assess the technical infrastructure around data sharing, it did not assess the ethical governance around data sharing in the first place. This shows the need for an evolved understanding of data governance as recommended below.

Recommendations

Privacy-enhancing technologies

The NHS could have explored various options to protect the personal health information that was shared with DeepMind during the testing phase of the app. These included privacy enhancing techniques that make it difficult to identify a data subject in a data set as well as granting the choice to patients to opt-in to a personal data wallet to share their personal data. However, this would have required transparency on the part of the parties in the first place during the design phase. Leveraging privacy-enhancing technologies (PETs) could improve data availability whilst respecting data privacy and minimising data misuse.

Transparency

A major problem with the NHS and DeepMind project was the lack of transparency by the NHS about the project. To deepen public trust in AI, governments need to be transparent in the conceptualisation, design and implementation of data-sharing projects especially in cases where the government will also be a user of AI. These transparent measures include full disclosure to oversight authorities for their review and guidance even when regulatory frameworks do not require such disclosures.

Defining the legal basis for processing

All data protection laws provide different bases for governments to process personal data through disclosure to third parties. It is important for any data sharing by the government to a third party to rely on the right legal basis. Consent of data subjects is not only the legal basis for sharing and other bases such as processing in the public interest could also apply.

Effective data governance mechanisms

Before the public denunciation of the NHS and Google's DeepMind data-sharing partnership, data subjects were not fully aware of the type of data that was being shared and public participation in the process was restricted. However, this partnership could have been a good example of a government using a data stewardshipto encourage patients to voluntarily share their information for better healthcare diagnosis.

Regulatory certainty

Both the Data Protection Act and various health laws applied in the NHS case study. DeepMind appeared to exploit the fact that various institutions had different and limited mandates in terms of their enabling laws to argue that they were not obliged to comply with any obligations set out by these oversight institutions. This lack of regulatory certainty can be addressed by recognising the Data Protection Commission as the primary oversight institution for data governance when there is a conflict in institutional mandates.

A Comparison: Deepmind and Moorfields Eye Hospital

The NHS subsequently entered into further data-sharing agreements with Deepmind, for which both parties acted with considerably greater attention to patient consent and medical and data ethics – for example in the case of Moorfields Eye Hospital,[54] and with University College London Hospital,[55] among others. By approaching data sharing responsibly, Deepmind and the NHS have since been able to successfully work together to improve patient health outcomes while being in full compliance with relevant data sharing legislation and guidance.

[54] https://www.theguardian.com/technology/2016/jul/05/google-deepmind-nhs-machine-learning-blindness
[55] https://www.theguardian.com/technology/2016/aug/30/google-deepmind-ucl-ai-radiotherapy-treatment-

3.2. The Health Passbook

Location: Taiwan
Players: National Health Insurance Administration, third-party AI App developers, and Health Health Passbook Users

Context

Taiwan introduced mandatory National Health Insurance (NHI) in 1995 and adopted a single-payer healthcare system covering more than 99% of the population. Data in the National Health Insurance Research Database (NHIRD)[56], such as personal health information, accompanied by a data subject's national ID number, gender, ethnicity, income, and medical information, can be directly authorised for sharing by the data subject with third parties for academic research.

The Health Passbook was launched by the National Health Insurance Administration (NHIA) in 2014 with the emerging mobile health trend of self-monitoring and self-care. Through the Health Passbook, users can transfer their health data collected and stored by NHIA to third-party apps. Users of the Health Passbook consent to the primary uses of their health data and the scope of consent may be limited to using data for better personalised health-related services, but, in some cases, it also includes research and development for Health AI technology. The use of shared data cannot go beyond specific purposes agreed upon by the parties.

Despite the high smartphone penetration rate, few Taiwanese citizens downloaded and registered the Health Passbook app in the first few years after it was released. The low usage was partly because of the technical hurdles people have to overcome throughout the authentication process.[57]

In 2018, NHIA partnered with telecommunication companies and adopted a mobile phone number authentication system for Health Passbook, which takes advantage of the existing real-name registration for SIM cards in Taiwan.[58] However, the turning point for the adoption of the Health Passbook among the general population was when the app included mask purchase and COVID testing and vaccination records during the pandemic.[59]

The Health Passbook enables a streamlined data flow from Health Passbook users to third-party apps. The government of Taiwan is not a direct provider of data to third parties but facilitates the sharing of data by a data subject with a third party provider through the Health Passbook. The model points to an alternative way of making data held by the government available for AI research. Such a tripartite model has benefits and risks compared to those where the government itself is the party of the data-sharing agreement, and this case study illustrates that.

[56] Cheng-Yang Hsieh, Chien-Chou Su and others (2019). *Taiwan's National Health Insurance Research Database: past and future*. National Library of Medicine. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6509937/
[57] See, Guan-Lin Ho, An Empirical investigation of the Factors that Influence Individuals to Download Medical Data from My Health Bank, I-Shou University, unpublished Master's thesis (2017) (in Chinese).
[58] See Chen Tzi-Fa, The Experience of Establishing National Health Insurance App Mobile Phone Number Identity Authentication System, Government Agency Information Communication No.357, p31-36 (2019) (in Chinese).
[59] See https://health.ltn.com.tw/article/breakingnews/3698882 (in Chinese).

Governance Setting

Taiwan has adopted a Personal Data Protection Act (PDPA) and Article 6(I)(4) of the PDPA, provides that, if necessary, personal health data can be used for academic research without data subjects' consent for public health and healthcare purposes.[60] In 2022, while upholding the constitutionality of Article 6(I)(4) of the PDPA, the Constitutional Court of Taiwan ruled that the absence of regulations enabling data subjects to opt out of the release of their data for academic research violated the constitutional protection of the right to information privacy.

Analysis

Data in the Health Passbook can be categorised into three categories: personal records, including but not limited to records concerning physiological measurement, allergy, vaccination, menstrual cycles, and a major illness or injury certificate; medical treatment and prescription records, including prescription, surgery, and records as well as outpatient clinical records in Western medicine, traditional Chinese medicine, and dentistry; and testing reports and examination results, ranging from cancer screening, blood glucose, cholesterol, medical images and pathology samples to a liver cancer risk prediction.

A feature of the Health Passbook is that users can transfer their personal health data, after their identities are authenticated, to a verified third-party app, which may provide services such as self-monitoring, health risk assessment, and personalised health virtual assistant. The NHIA developed a Software Development Kit (SDK) that is free of charge in facilitating data release. However, access is limited to (1) public administration agencies, (2) contracted National Health Insurance medical care institutions, (3) incorporated foundations, and (4) for-profit corporations.[61] China-invested enterprises are not eligible for the application. The SDK application process is divided into two phases for data security purposes. An applicant first has to demonstrate their app can function appropriately in an integrated test environment with the Health Passbook. In the second phase, with the successful function record at hand, the applicant needs to provide basic information about the app, contact information, Mobile Application Basic Security certificate, and a privacy policy and statement that will be disclosed to users before they transmit their health data to the third party app.[62] Only apps approved by NHIA can be integrated with the Health Passbook SDK via an NHIA-verified application programming interface (API).

Third parties' verified apps are restricted from directly connecting to the NHIA server. Instead, individual users who make such requests via third parties' verified apps would be redirected to the Health Passbook app, where they can freely determine the term and scope of shared health data and send their requests for sharing to the NHIA server through the app. Upon receiving the request from a Health Passbook user and confirming identity authentication, the NHIA server generates an encrypted health data file, which is downloaded and stored in the user's mobile device. The verified third-party app obtains the encryption key through the Health Passbook SDK, allowing it to access

[60] Article 6(I)(4) PDPA specifies an exemption for prohibition on collecting, processing and using personal health data when "it is necessary for statistics gathering or academic research by a government agency or an academic institution for the purpose of healthcare, public health, or crime prevention, provided that such data, as processed by the data provider or as disclosed by the data collector, may not lead to the identification of a specific data subject."
[61] Article 3, Regulations Governing Use and Management of Health Passbook Software Development Kit, National Health Insurance Administration, Ministry of Health, and Welfare (2022).
[62] Article 5, PDPA.

and de-encrypt users' downloaded health data files on their mobile devices (See Figures 1 and 2). No personal health data is exchanged between NHIA and third-party app developers throughout the process. The data transfer occurs between NHIA and Health Passbook users, and between Health Passbook users and third-party app developers. Third-party app developers have to enter a contract with NHIA to access Health Passbook SDK, but the contract does not obligate NHIA to share NHIRD datasets with the developers.

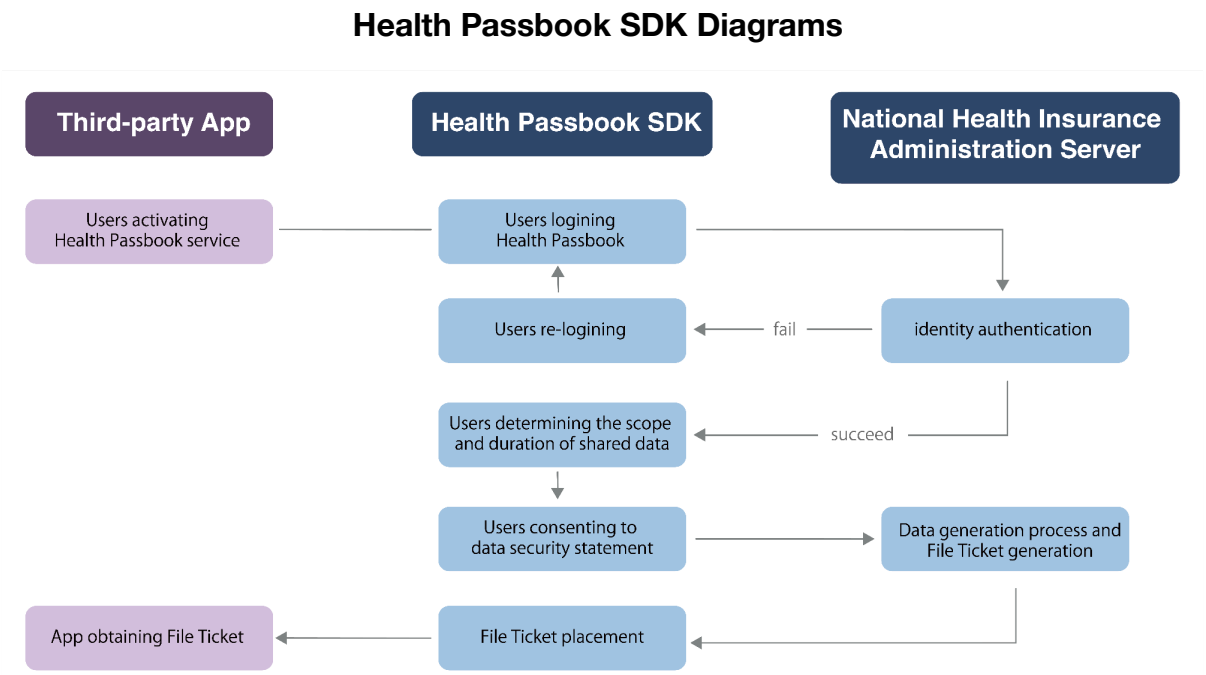

Figure 1: Health Passbook SDK Fronted Operation Diagram (Source: case study compilation)

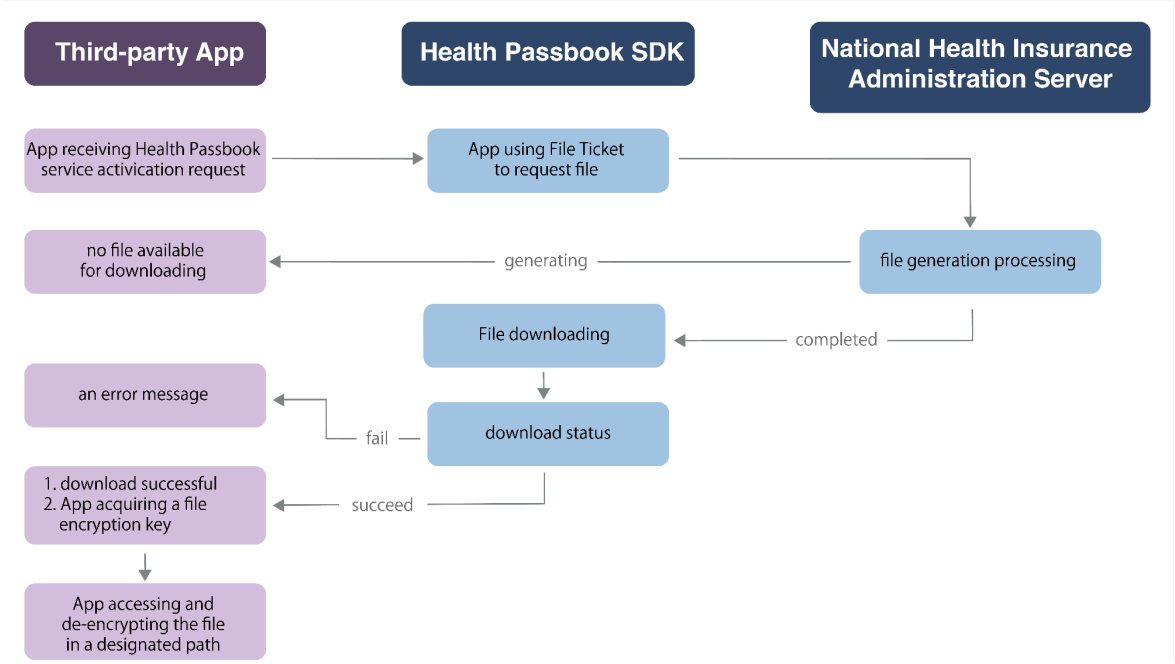

Figure 2: Health Passbook SDK Backend File Acquisition Procedure Diagram (Source: case study compilation)

To protect an individual's information autonomy, each instance of data sharing must receive the individual's authorisation and consent, and only then can a third-party app access the data. Furthermore, according to Article 3 of the PDPA, the user is entitled to the right to subsequently opt out of the collection, processing, or use of personal data – and the right to erase personal data that the private entity should respect. However, even though individuals' consent is essential in the NHIA Health Passbook data-sharing model, the exact nature through which informed consent should be taken is not precisely specified.

