From: Lewis Rangott
Sent: Tuesday, 9 April 2024 4:57 PM
To: Portfolio Committee 1
Cc: Jenny Ryan
Subject: CM: RE: Artificial intelligence (AI) in New South Wales – Post-hearing responses – 11 March 2024

Talina,

Answers to the two QoN are as follows. Please let us know if more detail is required.

The Hon. Dr SARAH KAINE: I'm also interested in procurement. We have a parallel inquiry in just about everything but one is about procurement at the moment. You talk about procurement and about an AI system used in Brazil to identify red flags in the public procurement process. I wondered if you could give any more information on that. You might need to take this on notice, but I can't see a reference for us to look up and get more detail about it. If you have any detail now, that would be great. If not, if you could provide it, that would be helpful.

LEWIS RANGOTT: I can provide the fine detail on notice but, as I understand that particular matter, it was relatively basic use of technology. It was comparing data in one pool over here and data in another pool over there and matching them and finding some red flags. I think in that particular case it was something that we see very commonly in our work. It was a public servant who was awarding contracts to their own company without disclosing the conflict. So that's rudimentary use of technology.

Response – The case study involving Brazil can be found in a report issued by The World Bank "Artificial Intelligence in the Public Sector - Maximising Opportunities Managing Risks" (2020), available at https://documents1.worldbank.org/curated/en/809611616042736565/pdf/Artificial-Intelligence-in-the-Public-Sector-Maximizing-Opportunities-Managing-Risks.pdf (see pp. 20-21). In addition, in March 2024, the OECD published its report "Generative AI for anti-corruption and integrity in government" which may be of interest. It is available at https://www.oecd.org/publications/generative-ai-for-anti-corruption-and-integrity-in-government-657a185a-en.htm.

[Note – Copies of these publications are attached.]

The CHAIR: In your submission you say that there is an increasing trend for data breaches. I think 20 per cent of the data breaches in 2023 that you are aware of involved social engineering schemes or impersonation. I'm aware of a major quite high-profile one recently—a Zoom meeting where a person was tricked into sending a large amount of money in a Zoom meeting online or something like that with fake personalities. You are saying that public agencies are already at risk and suffering from this. What do you mean by social engineering schemes and how does the Government protect its agencies against those?

PAUL LAKATOS: The social engineering schemes, as I understand it, are schemes which take the personal characteristics of a user using a computer and then wraps up a request or a demand in personal information, hence making it look like it's a genuine request by a genuine body. How you stop it is a question on

notice. I think we are all grappling with that. I don't think we have encountered a practical application in ICAC. Again, it's not likely to affect our work itself directly. It may affect what the people we are investigating do and how that's done.

Response – The NSW Department of Customer Service "*Cyber Security Guide*" defines social engineering as "attacks that aim to manipulate people to provide confidential or personal information, which can be used for fraudulent purposes. There are many forms, the most common being phishing" (see: https://www.digital.nsw.gov.au/sites/default/files/2023-09/cyber-security-guide-general.pdf, p. 2). A UK Parliament publication on fraud states "Social engineering is the process by which criminals groom and manipulate people into divulging personal and financial details or transferring money. Fraudsters use social engineering to bring a victim into what Brian Dilley called a "hot state". This is the point at which individuals stop thinking clearly and often feel rushed, anxious and mistrustful" (see: https://publications.parliament.uk/pa/ld5803/ldselect/ldfraudact/87/8706.htm).

Social engineering schemes require a degree of planning and effort because they are targeted at individuals or small cohorts. Advances in artificial intelligence make it easier to gather and deploy the personal information required to successfully manipulate a target.

One form of social engineering involves impersonation of a figure of authority (such as the head of the organisation or a senior manager), who requests or demands that a subordinate to provide sensitive information, change bank account details or make a payment. This is sometimes called "CEO fraud". A similar fraud can occur when someone impersonates a supplier or other payee of a public sector agency.

Social engineering can also target citizens by impersonating a public sector agency in an attempt to obtain personal information or money. In addition to harming the target, this can make it difficult for public sector agencies to carry out legitimate transactions with citizens because each party may question the authenticity of the other.

Kind regards,

Lewis

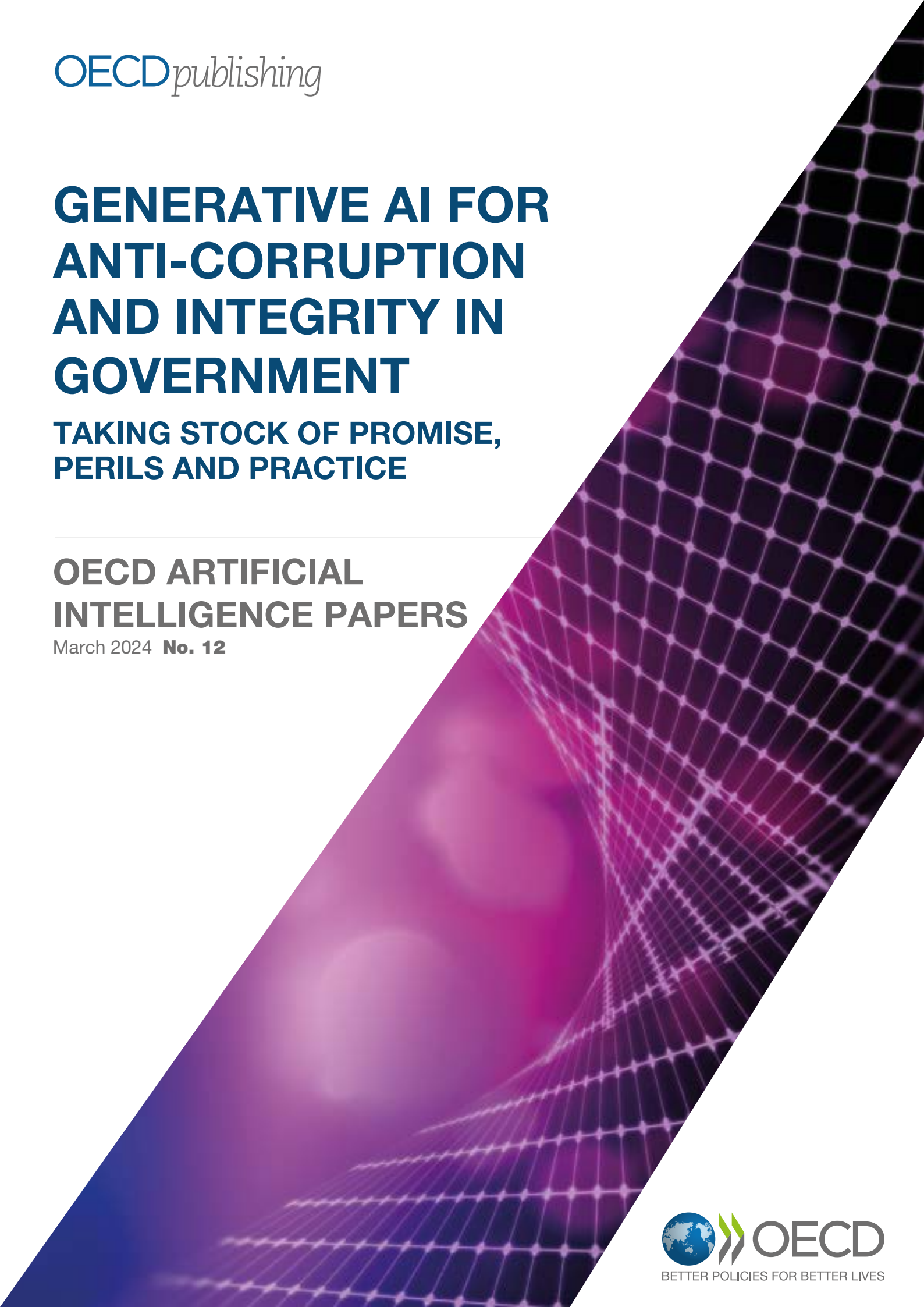
Lewis Rangott | Executive Director Corruption Prevention
NSW Independent Commission Against Corruption

OECD publishing

GENERATIVE AI FOR ANTI-CORRUPTION AND INTEGRITY IN GOVERNMENT

TAKING STOCK OF PROMISE, PERILS AND PRACTICE

OECD ARTIFICIAL INTELLIGENCE PAPERS
March 2024 **No. 12**



OECD
BETTER POLICIES FOR BETTER LIVES

OECD Artificial Intelligence Papers

Working Papers describe preliminary results or research in progress by the author(s) and are published to stimulate discussion on a broad range of issues on which the OECD works. Comments on Working Papers are welcomed, and may be sent to OECD Directorate for Public Governance, OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France; e-mail: gov.contact@oecd.org.

Authorised for publication by Elsa Pilichowski, Director, Public Governance Directorate.

Foreword

This paper examines the opportunities and challenges related to the use of generative artificial intelligence and large language models reported by a group of government actors engaged in anti-corruption and integrity efforts, termed “integrity actors.” The paper presents insights from responses to a questionnaire administered to integrity actors in government in early 2024.

Under the supervision of Elsa Pilichowski, Director of the Public Governance Directorate (GOV) and the guidance of Julio Bacio Terracino, Head of GOV’s Anti-Corruption and Integrity in Government Division, Gavin Ugale drafted this paper with contributions from Cameron Hall. Jamie Berryhill, Gallia Daor, Claire McEvoy, Mauricio Mejia Galvan, María Pascual Dapena, Karine Perset, Helene Wells and Ricardo Zapata provided valuable comments. Meral Gedik prepared the paper for publication.