To ensure data minimisation and transparency, the NHIA provides SDK applicants with a privacy policy template enumerating suggested provisions, including the specific purpose of health data collection, and the location, time, and purposes of data use. The privacy policy submitted is reviewed by the NHIA when considering the third party's application for the Health Passbook SDK. If the third party's use of the shared data exceeds the necessary scope of the specific purpose, the NHIA can terminate the SDK service. However, there is no independent mechanism to supervise and monitor the data processing activities of the third party.

As of December 2022, 149 organisations had applied for the Health Passbook SDK. The total number of apps submitted is 349. Among them, 64 (by 31 different developers) were already available. More than 30,000 Health Passbook users have transferred their health data to their chosen apps[63] (the term and scope of shared data are up to individuals' decisions). Some of these apps have used the data, shared by individuals through their Health Passbook, to facilitate AI-based personalisation. For example, a digital health start-up company, Lydia.ai, developed an app called AI Health Index to help users better understand their health conditions more intuitively. Furthermore, Lydia.ai partnered with an insurance company to help its app users evaluate and purchase health insurance products and plans based on the AI Health Index.[64]

Potential Benefits and Risks

Benefits

The major benefit of the Health Passbook is the control and autonomy it gives data subjects to make choices on who they want to share their data with. Data subjects may have their preferred digital health apps for self-monitoring and self-care. The Health Passbook allows them to share personal health data with these apps, and App developers are also able to provide more personalised health AI services, or further their AI research and development by utilising the health data shared by data subjects. This level of control given to data subjects is an excellent example of how data subjects rights – such as right to consent, access, and opt-out – can be fully protected in AI development. Furthermore, the role of government in facilitating access rather than as a direct provider of personal data significantly limits the liability of government in relation to the data processing once the government's due diligence on the third parties receiving the data is completed.

[63] Lin Huiqin (2022). *Open third-party APP to integrate health passbook and control health management information at once*.
https://health.ltn.com.tw/article/breakingnews/4166871 (in Chinese)
[64] Zheng Yuan (2021). *Combined with health passbook Lydia AI to build health ecosystem*.
https://shorturl.at/BCNRU (in Chinese)

Risks

Assumption of Informed Consent

There is an assumption that users of the Health Passbook fully understand the implications of sharing their data with third parties and how their data will be processed. The uses of data for AI research and development may come out as a surprise for many data subjects, despite their "informed consent." NHIA provides a privacy policy boilerplate and requires app developers to submit a privacy policy that will be used when applying for the Health Passbook. However, how most third-party apps display privacy policies does not make it easy for data subjects to navigate the fine print. Eager to access a better health self-management app, many data subjects may not read the terms and will not know using the app means they effectively agree to their data being shared for commercial AI development. In this regard, although the Health Passbook may help facilitate the secondary use of government health data for AI research, it falls short of ensuring data subjects' information autonomy, undermining the foundation of public trust in government stewardship of their health data.

Conclusion

The Health Passbook enabled NHI-insured persons in Taiwan to access their NHI health records and other related datasets, making health self-management more efficient and effective. Its rapid rise in popularity after 2019 was in part because of lower "effort expectancy" (after a mobile phone number authentication system was introduced) and stronger "performance expectancy" (because of the integration with mask purchase, testing and vaccination records during the COVID pandemic).[65] The multiple uses of the Health Passbook enabled user uptake but the number of data subjects who have opted to share their data with third parties is very low compared to the total number of users of the Passbook.[66] The facilitation model of the Health Passbook puts data subjects in control, but what individuals collectively want to share and the modest scale of the sharing, as illustrated in this case study, may be of limited use for a firm seeking access to large datasets to train AI.

Recommendations

Data governance checklist

For any government that is keen on embracing the role of facilitating access to rather than a direct provider of data, a data governance checklist such as a privacy-by-design tool, a consent template, a data subject notification template, or data subject rights-redress framework should be developed. This will help in harmonising the requirements with which all third parties must comply with if they want the government to facilitate access to data for them. This checklist will allow compliance with explicit fiduciary duties on third parties to act in the best interests of data subjects using the app.

[65] For the factors that influence Taiwanese people using Health Passbook under the model, see Pi-Jung Hsieh & Hui-Min Lai, *Exploring People's Intentions to Use the Health Passbook in Self-Management: An Extension of the Technology Acceptance and Health Behavior Theoretical Perspectives in Health Literacy*, 161 Technological Forecasting and Social Change 120328 (2020).

[66] By June 2022, the total number of users of the Health Passbook passed 10 million, making it the most used digital service offered by Taiwan's government. More than 40 percent of Taiwan's population downloaded and utilised the app.

Data Stewardship

The Health Passbook is typical of AI solutions emerging in LMICs, which provide socio-economic improvements in a variety of sectors. To fully achieve the objectives of such initiatives, a favourable regulatory and policy environment is needed. This will ensure the government either acts as a data steward by enabling access to data for public value, or that it creates the right public-private partnerships in procuring AI for the public good. AI procurement tools should be developed that ensure strong, transparent and responsible procurement processes are in place to acquire AI systems.

Data Trusts and Privacy-Enhancing Technologies

Taiwan adopted the model of data stewardship, and could have gone a step further by facilitating the creation of a data trust or data institution, so that data could be pooled to facilitate R&D by verified third parties granted access to the app. However, these would need to be safeguarded by data subject rights on access, portability, and erasure. These can be achieved by privacy-enhancing technologies (PETs), such as differential privacy, which masks an individual in a data set; or personal data 'wallets' (otherwise known as Personal Information Management Systems), which allow data subjects to manage who has access to their data.[67] Such PETs are reviewed in more detail in Section 6 of this report.

[67] GPAI (2022). Data Governance Working Group – A Framework Paper for GPAI's Work on Data Governance 2.0, November 2022, GPAI Tokyo Summit. https://gpai.ai/projects/data-governance/Data%20Governance%20-%20A%20Framework%20Paper%20for%20GPAI%E2%80%99s%20Work%20on%20Data%20Governance%202.0%20.pdf

3.3. The Rapid Response Register (RRR) for cash transfers in Nigeria

Location: Nigeria
Players: Nigerian government, Telecommunications companies

Context

In January 2021, the Nigerian government launched the Rapid Response Register (RRR) for COVID-19 cash transfers under the platform of the National Social Safety Nets Project (NASSP), funded by the World Bank.[68] The National Social Safety Nets Coordinating Office (NASSCO) administers this programme. According to the Vice President of Nigeria, during the launch of this programme, "the RRR is designed to focus mainly on the urban poor wards selected using scientifically validated methods of satellite remote sensing technology, machine learning algorithm and big data analysis."[69] The intention behind the RRR is to build a responsive framework for capturing the urban poor and vulnerable populations across Nigeria. This category of Nigerians are also described as the working poor who had an average of less than 5000 naira ($5) in their bank account and live in slum areas in Nigerian cities. This vulnerable group experienced high levels of informal employment, which was disrupted given the lockdowns and stay-at-home orders issued by countries such as Nigeria. Consequently, Nigeria implemented a rapid cash transfer programme to mitigate the devastating impact of the pandemic. Identifying the urban poor will present a significant challenge for any country given the high levels of daily mobility, homelessness, and, in the case of Nigeria, a lack of a reliable and centralised personal identity management system. Consequently, the government contracted an AI developer and telecommunications companies to use big data analysis to identify and target eligible beneficiaries. The government transferred its existing social protection register to the developer and telecommunications companies to supplement the list of eligible beneficiaries on the list.

Over 20 million people were identified by NASSCO as potential social protection beneficiaries for COVID-19 cash transfers.[70] Given Nigeria's census was last conducted in 2006, to generate more recent data on granular poverty estimates per location, "recent advances in deep learning were used to construct a high-resolution poverty map from satellite imagery and other sources of geospatial Big Data."[71] According to the World Bank, "these techniques work by learning how to predict poverty by being exposed to a large dataset that matches ground-truth labels of poverty (from geo-located household surveys) to imagery and other geospatial data. Intuitively, the algorithms learn the visible features that are predictive of poverty, such as road quality, building density, and land topology."[72] To rapidly scale up the NASSP to cover the urban poor, this led the Nigerian government to design the RRR to quickly identify and verify the urban poor using spatial and satellite imagery data analysis.

---

[68] The World Bank (2021). Nigeria to Scale-up Delivery of Social Assistance to 10.2 Million Households. https://www.worldbank.org/en/news/press-release/2021/12/16/nigeria-to-scale-up-delivery-of-social-assistance-to-10-2-million-households

[69] Y. Osinbajo, 'Launch of the Rapid Response Register for the economic Sustainability Plan January 2021'https://www.yemiosinbajo.ng/launch-of-the-rapid-response-register-rrr-for-the-economic-sustainability-plan-covid-19-cash-transfer-on-19-01-2021/

[70] I Apera ,'Rapid Response Register for COVID-19 Cash Transfer: The Nigerian Response to COVID-19 in urban communities' NASSCO Presentation

[71] Ibid.

[72] J. Blumenstock, J. Lein and others (2021). *Using Big Data and Machine Learning to locate the poor in Nigeria*. World Bank Blogs. https://blogs.worldbank.org/opendata/using-big-data-and-machine-learning-locate-poor-nigeria

Governance Setting

Nigeria's Data Protection Act was passed in June 2023 with the aim of boosting Nigeria's digital economy, according to the Nigerian President when he signed the law. The Act establishes data subject rights for Nigerians and a Commission that has already been established to exercise oversight over the implementation of the Act. Before adopting the Data Protection Act, the 2019 National Data Protection Regulation (NDPR) was issued by the National Information Technology Development Agency (NITDA) to regulate and control the use of data in Nigeria. This Regulation was applicable during the rollout of the RRR. The Regulation mandates all public and private organisations in Nigeria that control data of natural persons to make available to the public their respective data protection policies within three months after the date of the issuance of the Regulation.

Following the adoption of the NDPR, two frameworks were applied to support its implementation. These are the 2019 NDPR Implementation Framework and the Guidelines for Managing Personal Data by Public Institutions in Nigeria, 2020. According to the Guidelines for the Management of Personal Data by Public Institutions of Nigeria, a higher standard of consent-seeking methodapplies to the processing of sensitive personal data, which in this case includes ethnic and biometrics data. This higher standard of consent-seeking is not defined, and the Guidelines also carve out an exception to the requirement of consent, which includes cases of health emergency. This was the exception relied on by the government to process personal data for the RRR without the consent of the affected citizens.

NASSCO developed the RRR, and assume the responsibility of collecting and aggregating the data of poor and vulnerable households in accordance with established procedures to build a National Social Registry.

Analysis

The Nigerian government, through NASSP, shared its data for the RRR project in two ways. First, it appointed a private company to develop the AI around this project. Second, it shared data with telecommunications companies and banks responsible for profiling and reaching out to eligible beneficiaries. The Nigerian government collaborated with the telecommunications industry to deliver RRR by sharing data of potentially eligible beneficiaries. The industry was seen as "an enabler – able to reach the last mile in a particular location – and can help with identification and location."[73] As part of their operations, the telecommunications industry assisted in the geo-mapping of high-density areas and determined how to reach out to potential beneficiaries (through robocalls, text messaging or auto playbacks).

The NASSP has a data mining protocol in place with other sub-regional social safety net agencies in Nigeria, which requires a formal request to the NASSP and is followed by the signing of a memorandum of understanding with the entities on data usage protocols.[74] While NASSCO assume the responsibility of collecting and aggregating the data of the poor and vulnerable

[73] I Apera ,'Rapid Response Register for COVID-19 Cash Transfer: The Nigerian Response to COVID-19 in urban communities' NASSCO Presentation
[74] Procedure for Access, Mining, and Verification of the NSR by other Social Safety Nets Agencies https://nassp.gov.ng/wp-content/uploads/2021/06/Data-Mining-Protocoal-PROCEDURE-FOR-ACCESS-2020.pdf

households in accordance with established procedures to build a National Social Registry, and it approves the requests for the sharing of this data with other agencies, the MOU provides guidelines on the use of the data and prohibits disclosure to other third parties.[75] However, there is no indication that a similar data-sharing agreement was concluded with the private sector company appointed by the government to develop the AI for the roll out of the RRR.

In the initial design of the RRR, the collected data was to be verified against Nigeria's National Living Standards Survey data, collected alongside Nigeria's last census in 2006. To ensure the picture of social vulnerability was up to date, this would then be cross-checked with available data on vulnerable residents from existing databases of NGOs and other local support groups. Beneficiaries were also meant to be verified through visits from survey officials, and those without bank accounts were to be supported in opening one and mobile money payments were to be made an alternative payment option. However, the government decided to automate the entire process except mobile money payments. To identify poor Nigerians in urban and semi-urban areas, NASSCO obtained the verified list of urban communities from states across Nigeria and ranked 2650 urban and 6149 rural wards according to relative poverty and wealth indices. A total of 20 million people in 1163 wards were identified from the poorest wards across the country, and 43 variables were considered for this identification based on geographical satellite remote sensing technology deployed to locate poor urban wards and high-density settlements.

Enrolment of the urban poor in the RRR was conducted via mobile phone short messaging service technology, which allows residents of identified communities to register in a number of different ways, including by using a mobile phone's USSD codes – short codes that allow non-smartphones to interact with a mobile network's central computer. Using bank verification numbers, a scheme introduced by the Central Bank of Nigeria to give each bank customer a unique identity using biometrics, customers with an average balance of less than 5000 naira in their bank accounts were identified nationwide. Using USSD codes generated by telecommunications companies, these customers were sent automated messages to verify their identity. Ppon doing so, they started receiving direct digital payments from the government for six months.

Nigeria's RRR required high penetration of mobile phone usage among the urban poor. According to NASSCO, there is a high concentration of mobile phone usage among the urban poor because it is a primary means of conducting business. While this may be true, this also assumes that mobile phone users have a level of technological literacy to understand the invitations to be verified for eligibility for the RRR. In addition, NASSCO noted that there were significant levels of reactions from the contacted eligible beneficiaries to the enrolment text messages that included suspicions of fraud and disbelief, which meant many eligible Nigerians did not enrol.[76]

[75] Memorandum of Understanding for Data Sharing Agreement
https://nassp.gov.ng/wp-content/uploads/2021/06/MOU-for-Data-Sharing-Agreement.pdf
[76] I Apera ,'Rapid Response Register for COVID-19 Cash Transfer: The Nigerian Response to COVID-19 in urban communities' NASSCO Presentation

Potential benefits and risks

Benefits

The RRR is a programme that aims to maximise the power of AI, machine learning and big data analysis to connect the government directly to citizens. By cutting out inefficient and bureaucratic administrative agencies in delivering social protection through cash transfers, the government intends to make important cost savings and reduce wasteful expenditure. With the government's long-term objective of using the RRR as a method of cash transfers for future emergencies, Nigeria is on the cusp of changing a narrative about government inefficiency.

Risks

Exclusion of the urban poor

Many Nigerians were potentially excluded from the RRR programme due to the benchmarks introduced in automating the identification of eligible Nigerians. These included targeting only those within the identified location points, owners of a mobile phone and in several cases, bank account owners. This benchmark breeds distrust in Nigerians when there is confusion among potential beneficiaries on how some accessed the programme while others were excluded.

Algorithmic decision-making

The delegation of crucial decision-making in selecting and verifying beneficiaries from human beings to algorithms creates a sense of lack of accountability. Automated decision-making does notalways constitute 'AI', but the decision-making component is of particular interest when considering intersections of new technologies with the government's responsibilities. Many of the risks and harms associated with AI can be seen as deriving from risks associated with underlying data, such as inherent bias or omission, producing solutions that do not work equally well for all citizens., Likewise, the outcomes that may arise from automated decision-making in social protection raise important questions about justice and fairness that should inform policy design.

Privacy

There was limited understanding of the full scope of collection, use and processing of personal data by government officials and the private sector partners in the RRR project. For example, the decision to identify eligible Nigerians through certain criteria meant citizens were being profiled using their personal data without their consent, and created adverse consequences for some, such as those deemed ineligible for social protection.

Conclusion

Scaling up social protection coverage in Nigeria using AI was an effective way for the government to expand its reach to millions of Nigerians quickly and efficiently. However, there is also a cautionary tale here around the absence of human verification. The qualification criteria chosen exposed issues around using automated decision-making in social protection delivery, stemming from both data quality issues, and unequal access to mobile technologies among groups of people who do not fit into a predetermined definition of the urban poor, meaning many eligible citizens were not identified through the analysis.

Recommendations

Data Justice and Digital Inclusion

To facilitate inclusion, and to ensure marginalised groups are not rendered invisible in such an approach, the government needs to work towards first ensuring that "global digital public goods (such as the internet, cybersecurity and data) are more equitably available" before using AI as a solution for social protection.[77]

Transparency

It is important to have transparency of decision-making relating to the datasets needed to deliver the RRR using AI. During this research, we could not get a clear answer on the full spectrum of datasets used in determining the eligibility of potential beneficiaries. Given the acknowledgement in the pilot phase that one of the biggest constraints for the government was the public suspicion about the delivery of the project, it is recommended that a government's public campaign about such a programme should provide more details about the datasets shared with an AI developer. However, it is important to note that different communities have different needs and demand different mechanisms for transparency and accountability. This requires government agencies to be more cognizant of what kinds of information are being made available and how particular audiences may use, rely upon, or gain access to it in the first place.

Strong institutional oversight

The development of the RRR in Nigeria demonstrates how governments can utilise data and work with the private sector to provide tangible benefits for citizens. This case study underscores the need for strong institutional accountability where personal data is collected, used and disclosed. There must be mechanisms in place for holding both government and private sector data controllers and processors to account, and the new data protection commission must be backed by institutional resourcing to enable adequate oversight and enforcement of Nigeria's new data protection law.

Limited scope of processing

The ability of the government to identify and profile Nigerians through new AI capabilities developed by the private sector could lead to an expansion of data collection and surveillance, beyond the original purpose of identifying eligible Nigerians for social protection. Data governance frameworks in Nigeria need to prevent such an expansion, and while the adoption of a data protection law is a starting point, it also requires institutional cultures to change, and for the principle of data minimisation to be employed in delivering effective public services while all respecting the privacy of citizens

[77] GPAI (2022). Data Governance Working Group – A Framework Paper for GPAI's Work on Data Governance 2.0, Report, November 2022, Global Partnership on AI, Paris. https://gpai.ai/projects/data-governance/Data%20Governance%20-%20A%20Framework%20Paper%20for%20GPAI%E2%80%99s%20Work%20on%20Data%20Governance%202.0%20.pdf

3.4 The Aclimate agricultural data platform in Colombia

Location: Colombia
Players: Colombian government, The International Centre for Tropical Agriculture (CIAT) and farmers' collectives

Context

Aclimate[78] is a data commons initiative from Colombia, that exemplifies practical implementations of data sharing in Latin America. Data commons are based on the principle that data should be a collective asset, serving the broader community or the general public,[79] thus promoting equitable access. Within a common pool, data is viewed as a communal asset and shared resource.[80] The objective is to remove obstacles to information sharing, potentially leading to knowledge creation. Embracing this concept, Aclimate aims to empower farmers with the tools for informed decision-making, enabling them to consistently select the best planting options and adeptly respond to the evolving effects of climate change. This initiative views data as a communal asset, aiming to foster knowledge by removing information-sharing barriers. The primary objective is to provide farmers with timely and relevant insights in the context of a bimodal climate pattern, alternating between rainy and dry seasons. This is increasingly unpredictable, attributed to global climate change. The importance of these insights was evident in 2014 when rice farmers in Montería and Cereté averted potential losses amounting to approximately 3.6 million USD. Leveraging scientific guidance from the Aclimate platform, they strategically chose not to sow in the traditional sowing months of April and May.[81]

Aclimate was born from a partnership between the Colombian government, The International Centre for Tropical Agriculture (CIAT) – a non-profit agricultural research organisation – and farmers' collectives which enabled the merging of governmental and farmers' data to create the Aclimate platform, which integrates diverse datasets[82] with tools and machine learning insights that provide farmers with actionable information via a web platform or the Melisa chatbot[83] – a tool designed for farmers to easily interact with the information.

Governance Setting

Colombia's regulations seek to balance individual rights to privacy with the broader aim of transparency and open data initiatives. On one hand, it seeks to protect the rights of its citizens regarding their personal data; on the other, it encourages the free sharing of public data to boost innovation and transparency.[84] However, it is important to highlight that, in the Latin American

[78] The Aclimate Data Platform: https://colombia.aclimate.org/
[79] https://gobiernodigital.mintic.gov.co/692/articles-245714_recurso_2.pdf
[80] https://www.opendemocracy.net/es/loraculos-del-agua-epoca-variabilidad-climatica/
[81] https://gobiernodigital.mintic.gov.co/692/articles-238511_recurso_2.pdf
[82] Coming from weather stations of IDEAM- National Institute of Hydrology, Meteorology and Environmental Studies also, Climate Hazards Infra-Red Precipitation Stations and forecast Sea Surface Temperatures from North-American Multi-Model Ensemble.
[83] Melisa Chatbot
https://alliancebioversityciat.org/es/tools-innovations/melisa-chatbot
[84] The Colombian Political Constitution highlights personal rights, particularly pointing to the right to privacy and the right to amend data. This underscores the nation's commitment to safeguarding personal information and ensuring that individuals have the means to rectify errors (in the ACLIMATE case there were no privacy concerns that were documented). On the other hand, Law 1712 of 2014 promotes the right to access information and establishes provisions for open data. This

region, according to the Global Data Barometer,[85] even though open data policies scored high, data-sharing frameworks lagged.

In this context, several initiatives took place in the last couple of years in Colombia towards the design of a data governance model, in line with the National Data Infrastructure Plan (PNID).[86] Thus, the Data Infrastructure Governance Model (MGID) was proposed for Colombia as a distributed responsibility scheme, where various actors intervene to promote the democratisation of data infrastructure. It remains to be seen, if implemented, how it can encourage and enhance initiatives such as Aclimate.

Analysis

Several years before discussing the idea of a Data Infrastructure Governance Model, the value of data in averting agricultural crises became apparent. In 2013, the Colombian government began exploring options for strengthening growers' associations. With this goal in mind, the Minister of Agriculture signed an agreement with CIAT to "strengthen the capacity of Colombia's agricultural sector to adapt to climate vulnerability action." This agreement[87] includes evaluations of seasonal forecasting, and providing specific recommendations for increasing productivity using public data. CIAT proposed a collaborative approach with Fedearroz, a farmers' union, to merge association-specific data (i.e. annual rice surveys and harvest monitoring details) with government datasets. This was aimed at refining farmers' decisions. Thus, comprehensive statistics from the government and farmers' union databases were leveraged to provide advisories on optimal sowing periods to the Montería and Cereté farmers' collectives. The endeavour's success was recognised globally, receiving an award for the best work on climate change and data from the United Nations[88] in 2014.[89]

CIAT was responsible for conceptualising the project, forging partnerships with growers' associations, and leveraging available agricultural historical data. The Ministry of Agriculture and Rural Development,[90] through National Institute of Hydrology, Meteorology and Environmental Studies (IDEAM), shared national climate data for the development of the platform. Together with the aforementioned contributors, associations of crop growers such as Fedearroz (the Rice Growers Association) were instrumental in developing Aclimate. These organisations played a

suggests that while the country respects the rights related to personal data, it is also advocating for increased transparency in public data to bolster accountability and foster innovation.

[85] Global Data Barometer
https://globaldatabarometer.org/

[86] MinTIC publishes draft of the National Data Infrastructure Plan for Comments
https://mintic.gov.co/portal/inicio/Sala-de-prensa/179710:MinTIC-publica-para-comentarios-borrador-del-Plan-Nacional-de-Infraestructura-de-Datos

[87] ACLIMATE COLOMBIA. Open Data to Improve Agricultural Resiliency.
https://odimpact.org/case-aclimate-colombia.html

[88] https://www.unglobalpulse.org/challenges-hackathons/big-data-climate-challenge-2014/

[89] Using Big Data for Climate Change-Climate Challenge 2014.
https://www.unglobalpulse.org/challenges-hackathons/big-data-climate-challenge-2014/

[90] In an interview in 2016 with the CIAT project lead, conducted by Andrew Young and Stefaan Verhulst, he pointed to the Ministry of Agriculture and Rural Development (MARD) as the primary governmental advocate for Aclímate Colombia. However, while MARD played a vital supportive role, the chief governmental data source was the National Institute of Hydrology, Meteorology, and Environmental Studies (IDEAM) that collects and shares the country's climate data.(Verhulst and Young, 2017).

critical role in connecting the farming community with CIAT's resources, interpreting the data and tools, and ensuring their dissemination to the farmers.