The OECD Secretariat would also like to thank the following institutions for providing their insights via the OECD questionnaire: Court of Audit (Austria), Federal Internal Audit (Belgium), Global Affairs (Canada), Office of the Auditor General (Canada), Office of the Comptroller General (Colombia), Office of the Comptroller General (Costa Rica), Ministry of Finance (Costa Rica), Ministry of Justice (Czechia), Agency for Public Finance and Management (Denmark), Office of the Public Prosecutor (Denmark), National Audit Office (Estonia), Ministry of Finance (Finland), Ministry of Justice (Finland), National Bureau of Investigation (Finland), Court of Accounts (France), High Authority for Transparency in Public Life (France), Ministry of Digital Governance (Greece), National Transparency Authority (Greece), Integrity Authority (Hungary), National Tax and Customs Administration (Hungary), Anti-Corruption Authority (Italy), Board of Audit (Japan), Board of Audit and Inspection (Republic of Korea), Corruption Prevention and Combatting Bureau (Latvia), State Audit Office (Latvia), Ministry of Public Administration (Mexico), Ministry of Interior and Kingdom Relations (Netherlands), Court of Audit (Netherlands), Agency for Public and Financial Management (Norway), Government Security and Service Organisation (Norway), Court of Auditors (Portugal), Directorate-General for Administration and Public Employment (Portugal), Court of Audit (Slovenia), the General Comptroller of the State Administration (Spain), National Audit Office (Sweden), Internal Audit of the Embassy of Sweden to Guatemala (Sweden), Federal Statistical Office (Switzerland), National Audit Office (United Kingdom), Department of State (United States), Government Accountability Office (United States), European Court of Auditors (European Union), European Confederation of Institutes of Internal Auditing (European Union), Agency for the Prevention of Corruption and Coordination of the Fight against Corruption (Bosnia and Herzegovina), Office of the Comptroller General (Brazil), Federal Court of Accounts (Brazil), General Inspectorate of the State (Djibouti), Office of the Comptroller General (Ecuador), Integrity and Anti-Corruption Commission (Jordan), Ministry of Economy (Kosovo),[1] National Audit Office (Malta), Agency for Prevention of Corruption (Montenegro), National Anti-Corruption Centre (Republic of Moldova), General-Directorate of Anti-Corruption (Romania), Agency for Prevention of Corruption (Serbia), General Control of Finance (Tunisia), Presidency of the Government (Tunisia), Court of Accounts (Tunisia), State Audit Service (Ukraine).

The OECD Secretariat would also like to express its gratitude to Taka Ariga (Government Accountability Office, United States), Gutemberg Assuncao Vieira (Office of the Comptroller General, Brazil), Máté Benyovszky (Integrity Authority, Hungary) and Emanuele Fossati (European Court of Auditors, European Union) for their insights that helped to shape the questionnaire.

Table of contents

FIGURES

BOXES

Abbreviations and acronyms

ACA	Anti-Corruption Agency
AI	Artificial Intelligence
GPT	Generative Pre-trained Transformers
IT	Information Technology
LLM	Large Language Model
NLP	Natural Language Processing
OECD	Organisation for Economic Co-operation and Development
RAG	Retrieval-Augmented Generation
SAI	Supreme Audit Institution

Executive summary

Generative artificial intelligence (AI) has been part of the technological landscape for some time, but recent developments, particularly in large language models (LLMs) as one type of generative AI, have recently propelled it into a position of disruptive influence. Governments must keep pace with this innovation not only as regulators, but also as users. This paper explores the latter challenge with a focus on integrity actors, including anti-corruption agencies (ACAs) and oversight bodies, such as supreme audit institutions (SAIs) and internal audit functions.

The integrity actors who offered insights for this paper identified several opportunities and benefits of generative AI, focusing largely on their exploration and use of LLMs. For instance, integrity actors in Brazil are deploying LLMs to sift through massive datasets to identify patterns indicative of fraud, offering insights for investigations and risk mitigation measures. Integrity actors in Finland, France, Greece and the United Kingdom are using LLMs to support in drafting documents, analysing spreadsheets and summarising texts. These LLMs can make the day-to-day work of auditors and investigators more efficient, thereby freeing them from time-consuming organisational tasks.

Integrity actors also highlighted various challenges, ranging from technical ones concerning the integration of LLMs to strategic questions about ensuring trustworthy AI systems. Integrity actors recognise that LLMs are an evolving technology capable of "hallucinations," whereby they may generate convincing yet inaccurate, fabricated or misleading information, based on unclear reasoning. This inherent complexity in how LLMs generate outputs can perpetuate a lack of transparency and accountability in decision making, which can undermine the very principles that integrity actors seek to uphold. Failure to mitigate these risks, curb bias and promote responsible and ethical use of AI, can have harmful real-world impacts, such as the reinforcing of structural inequalities and discrimination.

To identify and explore these opportunities and challenges, the OECD sent a questionnaire to and interviewed organisations from several OECD communities, including the Working Party of Senior Public Integrity Officials, the Auditors Alliance, and a Community of Practice on Technology and Analytics for Public Integrity. Based on the responses of 59 organisations from 39 countries, the OECD collected key insights concerning the use of generative AI and LLMs. They included the following:

- Generative AI, particularly LLMs used for processing and generating text, can enhance the internal operations of integrity actors, with the most promising gains in operational efficiency and analysing unstructured data. For investigative and audit processes, integrity actors saw the highest value of LLMs in evidence gathering and document review, with a significant portion of respondents, especially those conducting performing audits, prioritising these activities.
- LLMs show promise for strengthening several anti-corruption and anti-fraud activities, but examples in government are limited and the return on investment is unclear. Integrity actors viewed document analysis and text-based pattern recognition as the most valuable use cases of LLMs for anti-corruption and anti-fraud. However, respondents reported few advanced initiatives in this area, and many organisations are still incubating ideas.

- Integrity actors cited a shortage of skills and IT limitations as the biggest challenges they face to implement LLMs. Many institutions expressed that they either lack sufficient financial, human, and technical resources to deploy LLMs entirely, or their staff does not have sufficient data literacy to use such tools. Concerns about budget constraints were comparatively more pronounced among internal audit bodies and ACAs relative to SAIs.
- Advice for piloting LLMs includes first incorporating them into low-risk processes and considering the requirements for scaling early on. Such an approach can build capacity where mistakes are not as costly before scaling generative AI to riskier, more resource-intensive and more analytical tasks. Having an early handle on the organisational needs for computational and storage resources can help an organisation to prepare for scaling.
- Integrity actors mostly rely on turnkey foundation LLMs developed by technology companies. Various options exist to develop LLMs, from open-source models to those created by private firms or government entities. In practice, integrity actors that responded to the questionnaire are either using an existing, turnkey model outright or they are fine-tuning a foundation model (i.e. further training a pre-trained LLM with specific datasets to adapt its capabilities for particular tasks).
- Overcoming language barriers inherent in using or fine-tuning off-the-shelf LLMs is a key challenge. Currently, most LLMs are trained in English, which poses limitations for many integrity actors who wish to deploy models in their native language. To address this challenge, some countries are investing in the development of local language LLMs.
- Integrity actors recognised the need for safeguards in some areas but can do more to ensure trustworthy AI systems, as well as the responsible and ethical use of generative AI as initiatives mature. Integrity actors can improve their focus and activities to mitigate the risks of bias and discrimination and address ethical concerns in how they use and apply LLMs internally.
- Integrity actors can put a greater emphasis on monitoring and evaluating LLMs, including considerations pertaining to the interpretability and explainability of a model's outputs. Evaluating LLMs and attempting to explain results poses complex challenges. However, addressing these challenges with multi-faceted solutions will be critical for the uptake of LLMs amongst integrity actors.
- Generative AI can enhance the work of integrity actors, but it also necessitates greater vigilance of evolving integrity risks. For instance, LLMs provide new ways for integrity actors to operate and assess risks, but they also can accelerate and amplify certain types of fraud and corruption.

The findings from the OECD's questionnaire are not generalisable to all integrity actors. Nonetheless, the paper describes common challenges and potential use cases that are transferable across contexts, providing inspiration to integrity actors as they consider how to make the most of this rapidly evolving technology. The OECD's policy-focused work offers inspiration throughout the paper, including the work of the OECD.AI Policy Observatory, as well as the OECD's Recommendation of the Council on Artificial Intelligence and the Recommendation on Digital Government Strategies.

1 Generative AI: Opportunities for enhancing anti-corruption and integrity in government

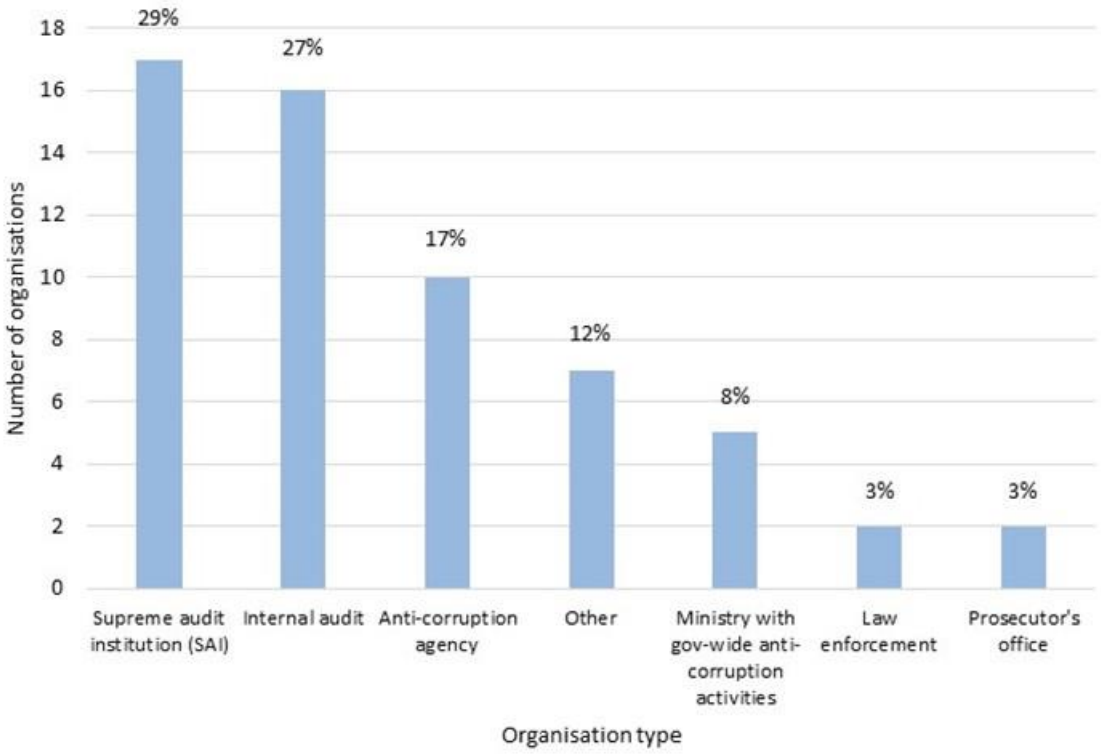
This section explores the opportunities for integrity actors to use generative AI, particularly LLMs, to enhance their internal operations as well as their anti-corruption activities. It presents the views of 59 integrity actors captured in an OECD questionnaire on generative AI for integrity and anti-corruption, including insights into the potential benefits the technology offers. The supreme audit institutions that responded to the questionnaire are generally the most advanced in their use of generative AI among questionnaire respondents. However, most integrity actors that responded to the questionnaire are still in the early stages of thinking about or developing generative AI tools.

1.1. The OECD's questionnaire on generative AI for integrity and anti-corruption

In January 2024, the OECD administered a questionnaire for integrity actors in government on the use of generative AI for public integrity and anti-corruption. To implement the questionnaire, the OECD relied primarily on three of its communities: the Working Party of Senior Public Integrity Officials, the Community of Practice on Technology and Analytics for Integrity and the Auditors Alliance. With the help of members of these communities, the OECD identified integrity actors in government with the relevant mandate and expertise to react to a questionnaire about generative AI for integrity and anti-corruption. Several participants from the Community of Practice piloted the questionnaire and select respondents provided additional insights via targeted interviews.

For purposes of the questionnaire and this paper, integrity actors include anti-corruption agencies (ACAs), supreme audit institutions (SAIs), internal audit or control functions, and ministries with government-wide integrity and anti-corruption activities (e.g. Ministry of Public Administration). They also include law enforcement and prosecutors' offices. These integrity actors together account for 88% (52) of the 59 organisations that responded to the questionnaire.[2] All but one of the other seven organisations to respond represent government entities that are responsible for AI policy. One Tax and Customs Administration responded to the questionnaire as well. SAIs and internal audit functions provided just over half of all responses. Figure 1.1 summarises some of these key features of the organisations that responded.

Figure 1.1. Number of respondents to the OECD's questionnaire by type of organisation



Note: The percentages show the proportion of organisations out of a total 59 that responded to the questionnaire. Internal audit bodies include both central internal audit bodies and internal audit units of individual institutions, as well as comptroller general's offices and ministries and agencies responsible for public financial management. The "other" category contains primarily ministries responsible for government-wide AI policy as well as one tax and customs agency.
Source: OECD questionnaire

Respondents to the questionnaire represent a broad range of government entities with different institutional mandates with regards to public integrity and anti-corruption. Most of the respondents have roles and responsibilities related to IT, data science, AI or digital initiatives within their organisation. The OECD did not attempt to identify or contact the entire sub-populations of integrity actors, as we define them in this paper. The ultimate purpose of this paper and the questionnaire is to explore current use cases and provide a snapshot of practices, opportunities and challenges. As such, the results are not generalisable to broader populations. All descriptive statistics that illustrate key findings reflect responses to the OECD's questionnaire without exception.

1.2. Overview of the maturity of generative AI initiatives

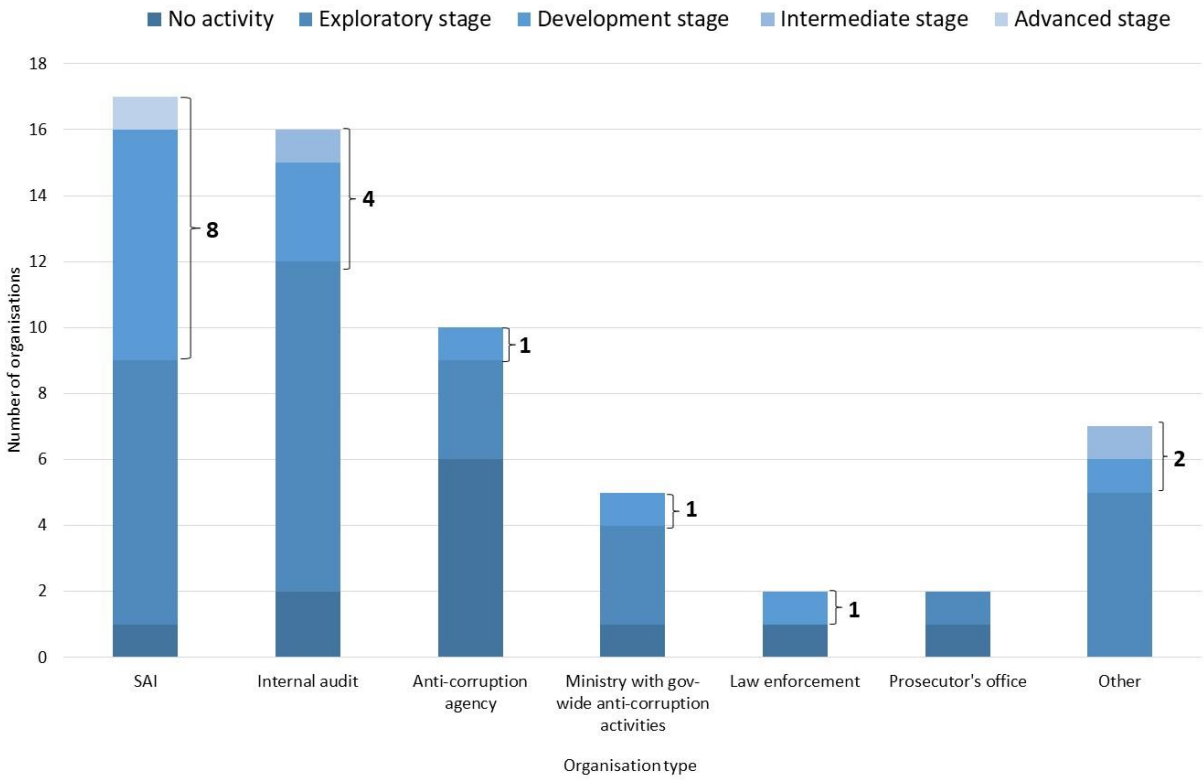
In recent years, generative AI surged in prominence with the rise of deepfakes and the introduction of transformative models like Generative Pre-trained Transformers (GPTs) and other large language models (LLMs), marking a significant leap forward in the field. LLMs are advanced machine learning algorithms proficient in interpreting inquiries or commands and producing responses in human-like language. These models function by processing extensive datasets during their training phase, allowing them to identify statistical correlations, such as how words relate to each other and the contextual importance of words within sentences. Utilising this insight, the models are capable of sequentially generating text, predicting each subsequent word in a sequence (OECD, 2023[1]) (Shabsigh and Boukherouaa, 2023[2]). The technology captured global interest in November 2022 with the introduction of text-to-image generators and the release of Open AI's ChatGPT (Lorenz, Perset and Berryhill, 2023[3]).

In this context, integrity actors have had little time to comprehend the opportunity generative AI presents for their work, let alone to fully integrate it into activities. When the OECD surveyed integrity actors, the expectation was that across the board the respondents would describe their organisations as being in the early stages of maturity concerning the use of generative AI and LLMs. Not only is the technology relatively new, but government entities—integrity actors included—are not known for being first movers in terms of technology adoption. The responses to the OECD's questionnaire reflect these expectations. Of the 59 organisations that responded from 39 countries, as well as two supranational organisations in the European Union, approximately 50% (30) reported they do not use generative AI in their operations, but they are exploring potential use cases. Another 24% (14) of respondents indicated their institutions are in the development phase. In other words, they have experimented with generative AI in a few projects, but it is not yet integrated into the organisations' operations.

SAIs' efforts to use generative AI were the most mature relative to other types of organisations, including one respondent who described their SAI's use of the technology as "advanced." Overall, 47% (8) of SAI respondents reported being at least in the development stage of using generative AI, the highest percentage of the different organisational types. After SAIs, 25% (4) of respondents working in internal audit bodies said their organisation is in the development stage or beyond, while only one institution in each of the other categories has reached at least the development stage. Figure 1.2 summarises these results and provides definitions for the different stages. The counts highlighted with brackets indicate the number and type of organisations that have reached at least the development stage, which is a subgroup of surveyed organisations that is the focus of subsequent analysis.

Figure 1.2. Stage of generative AI and LLM use by type of organisation

Which of the following options best describes the maturity of your institution's use of Gen AI and LLMs specifically, as a sub-domain of AI?



Note: The data label callouts highlight the number of institutions that have reached at least the development stage. Internal audit bodies include both central internal audit bodies and internal audit units of individual institutions, as well as comptroller general's offices and ministries and agencies responsible for public financial management. The "other" category contains primarily ministries responsible for government-wide AI policy as well as one tax and customs agency. Possible responses included the following: 1) Advanced Stage: Gen AI is deeply integrated into our core operations and we continuously seek ways to improve and expand its use. 2) Intermediate Stage: Gen AI is used in several areas of our activities, but it is not yet fully optimised or widespread. 3) Development Stage: We have experimented with Gen AI in a few projects but it is not yet integrated into our operations. 4) Exploratory Stage: We do not use Gen AI in our operations but we are currently exploring potential uses. 5) No Activity: We do not use Gen AI in our operations and we are currently not exploring potential uses.
Source: OECD questionnaire.

While these results are not generalisable, they align with the OECD's experiences working with these communities. Among SAIs that have successfully incorporated innovative approaches to the use of technology, data, and analytics into their audit work, a common thread is their openness to experimentation. In some countries, SAIs may also have access to more resources than other types of integrity actors, therefore enabling more experimentation, as highlighted later in the paper. This commitment to experimentation remains consistent even when other aspects of the SAI's work and culture tend to be risk averse. For those SAIs that have established dedicated "Innovation Labs," experimentation has become a strategic objective.

One notable advantage of an innovation lab is its role in institutionalising knowledge and expertise. This model can help to advance new methodologies that can benefit multiple departments within the SAI. For example, SAIs in countries like Brazil, the United States, and Norway have all established effective innovation labs to assist auditors in keeping pace with technological developments and drive continuous professional development (OECD, 2022[4]). This includes the integration of technology and data-driven

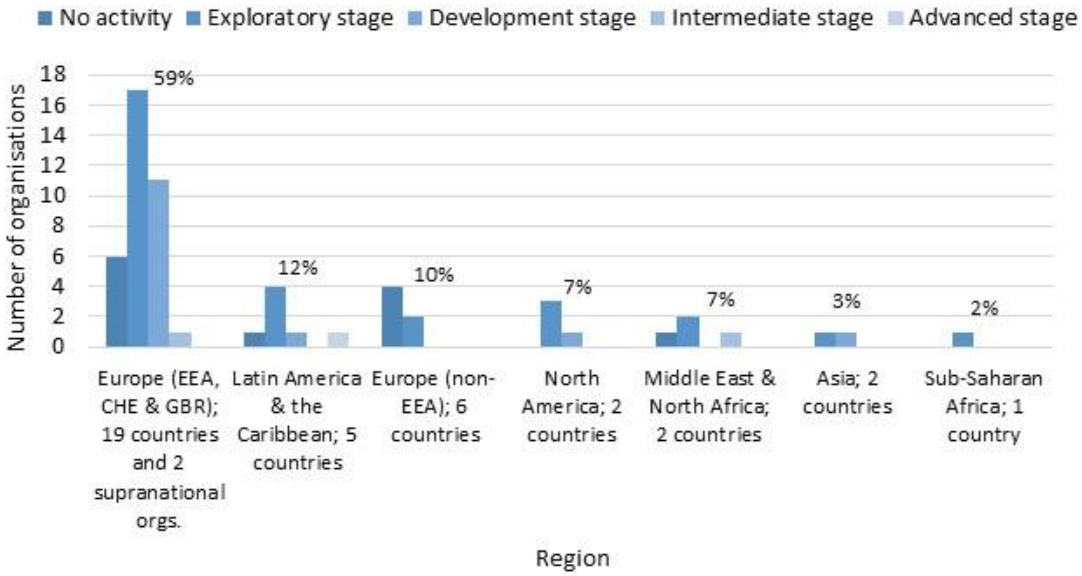
approaches into their auditing processes, as well as enhancing their knowledge for auditing emerging areas in government, such as the deployment of AI.