Aclimate offers probable insights into three categories (i.e. below average, average, and above average) for the following six months. These forecasts are refreshed daily, creating a range of scenarios crucial for crop modelling. Translating these probable forecasts into daily weather information involves a specific procedure. This approach entails reviewing historical data and assigning predictions based on probability percentages. The platform integrates diverse datasets such as field-specific yield data with comprehensive records of "cropping events" covering every stage from sowing to reaping to create crop simulation models to provide actionable information[91] for decision-making related to the choice of planting dates and seeds. Much of this data, such as large observational datasets about in situ crop performance from farmers and agricultural organisations, were already accessible in anonymised form but had to be centralised and digitised to be usable for Aclimate Colombia.[92]

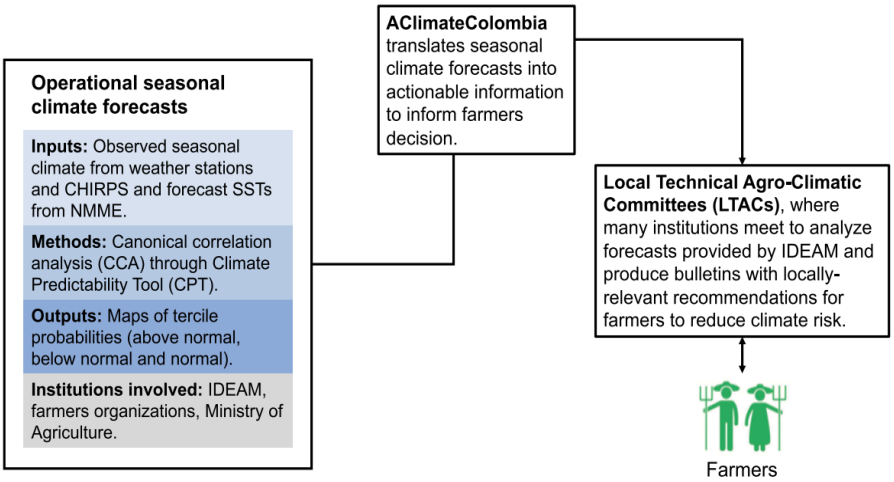

Figure 3: Diagram showing the process of production and use of seasonal agro-climatic forecasts. (Sotelo S, et al. (2020)[93,94,95]

CHIRPS: Climate Hazards Infra-Red Precipitation with Stations; SSTs: Sea Surface Temperatures; NMME: North-American Multi-Model Ensemble; IDEAM: Colombian Meteorological Service.

Additionally, crop growers' associations and Fedearroz (the Rice Growers Association) emerged as pivotal stakeholders. They bridge the gap between CIAT's offerings and the farming community by interpreting and relaying the insights and tools to the farmers.[96] Aclimate is available and accessible via a website and a mobile app. In 2022, Melisa chatbot, a user interface to easily interact with the Aclimate information and forecast was created. This AI-powered tool, compatible with social networks like WhatsApp, Facebook, and Telegram, is primed to deliver precise

[91] Steven Sotelo, Edward Guevara and others (2020). *Pronosticos AClimateColombia*: *A system for the provision of information for climate risk reduction in Colombia.* July 2020, Science Direct.
https://www.sciencedirect.com/science/article/pii/S0168169919315832
[92] ACLIMATE COLOMBIA. Open Data to Improve Agricultural Resiliency:
https://odimpact.org/case-aclimate-colombia.html
[93] CHIRPS: Climate Hazards Infra-Red Precipitation with Stations
[94] NMME: North-American Multi-Model Ensemble
[95] IDEAM: The National Meteorological Service
[96] ACLIMATE COLOMBIA. Open Data to Improve Agricultural Resiliency.
https://odimpact.org/case-aclimate-colombia.html

agro-climatic forecasts, revolutionising decision-making for farmers.[97] To access and use the Melisa chatbot, farmers must search for it via social network platforms or add it as one of their contacts. Once they are in their contacts, they can talk and chat with Melisa whenever needed. There is also a tutorial[98] on how to access and use The Melisa chatbot.

Benefits & Potential Risks

Benefits

The innovation of Melisa chatbot is significant in improving the way farmers access information on crop yield. To avoid the potential unequal access in the long term, it is expected that it can be used in areas with poor internet access In these circumstances, the tool will be adapted to other platforms, such as SMS (text messaging).

Risks

Unequal access

The commendable strides in tech-infused agricultural solutions have beneficiaries spanning the agricultural sector to research domains. Yet, there is a lurking danger. Emphasising technology might inadvertently marginalise the most vulnerable. According to Govlab,[99] "a move toward more technological and data-driven efforts to benefit the agriculture sector risks leaving behind those who need the most support." Adding to the above-mentioned unequal access, Lara Barangé[100] from the CIAT also points to some of the potential challenges and risks for the implementation of AI in the agriculture field. Several concerns arise from the potential reduction in farming diversity and environmental implications as well as loss of power by farmers. AI tools, with their emphasis on efficiency and yield, might inadvertently promote monoculture. Such a bias from AI could hinder efforts to emphasise agrobiodiversity.

Power dynamics

Furthermore, the advent of AI in the agriculture field could potentially bring with it significant shifts in power dynamics both between individuals and within organisations. Ideally, AI could be leveraged to equalise power structures and amplify individual autonomy. This is achieved through enhanced surveillance capabilities and the creation of incentives that compel individuals and entities to integrate AI or risk losing associated benefits. Consequently, this could further marginalise smaller players, including farmers, amplifying their power disparity with large corporations.[101] Overall, the benefits of AI efficiency could be counterbalanced by unexpected environmental consequences and unequal access promoting further unbalanced power relationships.

---

[97] https://www.opendemocracy.net/es/loraculos-del-agua-epoca-variabilidad-climatica/

[98] The Aclimate tutorial: https://www.youtube.com/watch?v=dysPl6vDncE

[99] The GovLab, at NYU, is an action-research centre that is dedicated to improving people's lives and enhancing democracy by revolutionising governance through the use of cutting-edge technologies. More information: https://thegovlab.org/about

[100] Lara Barange (2023). Artificial Intelligence: How could it transform agriculture. August 2023, Alliance Biodiversity & CIAT.
https://alliancebioversityciat.org/stories/artificial-intelligence-agriculture

[101] Robert Sparrow, Mark Howard & Chris Degeling (2021) Managing the risks of artificial intelligence in agriculture, NJAS: Impact in Agricultural and Life Sciences, 93:1, 172-196,
https://www.tandfonline.com/doi/pdf/10.1080/27685241.2021.2008777

Conclusion

Numerous stakeholders, including government agencies, civil society organisations, and private sector entities, hold large datasets. Shared responsibly, this data can greatly benefit the public. However, without standardised data-sharing protocols and regulations, the potential for misuse is real. Data sharing involves granting stakeholders access to data within defined use limitations and controls. The guidelines that direct and facilitate the expanded use of data, including sensitive, proprietary, or non-open data can range from laws and regulations to policies and guidance.[102]

Data-sharing arrangements[103] have become a popular topic with the increasing focus of the agenda on AI related initiatives. However, these guidelines primarily remain at the policy stage rather than being legally binding. Often, these strategies do not address agreements concerning sensitive or personal data.[104] In this context, it is important to clarify that while distinct, data sharing and data protection sometimes overlap, especially concerning personal data. Thus, without frameworks governing these arrangements, there is a risk of overlooking data's positive applications and allowing unchecked misuse. Countries should prioritise making more data accessible, implementing stringent data protection regulations, and safeguarding individuals' privacy. It is important, given the context of this case, as previously mentioned, to highlight that even though open data policies scored high in the region, data-sharing frameworks lagged. Given its importance, addressing complexities around data sharing is vital for maximising data's developmental potential as exemplified in the Aclimate case.

Recommendations

Data Trusts

To address the potential for further marginalisation of smallholding farmers as described earlier, s collaborative approach hs been taken by the Colombian government to work with CIAT and farmers' collective is laying the foundation for a future development of data trusts. These aim to collectively negotiate terms of use with potential data users. This not only re-balances power asymmetries but also enables active participation in the distribution of value that the data creates. This will redefine the role of government not only as a provider of data but also as a steward.

Open data

Voluntary agreements among parties for data sharing should embrace the objectives of open data in terms of democratising access and ensuring that exclusive data-sharing agreements with third parties are limited.

[102] Global Data Barometer Handbook
https://handbook.globaldatabarometer.org/2021/indicators
[103] The Global Data Barometer survey's qualitative findings also highlighted numerous global data-sharing governance strategies.
[104] Law 1712 of 2014 promotes the right to access information and establishes provisions for open data. This suggests that while the country respects the rights related to personal data, it is also advocating for increased transparency in public data to bolster accountability and foster innovation.

Accountability

The interface with government and the very structure of the bureaucracies which administer services are embedded with inherent accountability challenges. Automated decisions – particularly those which are fully automated or engage with subjects through chatbots – can exacerbateunaccountability when decisions are made based on incorrect feedback from the Melisa chatbot. Recourse mechanisms and accountability channels will need to address these issues for users, many of whom are adversely impacted by the results of the automation.

Digital inclusion

For digital developments such as Aclimate to be useful, access to the internet and digital literacy are both necessary. Government initiatives to use AI and other digital tools to accelerate the delivery of developmental priotities requires a deliberate policy shift on the part of the government to ensure equitable access for beneficiaries of the solutions.

4. Synthesis of Case Study Learnings

Public sector AI use can erode public trust

Governments' sharing of public sector data to enable AI is gaining ground across the world. The compiled case studies attest to various initiatives that use government data to develop AI solutions in response to societal challenges. However, much as AI has the potential to improve public sector efficiency and pave the way for innovative public services or private sector-driven innovations, it can also erode public trust if not delivered responsibly. The risk to public trust appears to be particularly high when governments enter into agreements with AI developers and providers outside of public procurement processes, especially when it involves the sharing of sensitive public data. The processing and sharing of data, especially personal data, should be done transparently and in close consultation with data subjects, noting their rights to know who holds their data and for what purpose. Citizens should remain informed about possible data transfers between governments and third parties.

As noted in the NHS case study, the government failed to protect citizens or data subjects, resulting in protests and the deterioration of patients' trust in the NHS.[105] Surprisingly, despite being one of the leading countries in AI adoption and regulation, the UK government in 2015/16 failed to provide oversight to protect its citizens or data subjects. Taiwan's approach to the health passbook, meanwhile, has been successful in building public trust and putting data subjects in the driver's seat in deciding when their health data can be used to inform wider public health decisions. Individual data rights are respected whenever the government intends to share health records with third-party AI companies. Data subjects provide consent on which public sector-held data can be shared and reserve the right to opt-out or ask for the erasure of previously shared data.

Embracing privacy-enhancing technologies that safeguard data subject rights and protect their identities and personal information will go a long way in building trust in AI use.

Data collaboration is important

AI development benefits from public sector and non-state actors' data. Where public sector data can be complemented by other datasets (i.e. from the private sector or researchers), new models for data sharing and governance are developed to enable AI. The flow of data needs to follow appropriate checks, oversight mechanisms, rules and audits. The Aclimate case study is an example of data collaboration between the government, a research institution, and farmers' cooperatives to centrally share data for the development of an AI tool. An ecosystem approach that seeks to integrate existing datasets from various actors can offer complex and rich data that fulfils the needs of AI models. At the same time, questions about power balance or data privacy that may come from any of the data contributors need to be carefully addressed to uphold fair and responsible data sharing.

[105] Natasha Lomas (2016). Concerns raised over broad scope of DeepMind-NHS data-sharing deal. Tech Crunch. https://techcrunch.com/2016/05/04/concerns-raised-over-broad-scope-of-deepmind-nhs-health-data-sharing-deal/

In the Taiwan case study, a tripartite model between the state, data subjects and private sector AI companies paved the way for the government to act as a facilitator for data sharing between the various parties. A similar model exists in Colombia for various actors to pool their data for public benefit which lays the foundation for the idea of data trusts as a data-sharing model to gain more traction for governments.

Responsible public sector data sharing needs to be the norm

Data capturing and sharing bolster automation and offer possibilities for algorithms to be deployed for decision-making. This can lead to biased decisions of exclusion, as noted in the Nigeria case study. Despite the possibilities around deploying AI and machine learning to fill data voids by offering alternative ways to measure poverty and target social protection beneficiaries, fully relying on automated decision-making led to cases of exclusion or inaccurate targeting of beneficiaries. At the same time, in the Nigeria case, existing data protection regulation provided less protection and rights for personal data collected by governments and the private sector (i.e. in this case, telecommunication companies). Undoubtedly, the government's agenda to help the poor is commendable. However, the way through which data was captured, shared and used needs to be done more responsibly and ethically. AI has the potential to drive socio-economic transformation and governments, as major data controllers and processors, can play an important role. At the same time, it needs to be done responsibly and ethically without exposing data subjects to risks, harm and other unintended consequences as a result of algorithmic decision-making.

Automated Decision-Making and Human Oversight

Some data protection frameworks prohibit or limit automated decision-making. However, if decisions are made by algorithms within a public sector context, this may constitute an improper delegation of authority in contexts where administrative justice frameworks exist. When institutions delegate their decision-making powers, the onus of proving, or disproving, why a data subject should be included within a system is unfairly shifted to the data subject. The accountability challenges extend not just to the data collection and retention itself, but also to the algorithms being used. While all four case studies had elements of automated decision-making, the Nigerian case study demonstrates the need for human oversight to address machine errors, exclusions, and the mitigation of adverse consequences.

For governments to maintain human oversight over the AI solutions, public sector bureaucrats need to acquire the required skills to fully scrutinise the potential consequences of AI against public trust and values (i.e. transparency, fairness, equity). Similarly, they need to be trained in specific areas, such as machine learning, data analytics, and automated decision-making to gain a deep understanding of the impact of the data shared with AI developers.[106]

Digital Inequalities and Inclusion

Economic literature is increasingly centralising inequalities not just as an understanding in income disparity, but also across social, political and technological spheres as a challenge to traditional

[106] M. Kuziemski, G. Misuraca, (2020) AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings, Telecommunications Policy, Volume 44, Issue 6.

narratives that centre growth as key goals. This aligns with notions from development economics that focus on the "Capabilities Approach", which widens the normative understandings of wellbeing – and the forms of institutional and agency restraints that can restrain development.[107] In the digital space, these inequalities can be used to frame unequal access to infrastructure and digital services for data sharing and access but are also central to contextualising how the impacts of technologies, both beneficial and harmful, may have differential outcomes.

Data and AI justice

Typically, notions of risks and harms dominate AI literature. However, these case studies demonstrate the need to move beyond just mitigation strategies in data sharing to embracing notions of justice, tackling inequalities and the enhancement of our collective wellbeing. These notions are also important for associating strategies in data sharing to legal and governance solutions. This means embracing data-sharing models such as data trusts and stewardship to address power asymmetries.

Accountability

In a political economy environment, policy effectiveness is scrutinised through an analytical lens: who is accountable for outcomes (intended and not intended), and how is this accountability facilitated?[108]

The question of accountability is a fundamentally important one particularly in those systems which act as a major access point to government services. There is another dimension in data-sharing around accountability:the development and implementation of public sector technologies through private sector service providers. Legal obligations in data protection laws which extend to private actors is one mechanism. However, the contracts between the government and private actors can be a mechanism for extending specific forms of accountability commensurate to the fulfilment of a private actor's obligations. This was present in the UK case study through an information sharing agreement but unclear if one was developed in the Nigerian case study.

Transparency

Access to data and information are crucial for equalising power imbalances. However, the lack of proactive disclosure of data and information was notable in relation to the UK and Nigerian case studies. Data subjects and users of these AI systems require a certain level of access to information to facilitate recourse and accountability.

There is another important imperative for transparency: to engender trust in an AI system, the public needs information and data about that system, about the political and economic dynamics involved in its creation (including procurement), about the outcomes it aims to achieve and processes it uses to do so, and even about the data (and the quality of that data) that underpins it. This is why data subject rights, like those in data protection frameworks, are an important component of creating a trusted environment for the implementation of data-sharing models.

[107] G. Razzano, A. Beyleveld, F. Adeleke (2021) " Automated Decisions and the Public Sector in Africa" Luminate Report
[108] Ibid.

Advancing an Economic Development Agenda

At a policy level, all four case studies sought to advance an economic development agenda. However, while governments' policies may reflect their developmental priorities, the primary purpose for data sharing is to achieve efficiency. Within government, efficiency and cost gains can result in gains for data subjects. However, ascertaining the actual efficiency of – or cost gains in – the implementation of these AI systems remains challenging. This is in part due to transparency challenges around the entire ecosystem. The Nigerian study is a strong example of whether the new AI system reduced corruption in social protection distribution – a key challenge that previous initiatives have had to contend with.

Regulatory certainty

All four case studies show laws and regulations need to extend obligations to both public and private sector actors. The way this can be facilitated will depend on the legal structures of the country concerned. It is clear, though, that there are avenues for developing existing frameworks to enable interoperability, cross-border flow of data, andflexibility in the legal basis for the processing of personal data.

Robust Public Procurement

The case studies show how strategic partnerships between government and the private sector can deliver the developmental priorities of governments. However, underlying such partnerships is the need for robust and transparent public procurement processes that enable the selection of the right partner to take custody of government data. A procurement checklist can be developed that facilitates decisions being made at the right level for the at risks involved, and then highlights what requirements may be needed in the procurement of private sector partners.

Institutional Oversight

All four case studies show the need for strong institutional oversight to enforce laws and regulations, and prevent the skirting of obligations by the players. Without the presence of these oversight institutions, implementation of the regulatory framework would likely be flagrantly ignored as demonstrated in the UK case study. Further research as part of this project will explore institutional capacity gaps in relation to the sharing of government data with AI developers, and discuss the particular role oversight institutions can and should play in safeguarding data transfers in such instances.

5. Key Enablers for Government Data Sharing

The provision of government held data for AI development requires three key enablers: a regulatory landscape which helps ensure equitable access to data by third-party AI developers; a policy landscape that facilitates the sharing of data within and outside of government, and; the availability of infrastructure to facilitate the accessibility of government held data.

5.1. Legal Landscape

While there are several legal aspects that intersect with data-sharing models and frameworks, there are five key areas that enable government data sharing.

5.1.1. Antitrust

Access to data holds high economic value for AI firms and gives them a competitive advantage in the development of products and services for economic gain.[109] It enables large AI developers to pursue anti-competitive practices that lock out small competitors from innovation, and which entrench their market power and dominance.[110] Recent examples of lobbying by large data holders wanting to block third party sharing of personal data shows the need for equitable approaches to sharing government data.[111]

In the UK case study, the selection of Google DeepMind as the exclusive government partner, without a transparent public procurement process, raises questions around equitable access to data by AI developers and the role of government in enabling market dominance in the creation of AI systems over other competitors. However, new data protection laws are addressing the effect of data being concentrated in the hands of a few market players by demanding interoperability.[112] Interoperability enables collaboration in data use to mitigate the effect of concentrating data in one firm.

Other measures, such as investments in digital infrastructure to enable competitive access, also equalise the field. In low and middle income countries, where public resources such as access to reliable energy and high speed Internet can be expensive even before investments in data processing, responsive regulation is needed to help new entrants to the market.

For AI systems, data is a critical input. Firms with exclusive access to government data can solidify their advantage. As the UK case study exemplifies, where scope of data use is vague, access to such large datasets can tempt the processing of datasets beyond the original purpose of data sharing to entrench market power. To address this, governments can use competition regulation to either penalise anti-competitive practices or to create an environment that is human-centred and protects smaller competitors.

---

[109] Thomas Tombal, 'Data Protection and Competition Law: Friends or Foes regarding Data Sharing?', 2021 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3826325#:~:text=Thomas%20Tombal,-Tilburg%20University%20%2D%20Tilburg&text=Consequently%2C%20an%20increasing%20call%20for,mixed%20in%20the%20same%20dataset.

[110] Robert Walters, Bruno Zeller, Leon Trakman, 'Personal Data Law and Competition Law - Where is it Heading', 2018 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3275832

[111] Thomas Tombal, 'Data Protection and Competition Law: Friends or Foes regarding Data Sharing?', 2021 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3826325#:~:text=Thomas%20Tombal,-Tilburg%20University%20%2D%20Tilburg&text=Consequently%2C%20an%20increasing%20call%20for,mixed%20in%20the%20same%20dataset.

[112] Jonathan Klaaren, 'Competition Policy and Data Protection in Africa', Policy Brief 5 (Mandela Institute, 2021) https://www.wits.ac.za/media/wits-university/faculties-and-schools/commerce-law-and-management/research-entities/mandela-institute/documents/research-publications/800427%20PB5%20Competition%20policy%20and%20data%20protection_REV%20Dec2021.pdf

To complement responsive regulation, collaboration in transnational enforcement is also necessary given the cross-border operations of most multinational enterprises. This will require cross-country harmonisation of regulation[113] to address gaps in competition enforcement.

5.1.2. Cross-border data flows

There is an increasing trend of countries using data protection laws to express a stringent vision of data sovereignty, such as the imposition of rigid adequacy requirements before data can be transferred to other countries for processing. IIn some jurisdictions, there are cross-border data restrictions that apply only to certain sectors.[114] Given the volume of data that is generated and transferred in the processing of AI systems, the application of data flow-restriction measures within the existing data ecosystem will prove challenging, particularly if there is not sufficient data specificity. In the regulation of data transfers across borders, it is important to distinguish between personal and non-personal data, between sensitive and non-sensitive data, anonymised or pseudonymised data, public or private data, stored or real-time data, open or proprietary data. The Colombia case study is an excellent example of sharing non-personal data in AI systems, including for that data to be shared for use outside Colombia.

There are non-economic arguments for regulation to enable the free flow of data. The sharing of government data for AI development can allow governments to make evidence-based decisions to improve social and welfare spending, and the delivery of public services to its citizens. As van der Berg has noted, “data’s value is maximised when it can flow with trust and permission across companies, sectors, and national borders to be used. That trusted and permissioned flow, with economic and legal frameworks to ensure safety, security, and equal access to opportunity, should be the goal of data policy.”[115] A presumption toward data flows rather than data restriction, in contexts of unspecified data (i.e. not personal or restricted data), is also consistent with presumptions of openness as expressed in many countries’ open government information regimes.

The global political economy suggests that given the concentration of dominant AI companies in the global north, LMICs should assert their data sovereignty through innovative mechanisms in sharing data equitably to maximise public benefits. However, the focus on where data is stored in some data protection laws does not enable deep dives in regulatory approaches on when and how governments should make data available to third parties.

---

[113] International laws set a framework for limiting anti-competitive practices by large data holders. Article 101 (1) of the Treaty of the Functioning of the European Union prohibits agreements or practices by large market players which have the effect of “preventing, restricting, or distorting competition” by small competitors. Practices by large data holders of refusing to share data with small competitors fall within the purview of these provisions and constitute anti-competitive behaviour. The existing regional competition authorities in Africa such as the Common Market for Eastern and Southern Africa Competition Commission, the Economic Community of West African States Competition Authority and the East African Community Competition Authority can serve as the basis for integrating competition policy across the African continent.

[114] The GDPR popularised the term ‘adequacy of data protection laws.’ The GDPR’s criteria for assessing the adequacy of the data protection laws of other countries is based on three factors. First, there must be the presence of rule of law, including respect for human rights and fundamental freedoms, as well as sectoral laws and their implementation. Second, there must be the existence and effective functioning of one or more independent supervisory authorities in the third country. Third, such countries must have international commitments through accession to binding international conventions.

[115] Shanelle van der Berg, (2021). ‘Data Protection in South Africa: The Potential Impact of Data Localisation on South Africa's Project of Sustainable Development (Policy Brief 2)’ https://www.wits.ac.za/media/wits-university/faculties-and-schools/commerce-law-and-management/research-entities/mandela-institute/documents/research-publications/800429%20PB2%20Data%20localisation%20and%20sustainable%20dev_REV%20Dec2021.pdf

One way of addressing the regulation of cross-border data flows is through transnational regulation and the adoption of the Cross-Border Privacy Rules (CBPR) model of the Asia-Pacific Economic Partnership (APEC).[116] This is an accountability-based mechanism that enables data flows through compliance with data privacy rules such as notice, consent, data minimization, defined uses of personal information, and preventing harm. However, each member country is responsible for enforcement. Companies accessing government data will be subject to audits and certifications from APEC-approved accountability agents in the same manner the Taiwan case study had a pre-approved mechanism for third parties accessing government-held data. This approach can serve as a global model for data governance.[117]

5.1.3. Intellectual Property

Consideration and protection of intellectual property (IP) rights is important in sharing government data. The data stewarded by governments may contain unique innovation in the compilation of datasets, and developmental processes of products and services. When such datasets are shared unilaterally by governments – even when they hold ownership over the datasets – it can have broader implications for copyright and trade secrets.

Databases are key in facilitating data access and re-use and are protected under intellectual property laws. International frameworks on IP such as the *Agreement on Trade-Related Aspects of Intellectual Property Rights* protects "compilations of data which by reason of their selection or arrangement of their contents constitute intellectual creations."[118] The protection afforded by these provisions require originality by the creators of the databases (i.e., databases constituting intellectual creations). This means that databases that do not meet the threshold of originality under copyright cannot be protected under these provisions.

Addressing who should hold IP rights for AI projects where governments have made their data available and where some public investment has been made is also important. The value created by a government's role in AI inventions should be recognised and adequately rewarded. Consequently, governments need to ensure the protection of IP rights when sharing data with third parties in a manner that promotes data partnerships with mutual financial benefits.

Licensing – and particularly forms of creative commons or other sharing licensing – is an important facilitating mechanism for governments as a prior step to sharing. To facilitate this, strong data governance practices are required.

5.1.4. Data Protection

Data protection laws have naturally become the default regulatory tool for AI systems as countries scramble to adopt AI strategies that will inform their future oversight model, given the foundational importance of data to the training, and ultimate functioning, of AI technologies. The UK and Nigeria case studies show the importance of robust data protection laws in protecting the rights of data subjects, and also of granting enforceable powers to independent oversight institutions to regulate AI systems. Data subject rights implicated in data sharing, such as the right to consent to the processing of data by third parties, to opt out of processing, and to data portability, are all embedded

[116] What is the Cross-Border Privacy Rules System?
https://www.apec.org/about-us/about-apec/fact-sheets/what-is-the-cross-border-privacy-rules-system
[117] There are currently nine participating APEC CBPR system economies: the US, Mexico, Japan, Canada, Singapore, the Republic of Korea, Australia, Taiwan, and the Philippines.
[118] Article 10(2), Agreement on Trade Related Aspects of Intellectual Property Rights
https://www.wto.org/english/docs_e/legal_e/27-trips.pdf

within data protection laws, although some exemptions may exist, such as use by governments or trustworthy entities for research, statistical purposes, or archives. Further, the legal basis for the processing of personal data will often be found in applicable data protection laws. The legal bases for governments to share data include prior consent of data subjects[119], interests of the data subject[120], public interest[121], the legitimate interest of the government in sharing data[122] and through existing legal mandates in other applicable laws.[123]

5.1.5. Access to Information

The right of access to information is now a well-established right that is recognised in several national laws and in international instruments. The main object of the Right to Information (RTI) is to provide access to government-held and government-produced information and, by definition, data is a raw element necessary to produce information. RTI laws cover adaptable/reusable data (the latter being the main object of Open Government Data) within their scope. There is a presumption of openness in government with the recognition of the right of access to information. Openness and transparency are not synonymous, although transparency is a *sine qua non* condition for the effective openness of any government.