In all, 59 organisations from 39 countries responded to the OECD's questionnaire. Respondents predominantly represented European countries, including 19 countries from the European Economic Area (EEA), Switzerland and the United Kingdom (UK), 2 supranational organisations in the European Union, as well as six non-EEA countries (i.e. from EU candidate and neighbourhood countries). Of the organisations representing the countries from the EEA, including the two supranational organisations, 34% (12) reported having reached the development stage (11) or intermediate state (1), although this number was 0% in non-EEA, European countries (see Figure 1.3). Among the five countries represented in the responses from Latin America, the two organisations that indicated a level of maturity at the development and advanced stages were both from Brazil. The only organisation of the 59 respondents that reported an advanced stage of generative AI and LLM use was based in Brazil. In other regions, while the number of responses was low, countries generally were in the exploratory or development stages.

This figure is not meant to allow for drawing comparisons about the digital maturity of integrity actors in different regions. As noted, the questionnaire only covered a subset of the global population of integrity actors, so any conclusions about digital maturity are only representative of the pool of respondents. Judging from exchanges between the OECD and integrity actors during the course of this analysis, it is likely that many other organisations that received the questionnaire chose not to respond because they did not have any activities or discussions concerning the use of generative AI whatsoever.

Figure 1.3. Maturity of generative AI and LLM use by region

Which of the following options best describes the maturity of your institution's use of Gen AI and LLMs specifically, as a sub-domain of AI? (*The percentages refer to the proportion of organisations from each region out of a total 59 organisations that responded to the questionnaire.*)



Note: "EEA" is European Economic Area, "CHE" is the country code for Switzerland and "GBR" is the country code for the United Kingdom. Possible responses included the following: 1) Advanced Stage: Gen AI is deeply integrated into our core operations and we continuously seek ways to improve and expand its use. 2) Intermediate Stage: Gen AI is used in several areas of our activities, but it is not yet fully optimised or widespread. 3) Development Stage: We have experimented with Gen AI in a few projects but it is not yet integrated into our operations. 4) Exploratory Stage: We do not use Gen AI in our operations but we are currently exploring potential uses. 5) No Activity: We do not use Gen AI in our operations and we are currently not exploring potential uses.
Source: OECD questionnaire.

Responses to the questionnaire suggest that digital maturity is higher concerning the broader use of AI than it is for generative AI specifically, suggesting that countries are employing strategic approaches to exploring and deploying AI use. Specifically, around 34% (20) of organisations are currently developing a strategy for the use of AI in their institution, while several others follow a government-wide strategy for the use of AI. Six institutions currently have an AI strategy in place, all of which were from EU countries with one exception.

The efforts of these integrity actors illustrate the value they place on formally recognising the need for a strategic approach to exploring and deploying AI. Having a digital strategy with clear goals, objectives, performance indicators and defined roles and responsibilities, among other features, is a critical aspect of digital maturity (see Annex 1.A), and an AI strategy is often a subset of such a digital strategy. As one example, Box 1.1 describes the efforts of the Netherlands to incorporate generative AI into its broader AI strategy as well as the work of public bodies, including those responsible for anti-corruption. Generative AI can also be incorporated into the strategies of specific institutions. For example, Norway's Office of the Auditor General (OAG) envisions increased use of AI in performance audits in its 2018-2024 Strategic Plan (Office of the Auditor General of Norway, 2018[5]).

Box 1.1. The government-wide vision on generative AI of the Netherlands

The Netherlands became one of the first countries to publish a strategy focused specifically on generative AI in January 2024. The government-wide vision on generative AI outlines the opportunities and challenges posed by generative AI, elaborates a vision for the use of generative AI in the public sector based on four principles, and establishes specific actions to ensure public sector generative AI use is responsible and effective. This strategy provides an example of how integrity actors can benefit from a broader strategic approach to generative AI in the public sector.

The four principles to guide the development of generative AI, as outlined in the strategy, are as follows:

1. Generative AI is developed and applied in a safe way
2. Generative AI is developed and applied equitably
3. Generative AI that serves human welfare and safeguards human autonomy
4. Generative AI contributes to sustainability and prosperity

Opportunities discussed in the strategy include generative AI's potential to automate administrative and legal processes, serve as a learning tool, and even solve problems requiring complex data analysis with many inputs. On the other hand, risks include the impact on citizens relating to bias and privacy, increased dependence on foreign tech companies with monopoly power, exacerbating job insecurity, and the proliferation of mis- and disinformation. Both sides of this issue are relevant for integrity actors. For example, the amount of complex data analysis required of many of these actors means that generative AI presents notable opportunities, while the sensitivity of this data means that risk mitigation is also necessary.

Moreover, some of the actions laid out in the strategy explicitly highlight the participation of integrity actors. For instance, the strategy advocates for pre-deployment audits of advanced models and assigns an action to the Ministry of Foreign Affairs to promote this practice—along with responsible use of generative AI more broadly—on the international stage. The action plan envisions using generative AI for legal and administrative processes and analysing large datasets, which would be relevant for integrity actors. Beyond this, since the Netherlands is taking a whole of government approach, all actions taken will support the responsible deployment of generative AI in integrity bodies as a subset of the public administration.

Source: (Government of the Netherlands, 2024[6])

1.3. Opportunities and benefits of LLMs for integrity actors

The OECD supports integrity actors in government to build their technological capacity and develop data-driven methodologies for assessing fraud and corruption risks. The digital maturity of these partner organisations varies widely, with a small group implementing advanced analytics and a larger group relying more on qualitative risk assessments. The work of other organisations reflects a similar reality where risk assessments typically involve manual analysis, which can be time-consuming, resource-intensive and inefficient, often relying on specific complaints or anecdotes (World Bank, 2023[7]). Advancements in the ability of governments to harness technology, data and analytics, as well as ever-evolving AI methodologies, are challenging this status quo.

While it may not be the norm, integrity actors in the public sector have for years successfully leveraged advanced analytics and AI, such as supervised machine learning, to uncover hidden patterns and anomalies that indicate potential corrupt or fraudulent behaviour. For instance, supervised machine learning helped the General Comptroller of the State Administration of Spain (*Intervención General de la Administración del Estado*, IGAE) to detect fraud and corruption by leveraging proven cases as training data, enabling the model to learn and identify complex patterns and anomalies in public grants indicative of fraud (OECD, 2021[8]). OECD members and partners across the globe, including public integrity partners from Brazil, Colombia, Korea, Lithuania and the United States, are advancing similar efforts (OECD, 2022[4]; OECD, 2021[8]). AI and data-driven assessments enable organisations to proactively mitigate risks and safeguard taxpayer money in ways that are more efficient and impactful than more manual approaches, while allowing for wider covering of the risk universe.

1.3.1. There are a variety of applications for LLMs in the integrity and anti-corruption space

The advent of generative AI, and in particular LLMs, creates new avenues for integrity actors to enhance the efficiency and impact of their work. This paper provides examples of some of these opportunities, which broadly cover two dimensions: the organisation's internal operations, and more specifically, anti-corruption and anti-fraud activities. Based on responses to the OECD's questionnaire, LLMs are a main focus of integrity actors' current exploration with generative AI, so much of the paper concentrates on this technique.

LLMs are well-suited to support integrity actors in automating certain fraud detection activities, such as querying documents and data sources for potential risk. LLMs can also help auditors and investigators to carry out many operational tasks that, while not unique to integrity actors, are particularly promising given the high volumes of documentation and data that audit, anti-corruption and investigative bodies typically process. For instance, LLMs can help to organise large volumes of text for easier prioritisation and consumption, and aid in root-cause analyses or pattern recognition (U.S. Government Accountability Office, 2024[9]). Some countries such as Sweden are developing government-wide virtual assistants that would help streamline these operational tasks in all public bodies, including integrity bodies (AI Sweden, 2024[10]). The efficiencies gained by these techniques can reduce both effort and error, allowing auditors and investigators to focus more on high-value tasks that require human judgement and expertise, which generative AI has yet to replace. By making anti-corruption and anti-fraud activities more effective, generative AI can also strengthen public integrity and accountability.

Academia offers additional inspiration for integrity actors to apply LLMs. For instance, financial and accounting literature provides numerous examples of using LLMs to assess financial texts. One group developed an LLM called FinBERT, based on Google's Bidirectional Encoder Representations from Transformers (BERT) algorithm and a large corpus of financial texts, for sentiment analysis and extracting specific discussions about environment, social and governance (ESG) (Huang and Yi Yang, 2023[11]). Another group of researchers took a case study approach and explored the adoption of ChatGPT by a

multinational company's internal audit function (IAF) across various stages of the audit process, including risk-based audit planning, audit preparation and data analysis. In this instance, the IAF observed promising results in tasks that involved scoping audits, brainstorming risks, drafting descriptions, interview preparation and report writing (Emett, 2023[12]). These texts highlight opportunities, but they also warn of risks and elaborate on challenges of deploying LLMs, some of which are covered in Section 2.

LLMs also have the potential to promote integrity in public spending if adopted by a broader range of actors that do not fit the definition of integrity actors used for this paper. For instance, LLMs, such as those that power ChatGPT, can support public procurement officials in analysing large amounts of data on a company and potential contractor to screen for fraud or corruption risks. One organisation that responded to the questionnaire highlighted the development of a pilot project to continuously identify risk indicators in public procurement processes using LangChain and an LLM to preprocess the unstructured data.[3] The organisation executes the preprocessing phase centrally, while leveraging the expertise of auditors in a more decentralised manner to provide prompts that pinpoint procurement features of interest.