One way of promoting the openness principle is through open data platforms. Open data is an important data-sharing model that can enable equitable data sharing of government data. To advance useful open data practices, Adams and Adeleke have noted that "government regulation is required to determine disclosure priorities, in terms of the creation, collection, organisation, analysis, publication and utilisation of information that prioritise the ability of citizens to exercise and protect their rights."[124] The idea of openness within the open data context implies the possibility of action, unlike the concept of transparency. In the case of a transparent government, information is transferred unidirectionally, whereas with openness, the establishment of a dialogue is presumed.

Access to information laws play a critical role in promoting openness in government data sharing and developing mechanisms for enhancing transparency. They provide a channel through which individuals can access government-held information such as agreements signed by governments

[119] The definition of consent covers three distinct elements: (a) the nature of the consent – freely given, specific, informed, and unambiguous, (b) the way the consent is given – by a statement or clear affirmative action, and (c) the purpose for which the consent is given – the processing of personal data. The requirement for consent is subject to exceptions provided under the applicable law. For instance, the Personal Data Protection Act of Argentina provides for instances when consent to process will not be required. This includes when the data is obtained from unrestricted publicly accessible sources or anonymised. In addition, for the consent to be valid, it is important for such consent to remain in force and not withdrawn.

[120] Under the UK and Nigerian data protection laws, processing to protect the interest of a natural person other than the data subject may also serve as a legal basis for data processing. See Article 6(1)(d), UK GDPR and Section 25(b)(iii) of the Nigerian Data Protection Act. A key issue to consider is whether a government agency should unilaterally decide the interests of a data subject. The UK and Nigerian case studies discussed in this report involved the processing of personal data to advance the interests of data subjects but also reveal the need to combine such legal basis with another layer of accountability by government parties.

[121] What constitutes "public interest" may vary from one legislation to another. Under the UK GDPR, activities necessary for the administration of justice or activities supporting or promoting democratic engagement will be classified as falling within the public interest.

[122] Data processing may be lawfully permitted if it is required to protect the legitimate interest of the data controller or a third party, if there is no overriding interest of the data subject's fundamental right or freedom, particularly where the data subject is a child. This legal basis is not applicable where data processing is carried out by a public authority in the performance of its tasks. While this legal basis for data processing is less common, the data protection laws of certain jurisdictions such as the UK permit public authorities to rely on it.

[123] In Article 19 of the Taiwanese PDPA, if the collection and use of personal data is specifically stipulated under a law or a regulation, such a law or regulation can be the legal basis to collect, process, and use the relevant personal data.

[124] Rachel Adams, Fola Adeleke (2016), 'Assessing the potential role of open data in South African environmental management', The African Journal of Information and Communication 19, 79–99. https://ajic.wits.ac.za/article/view/13605

and private actors involving data sharing, and procurement documents detailing contract awards. Some also mandate, or at least provide, significant guidance on the data required for proactive disclosure. Such disclosure enables individuals to know the type of data held about them, the measures being put in place to protect the data, the way the data is being used, the storage period, and the parties receiving and processing their data.

5.2. Policy Landscape

Models for data sharing need to be driven through both technical and policy processes, whilst being simultaneously supported by governance that incentivises sharing. Policy is a key mechanism for the advancement of data sharing, particularly in the context of the public sector. Policy provides the landscape for public sector actors to implement governance – and is in fact central to ultimately facilitating technical infrastructure.[125] Reflecting on the UK case study, sufficient guidelines and protocol should exist within an entity so that decision-makers are better able to understand the essential actors within the data ecosystem – such as data protection regulators – and the mechanisms for incorporating them within data activities. And policy is important for ensuring technology development (including the necessary data technologies) are consistent with a country's unique policy priorities, helping to support sufficient and effective investment.[126]

The African Union Data Policy Framework provides a high-level, thematic policy framework across the following areas: 1) Data Ownership, Control and Access; 2) Data Safety and Interference; and 3) Data-driven Value Creation.[127] Yet, within these overarching data policy frameworks, specific consideration should be given to essential components of policy for advancing the government's role as a data provider.

When considering policy approaches, it should be highlighted that data sharing is necessary not just for facilitating government-to-public data access, but also for facilitating intra-government sharing to advance the value of government data pools,[128] and to create mechanisms that also allow the government itself to benefit from receiving shared data from civil society and private sector actors.[129]

Key to any policy exercise is the understanding of data sharing as a "unique form of institutional interrelationship".[130] Thus, using policy to advance roles and responsibilities (which include permission levels), should be viewed as a priority. In this way, policy can be seen as a mechanism for driving internal accountability for data governance.

---

[125] Fitria Wahyuni and Rachma Fitriati, 'Why Is the Application Programming Interface the Backbone of a Smart City?', in *Journal of Physics: Conference Series*, vol. 1783 (IOP Publishing, 2021), 012029. https://iopscience.iop.org/article/10.1088/1742-6596/1783/1/012029

[126] Gang-Hoon Kim, Silvana Trimi, and Ji-Hyong Chung, 'Big-Data Applications in the Government Sector', *Communications of the ACM* 57, no. 3 (2014): 78–85. https://dl.acm.org/doi/10.1145/2500873

[127] African Union, 'AU Data Policy Framework', 2022, https://au.int/en/documents/20220728/au-data-policy-framework

[128] Gang-Hoon Kim, Silvana Trimi, and Ji-Hyong Chung, 'Big-Data Applications in the Government Sector', ACM Digital Library Volume 57 Issue 3 2014 https://dl.acm.org/doi/10.1145/2500873

[129] Bertin Martens and Néstor Duch-Brown, 'The Economics of Business-to-Government Data Sharing', 2020 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3540122

[130] David Tulloch and Francis Harvey, 'When Data Sharing Becomes Institutionalised: Best Practices in Local Government Geographic Information Relationships', *URISA Journal* 19, no. 2 (2007): 51–59 https://www.academia.edu/22303604/When_data_Sharing_Becomes_Institutionalized_Best_practices_in_local_government_geographic_Information_relationships

Interoperability remains a manifestly vital objective for public sector data policy. Open government data literature, in particular, has frequently cited the challenges of interoperability.[131] And facilitating interoperability means adopting clear data standards where relevant, which should include open data standards[132], in order to facilitate ultimate access activities like Application Programming Interfaces (APIs), portals, etc.

Specific technical elements of data infrastructure may come with their own specific policy needs, considering for instance APIs,[133] cloud servers, data servers and warehouses, etc., including associated digital infrastructure such as internet accessibility. There may even be software-associated policies, given the centrality of software to data storage and processing. Here,the use of open source software for public technology should be highlighted.[134]

But perhaps the most important component of policy is how it facilitates decision-making by public sector actors and implementers when sharing and protecting data in their everyday practice.[135] When the government is appreciating its role as data custodian, facilitating custodianship should become policy driven.

5.3 Infrastructure development

Ensuring accessibility, privacy and security are fundamentally technical questions. Investing in the infrastructure that facilitates data sharing is, ultimately, an investment in innovation. Data infrastructure can be narrowly understood as including the infrastructure necessary for the storing, processing, integrating, managing, accessing, securing, and analysing of data. But this is embedded with the need for the broader digital and IT infrastructures for the advancement of digital government. As the AU Data Policy Framework highlights, “data infrastructure that enables an integrated data system is a key strategic asset for countries, but the scale, extent and speed of change brought about by data-driven digital technologies make regulation complex and resource intensive.”[136]

Data storage and data management infrastructure should be responsive to the data concerned. For example, the warehousing needs of structured data, semi-structured data and unstructured data are different.[137] Technical choices must centre data needs – and specifically in this context, the ultimate

---

[131] Theresa A. Pardo, Taewoo Nam, and G. Brian Burke, ‘E-Government Interoperability: Interaction of Policy, Management, and Technology Dimensions’, *Social Science Computer Review* 30, no. 1 (2012): 7–23; Thomas Lodato, Emma French, and Jennifer Clark, ‘Open Government Data in the Smart City: Interoperability, Urban Knowledge, and Linking Legacy Systems’, *Journal of Urban Affairs* 43, no. 4 (2021): 586–600.

[132] George Kovacs et al., ‘Open Source Software and Open Data Standards in Public Administration’, in *Second IEEE International Conference on Computational Cybernetics, 2004. ICCC 2004.*, 2004, 421–28, https://doi.org/10.1109/ICCCYB.2004.1437766.

[133] Mark Boyd et al., ‘An Application Programming Interface (API) Framework for Digital Government’, *European Commission, Joint Research Centre*, 2020 https://op.europa.eu/en/publication-detail/-/publication/0e262d9b-ca32-11ea-adf7-01aa75ed71a1

[134] Kovacs et al., ‘Open Source Software and Open Data Standards in Public Administration’, 2004 https://www.researchgate.net/publication/4147504_Open_source_software_and_open_data_standards_in_public_administration

[135] Silja Eckartz, Wout Hofman, and Anne Fleur Van Veenstra, ‘A Decision Model for Data Sharing’, in *Electronic Government: 13th IFIP WG 8.5 International Conference, EGOV 2014, Dublin, Ireland, September 1-3, 2014. Proceedings 13* (Springer, 2014), 253–64 https://www.researchgate.net/publication/310770139_Electronic_Government_13th_IFIPWG_85_International_Conference_EGOV_2014_Dublin_Ireland_September_1-3_2014

[136] African Union, ‘AU Data Policy Framework’ https://au.int/sites/default/files/documents/42078-doc-AU-DATA-POLICY-FRAMEWORK-ENG1.pdf

[137] Gang-Hoon Kim, Silvana Trimi, and Ji-Hyong Chung, ‘Big-Data Applications in the Government Sector’, ACM Digital Library Volume 57 Issue 3 2014 https://dl.acm.org/doi/10.1145/2500873

accessibility needs (i.e. how data will be shared from its storage). Whilst some key features of public sector data have been described as 'silos and security', 'variety and velocity' will also become increasingly relevant. This is because velocity refers to the "speed data is generated, delivered, and processed", aspects that will be central for advancing real-time data as demand foreseeably grows.

Servers, whether local or cloud-based, are a significant feature of data infrastructure. Data governance questions around sovereignty, or even localisation, often feature heavily in discussions in this area. Fundamentally, data sharing – whether within a country, between a government and its partners, or outside the country – should prioritise security equivalencies between entities as the primary concern, whilst bearing in mind cost.[138] Usage of cloud computing minimises expenses by eliminating the need for hardware, software storage facilities, energy, and maintenance costs.[139] Its virtuality and lack of centrality enables multiple users in different locations in an organisation to access and use it.[140] This is because both servers and other infrastructure should be subject to continuous monitoring and system testing to be truly effective. Responsibilities for these tasks should consider the human capacity necessary to properly perform them.

In terms of access, there are a variety of ways that data might be accessed and shared. In the realm of big data, APIs are a vital accessibility feature. APIs are "...the connective nodes of digital components and thus instrumental enablers of this integration".[141] They are foundational for facilitating integrated systems, and for facilitating interoperability.[142]

Quality provision of data is reliant on many factors, not least of all investment into metadata,[143] with additional demand-side features including accuracy, completeness, validity, consistency, uniqueness, timeliness, and fitness for purpose.[144] Quality is relevant not simply as a standards issue, but because it affects the ability of users to garner productive use from the data concerned. Sharing may also be provided through different analytics and dashboards, i.e. through software aimed at generating meaning from the data for audiences. Generating meaning from data through analytics will often require the utilisation, in other words, of additional software or even using third-party software for the data management itself.

All hardware and software options considered in relation to data infrastructure necessary for sharing must, however, be considered in the context of what is available, and the context of the eventual users – whether government, private sector or civil society. This means acknowledging issues such as Internet access, primary devices, and digital capacities. Investments in data infrastructure will be judged against their effectiveness, and thus must be contextually responsive.

---

[138] Gabriella Razzano, 'Data Localisation in South Africa: Missteps in the Valuing of Data', Policy Brief 6 (Mandela Institute, 2021), https://www.wits.ac.za/media/wits-university/faculties-and-schools/commerce-law-and-management/research-entities/mandela-institute/documents/research-publications/800482%20PB6%20Missteps%20in%20valuing%20data_REV%20Dec2021.pdf.

[139] African Union, 'AU Data Policy Framework', 2022, https://au.int/en/documents/20220728/au-data-policy-framework

[140] African Union, 'AU Data Policy Framework', 2022, https://au.int/en/documents/20220728/au-data-policy-framework

[141] Mark Boyd et al., 'An Application Programming Interface (API) Framework for Digital Government', *European Commission, Joint Research Centre*, 2020 https://op.europa.eu/en/publication-detail/-/publication/0e262d9b-ca32-11ea-adf7-01aa75ed71a1

[142] Fitria Wahyuni and Rachma Fitriati, 'Why Is the Application Programming Interface the Backbone of a Smart City?', in *Journal of Physics: Conference Series*, vol. 1783 (IOP Publishing, 2021), 012029. https://iopscience.iop.org/article/10.1088/1742-6596/1783/1/012029
Mark Boyd et al., 'An Application Programming Interface (API) Framework for Digital Government', *European Commission, Joint Research Centre*, 2020 https://op.europa.eu/en/publication-detail/-/publication/0e262d9b-ca32-11ea-adf7-01aa75ed71a1

[143] Umbrich, Neumaier, and Polleres.