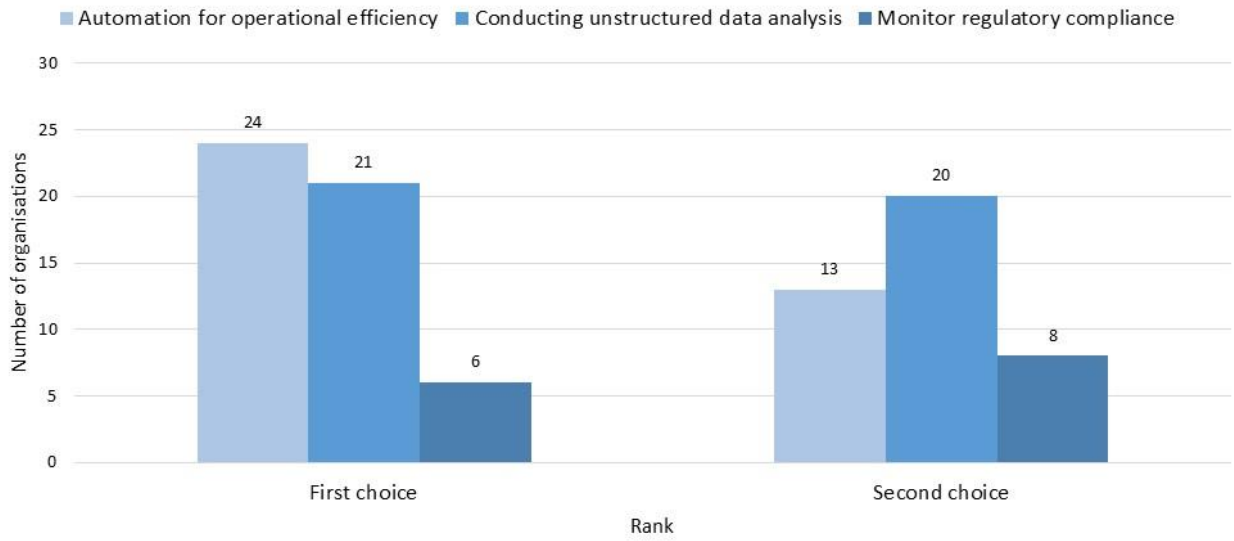
1.3.2. LLMs can enhance the internal operations of integrity actors, with gains in operational efficiency and analysing unstructured data being the most promising opportunities

The OECD asked integrity actors where they think generative AI, LLMs in particular, can add the most value for their organisation and its activities. As noted, the question focused on opportunities in two areas: 1) the organisation's internal operations; and 2) its anti-corruption and anti-fraud activities. These areas, as well as the options for responses, are difficult to artificially separate and they may not be mutually exclusive. Gains in one area of analysis or information processing can lead to efficiencies in others. With that in mind, OECD questionnaire respondents ranked operational efficiency and unstructured data analysis as the areas that could benefit the most from the use of generative AI and LLMs for internal operations (the first choice of 45 of the 59 organisations, or 76%). Figure 1.4 shows the areas of added value ranked first and second, with additional information about the various areas of internal operations surveyed.

A small group of integrity actors ranked monitoring of regulatory compliance as the main area of perceived value of generative AI and LLMs. Respondents viewed the contributions of generative AI to public engagement and transparency and training and capacity building as comparatively smaller. When breaking down the data by organisational type, over half of the 16 SAIs (9) that responded to the questionnaire ranked unstructured data analysis as the number one potential benefit of LLMs, but there were no other significant trends or patterns in the data by organisational type.

Figure 1.4. Perceived benefits of generative AI and LLMs for integrity actors' internal operations

Within your institution, which of the following areas of internal operations would benefit the most from the use of Gen AI and LLMs?



Notes: Possible responses included the following: 1) Operational Efficiency: Streamlining internal processes by automating routine tasks for core activities, allowing for more efficient allocation of human resources. 2) Unstructured Data Analysis: Leveraging Gen AI to effectively analyse and interpret unstructured data, such as text, images, and audio, which can provide deeper insights and inform decision-making processes. 3) Public Engagement and Transparency: Using LLMs to streamline communication with the public and stakeholders. 4) Training and Capacity Building: Using LLMs for training purposes, such as planning curricula and workshops. 5) Regulatory Compliance Monitoring: Employing Gen AI to continuously monitor and ensure compliance with relevant laws and regulations, reducing the likelihood of non-compliance issues. 6) Not sure. 7) Other.
Source: OECD questionnaire

Respondents also offered their views about the value of generative AI, includingLLMs, for investigative and audit processes. They ranked gathering evidence and document review as having the highest value in this respect. Specifically, 37% of respondents (22) ranked these activities as their top choice, followed by the use of generative AI and LLMs for selecting audits and investigations (ranked first by 25%, or 15 organisations). This was particular the case among SAIs and internal audit bodies, which as a group, ranked evidence gathering and document review higher relative to other integrity actors. As far as the value of generative AI and LLMs for other audit and investigative activities, fewer integrity actors ranked the following options at the top: drafting reports and producing graphics (7); none of the above/not sure (6); planning audits and investigations (5); generating content for public relations (3); and documenting processes (1). The initiative of Brazil's SAI (*the Tribunal de Contas da União*, TCU) to develop ChatTCU illustrates one approach for leveraging LLMs to enhance the efficiency with which auditors gather and review documentation. Box 1.2 describes the initiative and offers key lessons learned, many of which are broadly applicable to other types of organisations, even though ChatTCU is an SAI-led initiative.

Box 1.2. Lessons from Brazil's SAI and the development of ChatTCU

In February 2023, the Brazilian Federal Court of Accounts (TCU) launched ChatTCU based on OpenAI's ChatGPT. The TCU built the tool based on the view that LLMs are not a passing trend, and therefore it decided to take an institutional approach to consciously developing use cases while addressing potential risks. While the initiative is still developing, the TCU has already experimented with several applications. As of December 2023, the TCU reported over 1 400 users, demonstrating the extent to which the tool has been rolled out and adopted.

The current version of ChatTCU is integrated with TCU's systems, providing answers based on the Court's cases, selected precedents, and administrative system, coupled with the knowledge base of ChatGPT itself. For instance, ChatTCU allows auditors to request a summary of a case document, pose technical questions related to TCU's work and court decisions, and seek help for administrative services. ChatTCU v3 is based on GPT-4 32k, which grants better quality to the answers provided and fewer chances of errors or hallucinations.

The TCU plans to incorporate a range of new features, such as further integration with other systems and workflow automation through user prompts. TCU hosts ChatTCU on a dedicated instance of Microsoft Azure's cloud platform. This helps TCU to ensure the security and confidentiality of its data, and it allows auditors to use the tool without sending private data to OpenAI. Furthermore, hosting ChatTCU in this way helps facilitate integration with other systems. Key lessons learned from the TCU's experience, many of which are transferable to other integrity actors include:

- **Internalise technology**: The proactive development of ChatTCU, tailored to TCU's needs, suggests that integrity actors could consider building their own AI solutions rather than relying solely on external tools. This also helps build the digital literacy and capacity within the organisation that will prove valuable in other areas.
- **Integration with existing systems:** The integration of ChatTCU with TCU's existing systems allowed auditors to access administrative information and gain insights into audits more efficiently, underscoring the importance of seamless integration with existing workflows and systems.
- **Scalability and future-proofing:** TCU's plans to expand ChatTCU's functionalities demonstrate the need for scalability and adaptability in AI solutions, urging integrity actors to plan for future upgrades and developments.
- **Potential for standardisation:** TCU's consideration of incorporating audit standards into ChatTCU indicates the potential for AI tools to assist in maintaining standards, suggesting that other integrity actors may explore similar possibilities to enhance their processes.
- **Feedback-driven development**: TCU underscored the importance of collecting user feedback to continuously improve AI solutions, emphasising the need for integrity actors to create mechanisms for staff to provide feedback and suggestions for enhancements.
- **Multidisciplinary approach:** TCU formed a multidisciplinary working group to assess the risks and opportunities of using generative AI, involving representatives from various areas and promoting debates to help make informed decisions about AI implementation.
- **Invest in training and awareness**: TCU's emphasis on raising awareness among staff about the potential and risks of AI highlights the crucial need for training and educating staff members on how to effectively use AI technologies. Involving staff in developing AI solutions internally will also help them learn how to tackle these challenges firsthand.

Source: Responses to the OECD's questionnaire and https://portal.tcu.gov.br/en_us/imprensa/news/chattcu-integration-of-the-tool-into-the-courts-systems-improves-the-use-of-generative-artificial-intelligence-in-external-control-activities.htm

1.3.3. Generative AI and LLMs show promise for strengthening a variety of anti-corruption and anti-fraud activities, but examples in government are limited and the return on investment is unclear

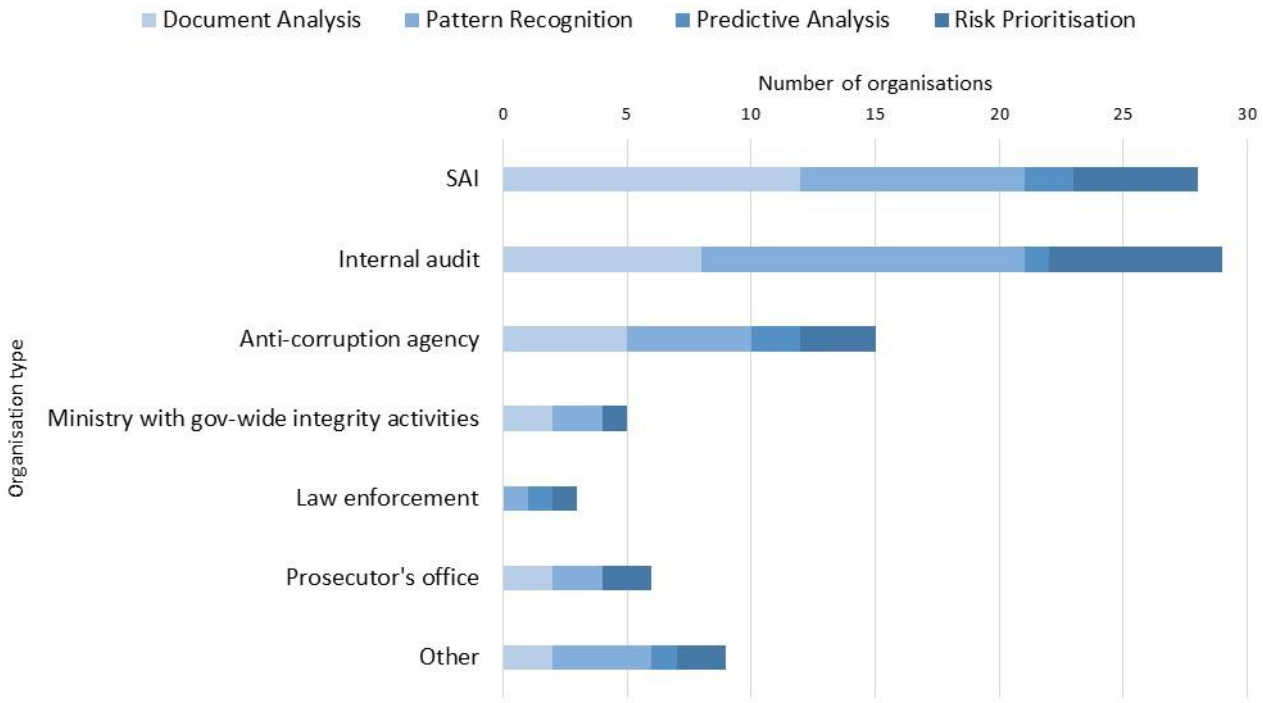
The OECD also asked respondents for their views on the perceived value of generative AI, including LLMs, for anti-corruption and anti-fraud activities. Document analysis and pattern recognition had the highest perceived value with most respondents ranking these activities as either their first or second option. They ranked risk prioritisation as third (15%, or 9 respondents). Following those top 3 selections, in smaller numbers, respondents chose either none of the above/not sure, developing training and simulation tools and conducting predictive analytics to prevent and detect fraud and corruption. ACAs showed the most interest in pattern recognition, with half of the 10 ACAs that responded to the questionnaire ranking this at the top of the list of perceived value of generative AI and LLMs. Several questionnaire respondents highlighted the fact that their mandate does not allow them to conduct anti-fraud activities.

Figure 1.5 summarises the top four responses by type of organisation for the perceived benefits of generative AI for anti-corruption and anti-fraud activities, and it provides definitions for different activities in the questionnaire. Some of the activities could overlap in practice and raise additional questions for future inquiry and research. For instance, activities to detect patterns and anomalies, which many respondents ranked as their top choice, could also inform risk prioritisation, which generally was ranked lower by most respondents. There could be practical reasons for this. Experimenting with LLMs and analysing unstructured data draws from finite resources. OECD's projects with integrity actors demonstrate that many organisations are already investing resources in response to other technological trends (e.g. "big data" analysis), including how to make better use of structured data for assessing risks.

Moreover, as an activity, risk prioritisation is an obvious candidate for experimenting with generative AI, but as an institutional process, it can already have its own set of established procedures, processes and tools in place. These are developed and refined over decades. In this context, whether it is an SAI, ACA, law enforcement body or other integrity actor, any new LLM-supported methodology would need to be thoughtfully designed and integrated if it is to become part of, or potentially disrupt, the status quo. This poses challenges that are not just technical in nature, but also organisational, legal and even political (internally), which may help to explain why many respondents did not rank risk prioritisation higher.

Figure 1.5. Perceived benefits of generative AI and LLMs for anti-corruption activities by type of organisation (top two choices)

Within your institution, which of the following anti-corruption or anti-fraud activities would benefit most from the use of Gen AI and LLMs?



Note: The numbers in the chart refer to the number of times an institution mentioned each activity as either their first or second ranked choice. Possible responses included the following: 1) Pattern Recognition: Identifying unusual patterns or anomalies in data that may indicate corrupt or fraudulent activities. 2) Document Analysis: Automating the review of large volumes of documents for potential corruption or fraud indicators. 3) Risk Prioritisation: Assisting in risk assessment and prioritisation of investigations based on AI-generated insights. 4) Predictive Analysis: Using LLMs for predictive analytics to anticipate and prevent potential corrupt or fraudulent activities. Other options not illustrated include: 5) Training and Simulation: Providing training and simulation tools to staff for better understanding and detection of corruption/fraud; 6) None of the above/not sure; and 7) Other.
Source: OECD questionnaire

One respondent with government-wide integrity and anti-corruption activities (i.e. a Ministry of Justice) highlighted the processing and review of asset declarations as one specific area of need and potential value of LLMs. Echoing the OECD's experience with many organisations responsible for asset declaration systems, the respondent described a high-volume of checks required for verifying the content of asset declarations. These checks are currently done manually in most countries. Having a tool to enhance the processing of these declarations is not just about creating more efficient processes and procedures. Such solutions would ultimately contribute to greater transparency in government and enhance public awareness about conflicts-of-interest concerning public officials.

Annex 1.A. Key dimensions for assessing institutional digital maturity

The dimensions and key practices below are based on reviews of academic literature, discussions with subject matter experts in government, industry and non-governmental institutions, insights from the OECD's technical support for governments to strengthen their digital strategies and data-driven risk assessments, as well as OECD Recommendations.[4] Numerous self-assessment tools for digital or technology readiness that are relevant or made for integrity actors at an institutional level, such as the Supreme Audit Institutions (SAI) Information Technology Maturity Assessment, also provided inspiration. These practices consider digital maturity and transformation from an organisational perspective, but many are applicable for designing and implementing digital projects.

Strategy and organisation

This dimension encompasses leadership's vision and strategy for digital transformation, including its goals for strengthening the use of digital technologies and data. A digital transformation strategy can stand alone or be integrated with existing organisational strategies. Either way, the aim is to ensure alignment of the digital strategy with other organisational priorities, audit processes and IT strategies (Bumann and Peter, 2019[13]). Clear delineation of roles and charting out responsibilities is imperative. This can include establishing an entity internally with an organisation-wide mandate to implement and co-ordinate digital initiatives. Data management and data governance are also key aspects of this dimension. They involve the policies, procedures, standards and controls that ensure data privacy, quality, consistency and security. Relative to project-based improvements, digital transformation by nature has a disruptive effect on an organisation's traditional approaches to data management and data governance.

Key practices in this dimension include:

- Align the Digital Strategy with the goals and objectives of other institutional strategies, such as the Strategic Plan and IT Strategy.
- Conduct assessments and establish a baseline for digital maturity, capabilities, IT infrastructure and architecture, and possible gaps.
- Identify key opportunities and challenges concerning data management and data governance, including priorities for ensuring data security and quality.
- Define roles and responsibilities internally, including the designation of an entity to implement the Digital Strategy that has an organisation-wide mandate and access to leadership.
- Engage with key stakeholders in the design of the Digital Strategy, including leadership, management and users of new tools, to understand digital maturity and priorities.
- Establish a plan and key performance indicators to monitor the implementation of the Digital Strategy.

People and culture

The expertise, skills and commitment of individual employees within an organisation are central to digital maturity on any level, whether the goal is transformation or introducing a new tool for using data. Core competencies often revolve around digital and data literacy, sometimes extending to advanced programming skills. Alongside these technical proficiencies, it is critical that employees understand the policies, processes and behaviours that promote the ethical use of data (OECD, 2020[14]). Furthermore, sector-specific knowledge and specialised expertise are also critical competencies, such as having sector-specific knowledge to understand the data landscape for risk analyses. Legal expertise is also valuable for navigating the legalities of data access, privacy, storage, and security. A digital-ready culture is not only about having the right set of skills and experiences available, but it demonstrates tangible ways that leadership and employees rally around digital goals. This can manifest in different ways, such as having policies that allow for the experimentation of new technologies or providing training for employees to improve their digital skills (OECD, 2022[4]).

Key practices in this dimension include:

- Ensure that leadership visibly endorses and partakes in digital initiatives, embodying a top-down commitment to the organisation's digital aspirations.
- Develop and implement a change management and continuous learning plan that focuses on enhancing digital and data literacy, as well as sector-specific knowledge.
- Introduce and encourage training programmes targeting technical proficiencies like advanced programming and data ethics.
- Institute clear policies that favour experimentation with new digital tools and technologies to foster innovation and a “trial-and-error” mentality.
- Establish guidelines on the ethical use of data, ensuring that staff understands and adheres to them.
- Prioritise and establish mechanisms for internal knowledge sharing, facilitating the dissemination of sector-specific, technical and legal expertise.
- Promote a culture of collaboration and digital empowerment, where employees at all levels feel engaged and invested in digital transformation objectives.
- Collaborate with legal experts to navigate the intricacies of data laws, ensuring the organisation remains compliant while maximising its digital potential.
- Implement feedback loops to understand employee challenges and needs in the digital landscape, adjusting strategies based on this feedback.
- Regularly evaluate the digital skills gap within the organisation and adjust training programmes accordingly.

Technology and processes

While technology, including IT systems, tools, and software, and the processes encompassing them are vital components of digital maturity, they should not be perceived as the primary objectives. The broader vision for digital transformation or the intent of any given project should inform technological advancements rather than being led by them. This underscores the importance of tailoring technology to specific needs. The term "state-of-the-art" is contextual, acknowledging that digital services, IT mechanisms, and tools differ in their complexity, resource needs, functionality, and alignment with various organisational goals. Given the rapid evolution of technology, expending resources without a lucid objective can lead to a waste of resources. Thus, a pragmatic approach involving cost-benefit analysis can guide judicious decision making about technological investments. This analysis can consider collaborative investments or

leveraging open-source technologies that may offer additional public benefits, such as the promotion of systemic transparency and collective technological development. Moreover, it is critical that considerations about adopting new technologies include assessments of their impacts on society, human rights and privacy, among other issues, to avoid exacerbating risks of discrimination and digital exclusion.

Key practices in this dimension include:

- Ensure any technology adoption aligns with the strategic objectives or specific goals of the organisation.
- Understand current capabilities, identify gaps, and ensure alignment with the organisation's digital maturity and objectives.
- Before investing in any new technology, gauge its potential return on investment and long-term sustainability.
- Start with a minimum viable product or proof-of-concept to test and validate new technologies or digital tools.
- Given the fast-paced nature of technological evolution, adapt and update tools and systems based on changing needs and feedback.
- Regularly research and explore new technological advancements that could replace legacy systems and enhance organisational effectiveness and efficiency.
- When selecting technologies, consider how easily they can be scaled or adapted to changing organisational needs or goals.
- Ensure that technologies are user-friendly, meet the needs of the organisation, and are accessible to all relevant stakeholders.
- Ensure that any new technology or process integrates robust cybersecurity protocols to safeguard organisational data.
- Define roles, responsibilities, and decision-making processes related to technology adoption and usage.
- Facilitate channels for sharing best practices, lessons learned, and feedback regarding technology tools and processes.
- Establish mechanisms to assess the social, human and ethical impact of adopting new technologies and mitigate associated risks, including those concerning data privacy, discrimination and digital exclusion.

Environment and partnerships

National frameworks, ranging from laws to directives, can considerably shape the trajectory of institutional digital transformation. For instance, a country's legal and policy framework can lay the foundation for managing digital governance, data stewardship and the sharing of data. These external parameters can influence the effectiveness of a digital transformation project, either limiting or propelling the adoption of digital technologies. Within an organisation, while robust data governance streamlines the sharing and accessibility of data, the true essence of data sharing transcends just infrastructure or processes. Cultivating collaborative relationships between entities, including industry, academia and civil society organisations, is an indispensable cornerstone for advancing digital maturity and ensuring that necessary safeguards are in place.