[144] IBM, 'What Is Data Quality?', IBM, accessed 28 November 2023, https://www.ibm.com/topics/data-quality.

6. Mitigating Barriers to Accessing Government Data

The case studies in this report show that various challenges in government data sharing may arise depending on the unique context of each country and AI project. However, there are five common barriers that deserve closer attention: use of personal data to train AI systems; regulatory design; value sharing; technical capacity; and public attitudes. The recommendations below address possible ways of mitigating these barriers.

6.1. Privacy-Enhancing Technologies

The exploitation of large datasets for AI development continues to attract privacy and human rights concerns, especially when the data has personally identifiable or sensitive attributes. Recent developments have proven that data can cause harm, human rights abuse, and other unintended negative consequences if not handled responsibly or ethically. Privacy-enhancing technologies (PETs) are an important means for governments to responsibly share data for AI development. Techniques such as anonymisation, differential privacy, or homomorphic encryption, sometimes in conjunction with federated learning,[145] enable the protection of sensitive information while still extracting valuable insights from datasets.

Anonymization serves as a crucial privacy-enhancing technology by removing personally identifiable information from datasets, ensuring that individuals remain unidentifiable. This method protects privacy by reducing the risk of personal data exposure and facilitates the safe sharing of data for analysis without compromising confidentiality. Effective anonymization helps organisations comply with privacy regulations, such as the GDPR, by mitigating legal and security risks associated with handling personal data.

Synthetic data generation[146] also offers a promising solution, providing realistic yet privacy-preserving data for collaborative AI research and innovation. Governments can strike a crucial balance between data sharing for societal benefits such as improving public services and safeguarding citizens' privacy, thus fostering trust and transparency in the use of AI technologies. Synthetic data is created using algorithms to clone or mimic real data, to replicate its basic properties despite being fabricated.[147] Whilst there are significant policy challenges in the realm of data, it is worth highlighting that synthetic data should not be conflated with the utility present in "dummy data" for development practices.[148]

Data can be fully or partially synthetic. For the first category, as the name suggests, the data is fully synthetic. This means that all the data variables (i.e., categorical, numerical) are synthesised and, in theory, not identifiable (although there may be individual synthetic data points that match real-world

---

145

Federated learning is an innovative approach to machine learning that allows multiple entities to contribute to a collective learning process without sharing rawdata.Take for example a consortium of hospitals. Instead of centralizing raw data, the learning algorithm travels to each location, learns from the data on-site, and only the updated model parameters are shared centrally. This iterative process continues until the algorithm is effectively trained, ensuring that raw personal data remains at its original location.

146 Synthetic Data Generation: Definition, Types, Techniques, and Tools: https://www.turing.com/kb/synthetic-data-generation-techniques

147 Goldstein, R., Woolley, M.E., Stapleton, L.M., Bonnéry, D., Lachowicz, M., Shaw, T.V., Henneberger, A.K., Johnson, T.L. and Feng, Y., 2020, 'Expanding MLDS Data Access and Research Capacity with Synthetic Data Sets' *Maryland Longitudinal Data System Center, Baltimore, MD* https://mldscenter.maryland.gov/egov/Publications/ResearchReports/SDPReportFINAL.pdf

148 Jono Bosman, 'Supercharging Google Sheets with the Power of ChatGPT', OpenUp Blog, 7 March 2023, https://openup.org.za/blog/supercharging-google-sheets-with-the-power-of-chatgpt

ones), which allows for confidentiality to be strongly maintained whenever the data is used. In contrast, the second category of synthetic data is partially synthetic: non-sensitive variables are unchanged with respect to the source.

Data-sharing requests to feed algorithms, train AI models, or fill data analysis models are unprecedented in today's digital economy. The global firm Gartner claimed that synthetic data might overtake actual data in training AI models by 2030.[149] Moreover, Gartner posits that synthetic data will continue to be produced and used, thanks to computer simulations and AI generative models. Synthetic data is poised to facilitate more widespread AI model training by being a substitute for the hard-to-get, expensive, sensitive and scattered real-world data; but it does often depend on real-world data, at least initially. Synthetic data attracts a lot of interest because it helps navigate the complex issues around real data availability, accessibility and compliance requirements. In critical sectors of the economy (i.e., health, financial systems, etc.) where data has personally identifiable information and is sensitive, synthetic data is emerging as being a viable complementary option alongside real data for early-stage testing, and deployment of AI technologies.

Using synthetic data comes with risks and must be pursued with precautions. The European Commission has developed an analysis of the positive and foreseeable negative impacts of synthetic data on data protection.[150]

Negative foreseen impacts	Positive foreseen impacts
Output control could be complex: A way to ensure the output is accurate and consistent is by comparing synthetic data with original or human-annotated data. However, access to the original data is required for this comparison.	***Enhancing privacy in technologies***: Through a privacy by design approach, this technology could provide an added value by not disclosing the personal data of subjects. [note: this is also offered by other PETs]
Difficulty mapping outliers: Synthetic data can only mimic real-world data; it is not a replica. Therefore, some synthetic data generation methods may not cover some outliers that the original data has. However, outliers in the data can be more important than regular data points for some applications, but the downside is that these may correspond to identifiable individuals.	***Improved fairness***: Synthetic data might contribute to mitigating bias by using *fair synthetic datasets* to train artificial intelligence models. These datasets are manipulated to better represent the world (to be less as it is and more as society would like it to be).

[149] Rob Toews, 'Synthetic Data Is About To Transform Artificial Intelligence', *Forbes* 2022 https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/?sh=37b04bfe7523

[150] Synthetic Data: https://edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en

Negative foreseen impacts	Positive foreseen impacts
***The quality of the model depends on the data source*:** The quality of synthetic data is highly correlated with the quality of the original data and the data generation model. Synthetic data may reflect the biases in the original data. However, manipulating datasets to create *fair synthetic datasets* might result in inaccurate data.	

Table 2: Foreseen positive and negative impacts of synthetic data

6.2. Regulatory sandboxes

The proliferation of emerging digital technologies has introduced novel risks. New, disruptive, opaque AI tools are created daily, and regulators struggle to maintain oversight. Governments are adopting anticipatory regulatory instruments known as regulatory sandboxes to support innovation, harness the potential of technology, and control possible adverse effects or risks.[151] Regulatory sandboxes for AI offer an environment that allows the development, testing and validation of AI models and innovations before their deployment to the market, or before they start being used at large scale. Such regulatory landscapes have also been proposed (and explored) for more specific regulatory contexts, like data protection. This environment is generally controlled and serves as a safe space or testbed for regulatory bodies to test and scrutinise AI innovations to validate them and reduce the time to market or internal adoption.[152]

Although generally housed and mainly operated by state regulatory agencies or data protection offices, the sandboxes must be run following multi-stakeholder and multi-disciplinary approaches across public national agencies and coordinating with market actors.[153] This is required to meet AI innovations' technical, social and economic specificities.[154] Public sector bodies in charge of the sandboxes need the necessary AI expertise to support optimal testing that results in maximum benefits for the market, regulators and end users. They also need to be able to assess market dynamics and behaviours through emerging technologies and thus support competition and innovation via regulatory sandbox frameworks.[155] Where cross-border data flow is necessary to enable AI development, regulatory sandboxes can be leveraged for experimentation, paving the way for trust and bolstering international cooperation between regulators, innovators and countries.[156]

---

[151] Deborah Morgan, 2023, August. Anticipatory regulatory instruments for AI systems: A comparative study of regulatory sandbox schemes. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 980-981) https://dl.acm.org/doi/10.1145/3600211.3604732

[152] Katerina Yordanova (2022), The EU AI Act–Balancing human rights and innovation through regulatory sandboxes and standardisation. *Competition Policy International* https://www.competitionpolicyinternational.com/wp-content/uploads/2022/03/TechREG-Artificial-Intelligence-March-2022.pdf

[153] Regulatory Sandboxes in AI, OECD Digital Economy Papers: https://www.oecd-ilibrary.org/docserver/8f80a0e6-en.pdf?expires=1700740909&id=id&accname=guest&checksum=41485045F123CE751B8D66F026759D44

[154] Government of the United Kingdom (2021), National AI Strategy, https://www.gov.uk/government/publications/national-ai-strategy.

[155] Andy Chen. (2019), "Regulatory Sandbox and Competition of Financial Technologies in Taiwan", Competition Policy International. https://www.competitionpolicyinternational.com/wp-content/uploads/2019/01/Asia-Column-February-2019-Full.pdf

[156] BIAC (2020), Regulatory Sandboxes for Privacy Analytical Report, https://biac.org/wpcontent/uploads/2021/02/Final-Business-at-OECD-Analytical-Paper-Regulatory-Sandboxesfor-Privacy.pdf

6.3. Fair Financial Models

The responsible sharing of public sector data to advance AI development requires fair financial models. The economic value of data sharing is dependent on the ability of third-party AI developers to process them in specific ways to generate value. This means the value of data differs from one third party to the next. The nature of the applicable data that is being shared by a government also determines the value of the data. The structured nature of data and the size are relevant factors in the valuation of data. In all the analysed case studies, data was purportedly being shared in the interests of the public. However, there is also a monetary value that accrues to third parties who can process government data to create financial gain. Negotiating a fair share of the monetary value accruable to governments and its business stakeholders in a data partnership can be challenging. Public sector data can be fragmented, incomplete, incoherent, and inaccessible in reusable formats. This can undermine its value and the possibility of using it for secondary purposes such as training AI models and third parties may have to invest in raising the quality of the data itself before it can generate value.

Governments should consider adopting a principles-based approach towards deriving monetary value from the outcomes of sharing data with third-party AI developers in commercial and non-commercial data partnerships.[157] This approach must nevertheless have public purpose at the centre of the value creation process and the equitable distribution of the returns.[158] Governments can achieve this by creating data partnerships with conditionalities that can influence the behaviour of third parties, improve outcomes and ensure fair financial value sharing in the future.

6.4 Skill and capacity gaps

The public sector can only enable AI if it has the required human capabilities. With AI being a very progressive field, it can be difficult to stay informed. For governments to share data to enable AI, public sector bureaucrats must acquire the required skills to fully scrutinise the potential consequences of AI against public trust and values (i.e. transparency, fairness, equity, etc.).[159] Similarly, they need to be trained in specific areas, such as machine learning, data analytics, automated decision-making, etc., to understand the possibilities AI can help unlock.

An oftenunderappreciated skill is being able to consider the level of digital and development capacity needed bygovernments to effective in-sourcr technology or software development. Whilst there are often agencies mandated in government to assist in these functions, for data sharing to be effective: it needs to be functionally spread across departments and agencies.

6.5. Public and Cultural attitudes

The development and enforcement of data governance for AI development have predominantly followed an ethical and principled approach, providing prescriptions for data sharing across the data lifecycle. This is the case for governments such as the UK and Germany, as well as private sector

[157] See for example, 'Value Sharing Framework for NHS data partnerships' https://transform.england.nhs.uk/key-tools-and-info/centre-improving-data-collaboration/value-sharing-framework-for-nhs-data-partnerships/

[158] Mariana Mazzucato, (2022). Missions and public purpose: a new social contract between business, labour and the state. UCL Institute for Innovation and Public Purpose (IIPP) https://www.ucl.ac.uk/bartlett/public-purpose/sites/bartlett_public_purpose/files/mazzucato_m._2022._rethinking_the_social_contract_between_the_state_and_business_a_new_approach_to_industrial_strategy_with_conditionalities.pdf

[159] Maciej Kuziemski, and Gianluca Misuraca, 2020. AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications policy*, *44*(6), p.101976 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7164913/

actors such as Accenture and IBM.[160] On the other hand, ethics, principles, or norms may vary from one cultural context to another. Cultural values such as public trust can be interpreted differently, and comprehension of AI and its implications can vary depending on individuals, countries, or contexts. Cultural values may also include openness, transparency, ethics, democracy, privacy, and can be enablers for governments' efforts to share data they hold.[161] As Robinson has noted, "it is critical that policymakers, legislators, industry, and citizens have the opportunity to understand how cultural values interact with policy discussions about technologies such as AI."[162]

An aspect of cultural attitudes for data sharing worth considering includes those specific to political cultures. This supply-side imperative is important, as cultures of secrecy are frequently cited as key inhibitors to data-sharing practice both within, and outside of, government. This attitude also extends to how much governments are willing to invest in data collection to ensure structured data is available to AI developers, and how much data collection is recognised as a strategic pillar requiring dedicated and sufficient resources that can create future value.

[160] GPAI 2022, 'Data Governance Working Group: A Framework Paper for GPAI's Work on Data Governance 2.0, GPAI Tokyo Summit (November 2022) https://gpai.ai/projects/data-governance/Data%20Governance%20-%20A%20Framework%20Paper%20for%20GPAI%E2%80%99s%20Work%20on%20Data%20Governance%202.0%20.pdf

[161] Stephen Robinson, 2020. Trust, transparency, and openness: How inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI). *Technology in Society*, *63*, p.101421 https://www.sciencedirect.com/science/article/pii/S0160791X20303766

[162] Note Stephen Robinson above.

7. Recommended Principles for Data Sharing by Governments

In deepening the conceptual idea of governments as a provider of data for AI development, there are important lessons arising from our case study analysis. These include the beneficial value of the AI systems that are utilising government data, the relevance of data collaborative approaches that adopt equitable data-sharing models, and the importance of protecting human rights and promoting data justice. The case studies also show the need to minimise data sharing with third parties and the existence of a legal basis for sharing data with third parties. In addition, any data sharing should be governed by agreements with a defined scope for processing government-held data and independent oversight institutions. Across all case studies, the importance of transparency and data subject participation in building public trust in data sharing was also evident. Finally, underlying data sharing by governments is the need to promote digital equity to maximise the beneficial value of the AI systems.

The analysed case studies from different regions with different stages of development in AI strategies, legal frameworks, technological capabilities, and technical capacities show the difficulty in developing universal principles that would apply to very diverse country contexts. However, as countries adopt their respective Responsible AI principles, and as global institutions with member states such as the OECD[163] and UNESCO[164] also publish Responsible AI principles, there is a convergence by governments around key themes on how data should be used in AI systems. The principles set out below have identified areas of further convergence with Responsible AI principles and integrated them where necessary.

The following apply to ***when*** governments should share data:

1. **Public Benefit**

   AI systems aiming to use government data should have beneficial outcomes for the public and data subjects.[165] Use of government-held data to develop AI can create financial value, and governments should ensure data-sharing partnerships are created fairly to prevent further wealth inequalities. AI systems aiming to use government data should also have beneficial outcomes for the public and data subjects in order to safeguard sustainable development and individual well-being. They should contribute as far as possible to achieving the Sustainable Development Goals.

2. **Accountability**

   Government data shared for AI development should be fully auditable, with scope of data use defined and a clear legal basis for processing. Governments should ensure full compliance with legal frameworks, and data-sharing agreements should clearly identify the obligations of third parties.

[163] OECD AI Principles Overview
https://oecd.ai/en/ai-principles
[164] UNESCO Ethics of Artificial Intelligence
https://www.unesco.org/en/artificial-intelligence/recommendation-ethics
[165] Stahl argues for the use of AI for the furtherance of human well-being. The EU and OECD have also adopted human, societal and environmental well-being as part of their Responsible AI Principles. See Bernd Stahl, (2020) *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AU and Emerging Digital Technologies*' Springer.

3. **Data subject participation**

   Data subjects should have the option to consent to the sharing of their data by governments with external parties except under some very specific and regulated regimes like research, governmental statistics and archives. Governments should only enter data-sharing partnerships where the rights and interests of data subjects can be fully protected.

4. **Data Equity and Data Justice**

   Governments should adopt accessible data-sharing models and necessary infrastructure that enables equitable access to relevant government data through fair public procurement, investment, and collaborative approaches in data sharing.

The following principles apply to *how* governments should share data:

5. **Transparency**

   Governments should be transparent with the public about the purpose and means of data sharing through notice and consent where necessary. Consultative engagements with the public are also encouraged in planning for sharing data with third parties.

6. **Safety and Security**

   AI systems using government data should be technically robust and safe. They should aim to prevent potential breaches of data, including through use of privacy-enhancing technologies where relevant and through use of facilities that are equally safe and secure. They should address issues around accuracy, reliability and reproducibility.

7. **Data Minimisation and Proportionality**

   Sharing of government data for AI systems should not go beyond what is necessary to achieve the legitimate aim of the system.

8. **Human Rights Protection and Privacy**

   AI systems that use government data must promote digital inclusion, protect human rights, and have built-in processes to ensure fairness and prevent harmful bias through user-centric design that ensures accessibility and keeps a human-in-the-loop. Independent oversight of data sharing should be present and the full spectrum of the human rights should be protected.

8. Conclusion

Governments hold important data about the populations they serve, which can be used to develop AI applications that support addressing social challenges, public interest issues and developmental priorities.[166] This is a comprehensive report of the research conducted by Research ICT Africa and commissioned by CEIMIA and GPAI, to better understand the role of government as a responsible provider of data for AI development. It defines what is meant by "government data sharing for AI". The report summarises the key findings from a review of four illustrative case studies, synthesising the findings to discern the key enablers that should support the sharing of government data with AI developers. It also identifies possible mitigation strategies to overcome barriers to data sharing, and also recommends key principles that should be considered by governments as part of data sharing for AI.

The case studies in this report show that use of government-held data in the development of AI can achieve important developmental goals if data is shared in a way that earns public trust and protects the privacy rights of data subjects. This report found that governments and their private sector partners should be transparent about data use and should actively seek data subject participation. It also highlights the importance of embracing privacy-enhancing technologies and adopting data-sharing models that can foster accountability in data use. In addition, the adoption of any type of data-sharing model by governments needs to consider the implications of third-party access for data justice, inclusion, and equity. This means progressive regulation which establishes strong oversight institutions, and recognises data subject rights such as the right to portability, and appropriate data-sharing methods. Transparent public procurement practices for AI and new methods of data sharing with governments as data stewards and facilitators of access to data should also be considered.

To develop sound governance frameworks that safeguard against risks, and to create enabling environments for the opportunities of government data for AI to be catalysed, further insights are needed into what the future of government as a provider of data might entail. This includes assessing the factors that are driving change, as well as the degree to which these factors are desirable, probable, and feasible in the context of the responsible provision of government data for AI development.

[166] Stefano Sedola, Andrea Junior Pescino, and Tira Greene (2021). Blueprint - Artificial Intelligence for Africa. Smart Africa.
https://www.bmz-digital.global/wp-content/uploads/2022/08/70029-eng_ai-for-africa-blueprint.pdf

References

1. Agreement between: Taunton and Somerset NHS Foundation Trust and DeepMind Technologies Limites and Google Health UK Limited https://www.whatdotheyknow.com/request/607622/response/1449343/attach/5/TSFT%20Redacted%20GHUK.PDF.pdf

2. Amy Dickens, 'The Right to Health: Implications of Data Driven Health Research Partnerships' (July 2021) https://repository.essex.ac.uk/31194/1/PhD-%20FINAL%20VERSION%20(w.%20corrections).pdf

3. Andrew Prismal vs Google UK Limited, DeepMind Technologies Ltd and LCM Funding UK Limited [2023] EWHC 1169 (KB) https://www.judiciary.uk/wp-content/uploads/2023/05/Prismall-v-Google.pdf

4. Regulations Governing Use and Management of Health Passbook Software Development Kit, National Health Insurance Administration, Ministry of Health and Welfare (2022).

5. B. Prainsack, S. El-Sayed, N. Forgó, Ł. Szoszkiewicz, P. Baumer (2022) 'Data solidarity: a blueprint for governing health futures' Volume 4, Issue 11

6. B. Snaith & J. Massey. (2021), 'Data Institutions for Climate Action', https://theodi.org/article/data-institutions-for-climateaction/

7. Chen Tzi-Fa, 'The Experience of Establishing National Health Insurance App Mobile Phone Number Identity Authentication System', Government Agency Information Communication No.357, p31-36 (2019) (in Chinese).

8. C. van Ooijen, B. Ubaldi and B. Welby (2019), 'A data-driven public sector: Enabling the strategic use of data for productive, inclusive and trustworthy governance', *OECD Working Papers on Public Governance*, No. 33, OECD Publishing, Paris, https://doi.org/10.1787/09ab162c-en.

9. Gregory J. Mathews, Ofer Harel, 'Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy' (2011) https://projecteuclid.org/journals/statistics-surveys/volume-5/issue-none/Data-confidentiality--A-review-of-methods-for-statistical-disclosure/10.1214/11-SS074.full

10. Data Governance in The Public Sector (OECD, 2020): https://www.oecd-ilibrary.org/sites/9cada708-en/index.html?itemId=/content/component/9cada708-en

11. Emerging models of data governance in the age of datafication: https://journals.sagepub.com/doi/full/10.1177/2053951720948087

12. Expanding MLDS Data Access and Research Capacity with Synthetic Data Sets: https://mldscenter.maryland.gov/egov/Publications/ResearchReports/SDPReportFINAL.pdf#:~:text=The%20MLDS%20Center%20synthetic%20data%20project%20%28SDP%29%20was,creating%20cluster-specific%20synthetic%20versions%20of%20the%20MLDS%20data

13. G. Razzano, A. Beyleveld, F. Adeleke (2021) 'Automated Decisions and the Public Sector in Africa' Luminate Report

14. G. Razanno, 'Data Localisation in South Africa', (2021) Mandela Institute Policy Brief 06. https://www.wits.ac.za/media/wits-university/faculties-and-schools/commerce-law-and-manag

ement/research-entities/mandela-institute/documents/research-publications/800482%20PB6%20Missteps%20in%20valuing%20data_REV%20Dec2021.pdf

15. G. Razanno, 'AI4D - Digital and Biometric Identity Systems', (2021), Research ICT Africa. https://researchictafrica.net/publication/policy-paper-ai4d-digital-and-biometric-identity-systems/

16. GPAI (2022), 'Data Justice: A Primer on Data and Social Justice, Report' (2022), Global Partnership on AI. https://gpai.ai/projects/data-governance/primer-on-data-and-social-justice.pdf

17. GPAI (2022), 'Data Governance Working Group – A Framework Paper for GPAI's Work on Data Governance 2.0', (2022), GPAI Tokyo Summit. https://gpai.ai/projects/data-governance/Data%20Governance%20-%20A%20Framework%20Paper%20for%20GPAI%E2%80%99s%20Work%20on%20Data%20Governance%202.0%20.pdf

18. Guan-Lin Ho, 'An Empirical investigation of the Factors that Influence Individuals to Download Medical Data from My Health Bank', I-Shou University, unpublished Master's thesis (2017) (in Chinese).

19. Information Commissioner's Office, 'RFA0627721 - Provision of Patient Data to DeepMind' https://ico.org.uk/media/action-weve-taken/undertakings/2014353/undertaking-cover-letter-revised-04072017-to-first-person.pdf

20. Is Synthetic Data the Future of AI: https://www.gartner.com/en/newsroom/press-releases/2022-06-22-is-synthetic-data-the-future-of-ai

21. Julia Powles, Hal Hodson, *Google DeepMind and Healthcare in an age of algorithms* (March 2017)

22. Memorandum of Understanding for Data Sharing Agreement https://nassp.gov.ng/wp-content/uploads/2021/06/MOU-for-Data-Sharing-Agreement.pdf

23. M. Kuziemski, G. Misuraca, (2020) 'AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings', Telecommunications Policy, Volume 44, Issue 6.

24. Open Data to Improve Agricultural Resiliency: https://odimpact.org/case-aclimate-colombia.html

25. OECD (2019), 'Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies', OECD Publishing, Paris. https://www.oecd.org/sti/enhancing-access-to-and-sharing-of-data-276aaca8-en.htm

26. Pi-Jung Hsieh & Hui-Min Lai, 'Exploring People's Intentions to Use the Health Passbook in Self-Management: An Extension of the Technology Acceptance and Health Behavior Theoretical Perspectives in Health Literacy', 161 Technological Forecasting and Social Change 120328 (2020).

27. Privacy Impact Assessment Google DeepMind Streams at Royal Free NHS Foundation Trust http://s3-eu-west-1.amazonaws.com/files.royalfree.nhs.uk/Privacy_Impact_Assessment___Streams___Royal_Free_Hospital.pdf

28. Procedure for Access, Mining, and Verification of the NSR by other Social Safety Nets Agencies https://nassp.gov.ng/wp-content/uploads/2021/06/Data-Mining-Protocoal-PROCEDURE-FOR-ACCESS-2020.pdf

29. Regulations Governing Use and Management of Health Passbook Software Development Kit, National Health Insurance Administration, Ministry of Health, and Welfare (2022).

30. Robert Sparrow, Mark Howard & Chris Degeling (2021), 'Managing the risks of artificial intelligence in agriculture', NJAS: Impact in Agricultural and Life Sciences, 93:1, 172-196, https://www.tandfonline.com/doi/pdf/10.1080/27685241.2021.2008777

31. S. El-Sayed & B. Prainsack (2022) 'PLUTO/PublicVal - Public Valut Tool' https://digitize-transformation.at/news-und-events/detailansicht/news/plutopubval-public-value-tool/

32. Steven Sotelo, Edward Guevara and others (2020), 'Pronosticos AClimateColombia: A system for the provision of information for climate risk reduction in Colombia' (July 2020), Science Direct. https://www.sciencedirect.com/science/article/pii/S0168169919315832

33. Stefano Sedola, Andrea Junior Pescino, and Tira Greene (2021), 'Blueprint - Artificial Intelligence for Africa', Smart Africa. https://www.bmz-digital.global/wp-content/uploads/2022/08/70029-eng_ai-for-africa-blueprint.pdf

34. Taiwan's National Health Insurance Research Database: Past and Future: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6509937/

35. The UK Data Protection Act, 2018

36. The UK Data Protection Act, 1998

37. The UK Information Governance Review (March 2013)

38. The UK Health Service (Control of Patient Information) Regulations 2002

39. The Global Data Barometer (2022): https://globaldatabarometer.org/the-global-data-barometer-report-first-edition/

40. The Royal Free London NHS Foundation Trust, 'Audit of the Acute Kidney Injury detection systems known as Streams' https://s3-eu-west-1.amazonaws.com/files.royalfree.nhs.uk/Reporting/Streams_Report.pdf

41. Ting-Chi Liu, 'The Past, Present, and Future of the Right to Information Privacy: A Comparative Law Perspective', (2010) J. Institutional & Theoretical Eco. 120, 163-64 (2010).