Key practices in this dimension include:

- Stay abreast of updates to knowledge of laws, policies, and guidance related to digital governance, data management, and sharing.
- Engage with policymakers to advocate for supportive laws, regulations and policies that bolster the goals of digital initiatives.
- Establish data-sharing protocols that align with both internal goals and external legal requirements, allowing for efficient and timely exchange of information.
- Define institutional roles, responsibilities and expectations for all digital initiatives that involve collaboration with external stakeholders.
- Ensure that partnerships are mutually beneficial, fostering a sense of shared ownership and collective achievement.
- Establish channels to gather feedback from partners, ensuring continuous improvement in collaborative endeavours.
- Encourage an organisational mindset that values partnerships as a key enabler of digital growth.
- Establish relationships with other organisations, both within and outside the government (e.g. industry, academia, civil society), to promote collaborative digital initiatives, share best practices, and ensure ethical use.

2 Generative AI: Challenges, risks and other considerations for integrity actors in government

Many of the opportunities and benefits that generative AI and LLMs offer, as discussed in Section 1, come with a unique set of challenges, risks and technical considerations. The 59 integrity actors that responded to the OECD's questionnaire highlighted issues that are relevant for their own context. However, many of the challenges and concerns they raised regarding the development, deployment and scaling of LLMs are relevant across different types of organisations and regions. This section explores these aspects of generative AI, particularly LLMs. It provides insights that can help integrity actors to understand and anticipate the range of challenges that this new area presents and be better positioned to overcome them if and when they arise.

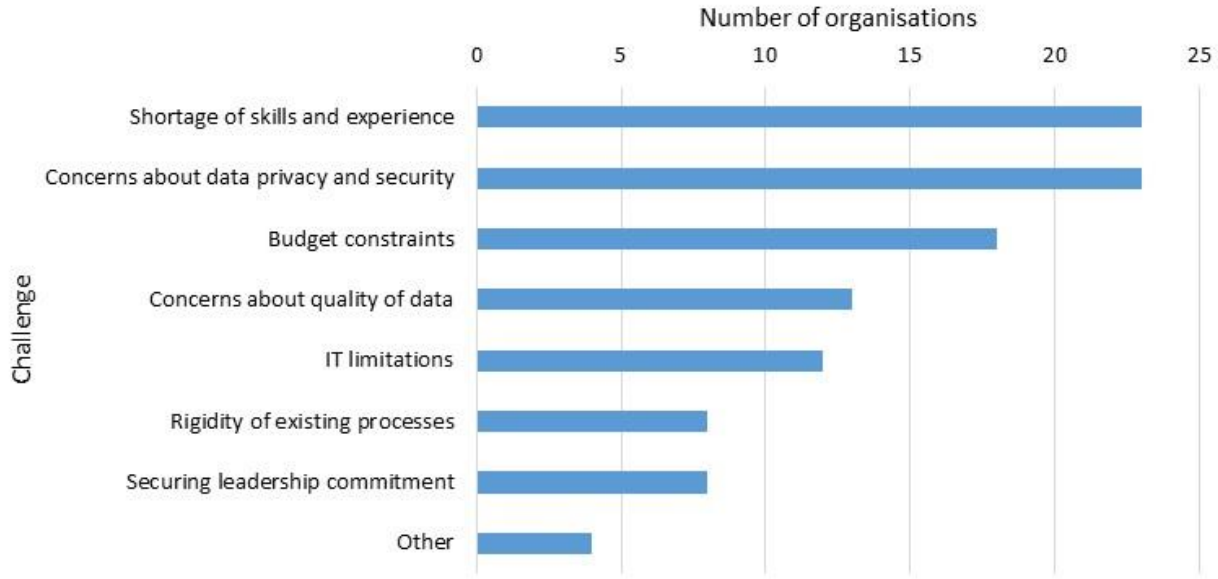
2.1. Overview of main challenges for integrity actors to adopt generative AI and LLMs

2.1.1. Integrity actors cited a shortage of skills and IT limitations as the biggest challenges they face to implement generative AI and LLMs

The OECD asked questions to understand the nature of challenges that integrity actors face to adopt generative AI and LLMs. Shortage of skills and experience ranked at the top of organisations' concerns, and this was the main challenge identified by anti-corruption agencies (see Figure 2.1). Organisations identified challenges related to preserving data privacy and security just as frequently; this issue was of particular concern to SAI respondents. Budget constraints, quality of data and IT limitations were also flagged as either the greatest or second greatest challenge by at least 10 of the organisations. Relatively fewer respondents highlighted concerns about the rigidity of existing processes or securing leadership commitment. One SAI noted that creating a business case for using and integrating generative AI into its operations is a challenge.

Figure 2.1. Main challenges for deploying generative AI and LLMs

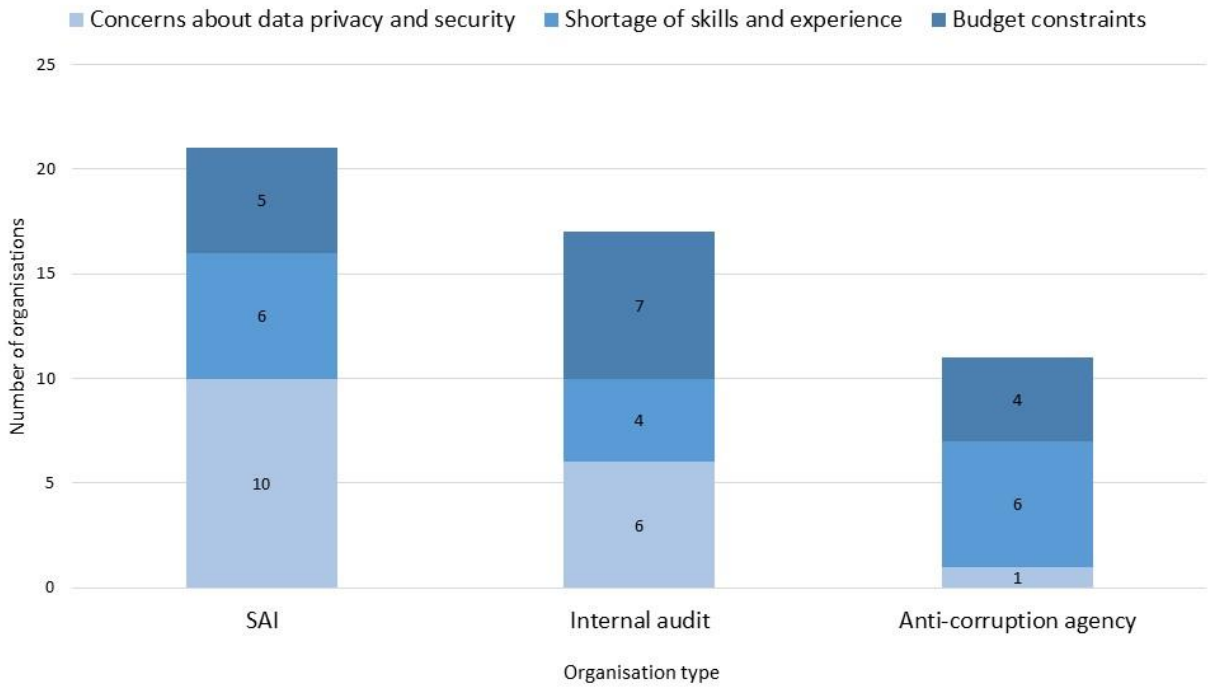
What are the biggest challenges your institution faces concerning the adoption of Gen AI and LLMs in general?



Note: "Number of organisations" refers to the number of organisations that selected each challenge as either their greatest or second greatest concern. Possible responses included the following: 1) Shortage of skills and expertise; 2) Concerns about data privacy and security; 3) Budget constraints; 4) Concerns about the quality of data inputs and outputs (e.g. biases and "hallucinations"); 5) IT limitations for developing and maintaining LLMs (e.g. IT systems and computing capacity); 6) Rigidity of existing structures or processes; 7) Securing leadership commitment and support; and 8) Other.
Source: OECD Questionnaire

Concerns about budget constraints were somewhat more pronounced among internal audit bodies relative to SAIs and ACAs, taking into account responses as a percentage of the total number of institutions by type (see Figure 2.2). Many institutions expressed that due to resource constraints, they either lack the sufficient financial, human, and technical resources needed to employ LLMs entirely or their staff does not have sufficient data literacy to make the use of such tools possible. One category of challenges the questionnaire did not clearly capture was that of methodological limitations. For instance, one ACA noted that its biggest challenge was having sufficient data to be able to develop an LLM. This points to a hierarchy of needs when it comes to developing LLMs. Given many institutions' early stage of development, in practice, some institutions appear to be focused on more technical challenges of developing viable proof-of-concepts, while recognising other challenges lie ahead (e.g. ensuring data privacy and security), particularly as they scale and roll-out LLMs.

Figure 2.2. Main challenges for deploying generative AI and LLMs by type of organisation



Source: OECD questionnaire

Tailored education is pivotal for overcoming challenges associated with skill and expertise deficits for using generative AI and mitigating associated risks. As illustrated by the experience of the European Court of Auditors (ECA), tailoring trainings involves adapting the curriculum to make courses available that are dedicated to generative AI and LLMs. Additionally, it means ensuring training content illustrates concrete uses cases and links generative AI tools to processes that are familiar to the trainees, which in the ECA's case would be auditors. Box 2.1 further describes the ECA's initiative to develop its trainings for generative AI.

Box 2.1. The generative AI training programmes of the European Court of Auditors (ECA)

The ECA is the supreme audit institution (SAI) of the European Union (EU) and is responsible for auditing the EU's finances as well as co-ordinating good practices across the SAIs of the 27 EU member states. The ECA is exploring how generative AI can be employed to make its audits more efficient and effective. As of February 2024, the ECA has developed two trainings on generative AI and is preparing several more. The trainings were developed in response to increasing demand among staff for guidance on how to employ generative AI tools in light of ChatGPT's growing popularity.

The ECA therefore developed an introductory training on generative AI that covers both how it works and ways in which existing generative AI tools can be used in auditing. It also offers advanced training where staff can develop their own machine-learning tools. The ECA has repeated the introductory training in response to high demand from staff, demonstrating that staff are eager to employ these tools once they have the proper knowledge.

A key distinguishing feature of the trainings offered by the ECA is their focus on integrating examples from existing audit work. The trainings explain how generative AI could have been used at different stages of past audits that staff are already familiar with, which promotes an understanding of the benefits and risks of generative AI on a practical basis, rather than a theoretical one. The training also teaches staff how to critically evaluate the outputs of generative AI.

The ECA is planning to develop more trainings based on areas of high demand and/or high risk. These include legal and copyright risks related to generative AI, conducting cybersecurity audits using AI, and including AI-based risks in IT audit methodologies. A training on prompt engineering in the context of generative AI is also under consideration.

Source: OECD interview with the European Court of Auditors

Concerning regional challenges highlighted in the questionnaire responses, several institutions in EU countries highlighted the need to ensure compliance with the General Data Protection Regulations (GDPR) and the EU's AI Act. The GDPR restricts the terms under which organisations in EU countries can reuse personal data, namely by requiring user consent. Understanding the full impact of the GDPR on AI or the integrity actors' use of it is beyond the scope of this paper. Nonetheless, respondents to the OECD's questionnaire highlighted this issue as a key consideration in their implementation of LLMs, which has resulted in them taking a more cautious and deliberate approach. As discussed later, in the context of AI, integrity actors may also face tensions between the need for algorithmic transparency and protecting data privacy.

The European Parliament approved the AI Act at the time of writing this paper in March 2024, with a formal endorsement by the Council of the EU needed before it enters into force. The Act establishes obligations for AI developers and users based on potential risks and the level of impact of the AI system, with the aim of protecting fundamental rights, democracy, the rule of law and the environment from high-risk forms of AI (European Parliament, 2024[15]). The effect of the AI Act on government entities as users of AI, including integrity actors relying on foundation LLMs of private companies, remains to be seen as EU countries turn towards implementation and enforcement of the Act. Section 2.3 below explores themes related to the purpose of the EU's AI Act, including challenges and considerations for integrity actors concerning the promotion of trustworthy AI and responsible use of LLMs.

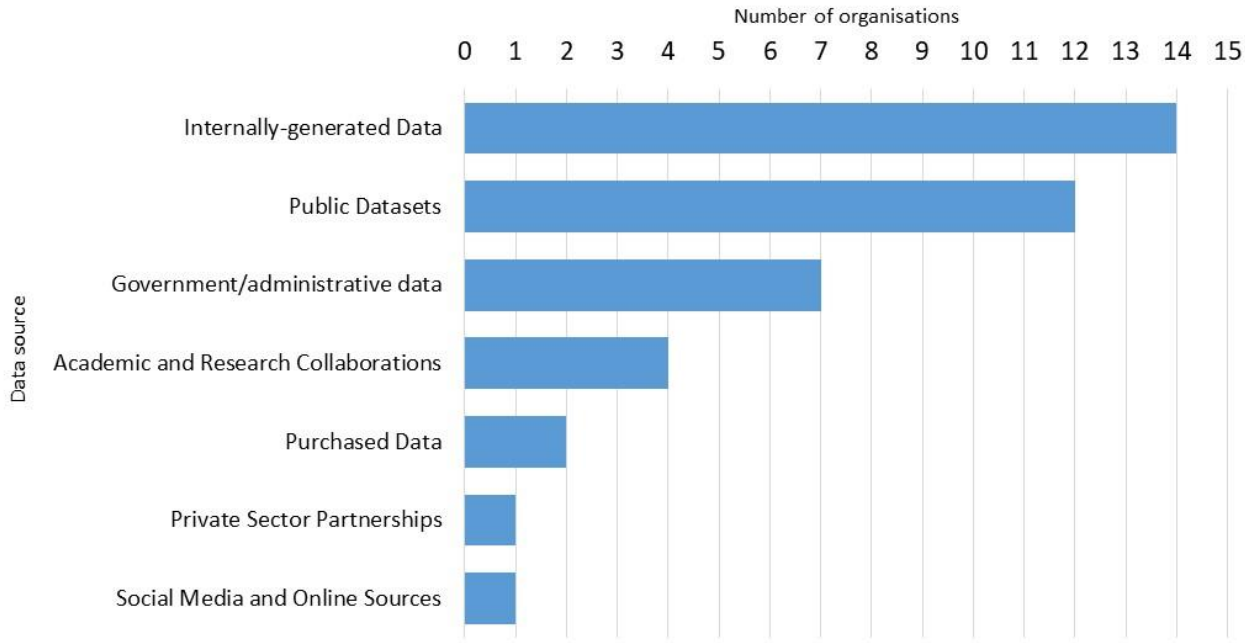
2.1.2. Advice for piloting LLMs includes first incorporating generative AI into low-risk processes and considering the requirements for scaling early on

The questionnaire asked specifically about challenges related to piloting and scaling generative AI initiatives, focusing on the 17 respondents that are either experimenting with or integrating generative AI tools. Many of the main challenges echoed those highlighted in the figure above, including shortage of skills, IT limitations and concern about the quality of data inputs and outputs (e.g. biases and hallucinations). Only one institution expressed concerns about securing leadership commitment, although respondents may have been less likely to select this option if superiors monitored their responses. However, since few institutions have reached this level of maturity and there were therefore fewer responses to this question, the spread in responses was fairly even. Several themes came to the forefront in the responses to questions about piloting and scaling generative AI initiatives:

- Start by incorporating generative AI into low-risk areas and processes. Many of the institutions that have reached the piloting stage seem to be focusing on incorporating generative AI, with a focus on LLMs, into relatively low-risk processes, such as document querying, writing document summaries and press releases, and answering user questions. Such an approach can help build capacity in areas where mistakes are less costly—either financially or from a compliance perspective—before they scale LLMs to riskier and more analytical tasks, including those that require more financial resources.
- Consider the IT requirements not only for piloting, but for scaling as well. When piloting LLMs, it is first necessary to establish certain prerequisites for IT infrastructure. This includes computational and storage resources, including the availability of high-performance computing power, data storage, and data management capabilities. Over half of the 17 respondents with LLM initiatives ranked this as the number one IT challenge, followed by challenges related to software tools as well as system scalability and integration.[5] One respondent noted that having the rights tools in place first is just as important as having the right algorithms.
- Consider internally-generated data to demonstrate value and establish quick wins. This data could be internally held and/or produced by the government body itself or come from another source. The integrity actors that responded to the OECD's questionnaire are primarily relying on internally-generated data or public open datasets, potentially because they view this approach as a lower risk than using other data sources (see the next section for a discussion on Retrieval-Augmented Generation). Comparatively fewer organisations are using other government data, while only a limited number of organisations are using data purchased from the private sector, obtained through a public-private or academic partnership, or obtained from social media or another online source (see Figure 2.3).

Figure 2.3. Primary data sources for building LLMs among questionnaire respondents

From which sources does your institution primarily acquire the data used for building and training your LLM(s)?



Note: Possible responses included the following: 1) Internally Generated Data: Data generated from within our own institution (e.g. reports, administrative records, etc.); 2) Public Datasets: Data sourced from publicly available datasets (e.g. government open data portals, public research datasets). 3) Government/administrative data: Data sources produced or owned by government entities, but are not public or open. 4) Private Sector Partnerships: Data obtained through partnerships or agreements with private sector entities. 5) Purchased Data: Data procured from commercial data providers or brokers. 5) Academic and Research Collaborations: Data obtained through collaborations with academic or research institutions. 7) Social Media and Online Sources: Data extracted from social media platforms, websites, and other online sources; 8) Other.
Source: OECD questionnaire

One notable difference between the challenges identified for piloting LLMs versus scaling them is the emphasis on both data privacy and security as well as budget constraints. In the initial phases of testing LLMs, the primary challenges highlighted involve data privacy and security, alongside concerns about data quality. Out of 17 organisations, only two mentioned budget constraints as an issue during this pilot phase. Conversely, when it comes to expanding the use of generative AI and LLMs, budget constraints emerge as a more significant challenge, with fewer organisations expressing concerns about data privacy and security at this stage. This may reflect an evolution in maturity in terms of managing data privacy and security issues, as well as the increased resource needs when scaling LLMs that are not present early on.

2.2. Building a generative AI and LLM capacity within institutions responsible for integrity and anti-corruption

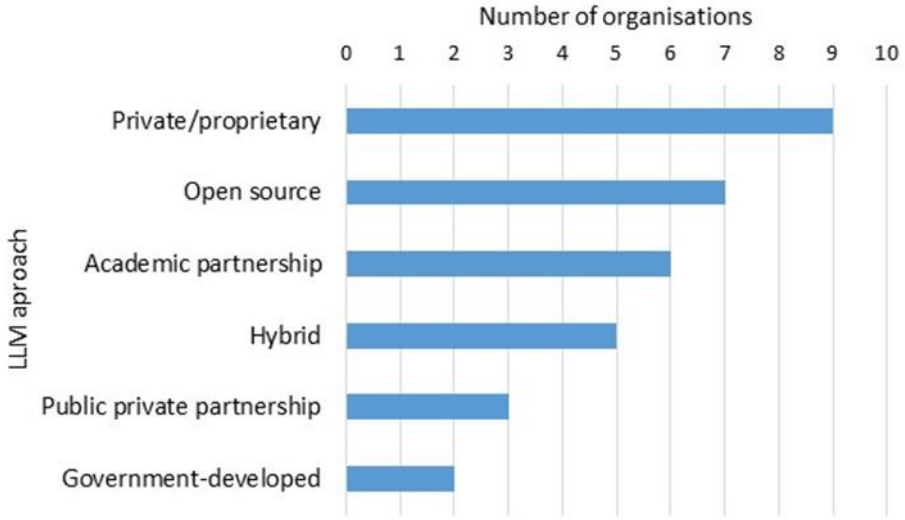
2.2.1. Integrity actors mostly rely on turnkey foundation LLMs developed by technology companies

Integrity actors have multiple pathways for piloting and scaling LLMs. These include leveraging open-source LLMs; utilising models developed by private companies for their advanced capabilities; or embarking on their own development projects. Collaborative efforts with the private sector or academia, as well as hybrid approaches that combine these elements, present viable alternatives. Of these options, open-source LLMs offer some algorithmic transparency. However, due to intellectual property restrictions, it can still be difficult for users to review their source code and training data, thus limiting the degree of transparency. Other mechanisms that promote interpretability, explainability, such as user-friendly explanations of decisions made, can help promote transparency in decision making internally and to the public at large (see Section 2.3).

The integrity actors that responded to the OECD's questionnaire predominantly use open-source and private sector models, which according to several respondents, helped to overcome constraints in financial and human resources (see Figure 2.4). Several of the integrity actors that reached the development stage of using LLMs (see Section 1) highlighted the use of multiple approaches to testing or using them.

Figure 2.4. Integrity actors' approach for using LLMs

What is the general approach your institution is taking to test and/or use LLM(s) for your operations?



Note: This question was only asked to questionnaire respondents who have reached the stage of developing generative AI models. Possible responses included the following: 1) Open Source Model: We use or develop LLMs based on open-source platforms or technologies. 2) Private/Proprietary Model: We use LLMs developed by private companies (e.g. ChatGPT by OpenAI). 3) Hybrid Model: We use a combination of open-source and private/proprietary LLMs. 4) Public-Private Partnership: Our LLMs are developed or used in collaboration with private entities under a public-private partnership model. 5) Government Developed and Maintained: Our LLMs are exclusively developed and maintained using government resources without private sector involvement. 6) Research and Academic Collaboration: We are engaged in collaborations with academic or research institutions for the development or use of LLMs.
Source: OECD questionnaire.

Integrity actors are leveraging a variety of LLMs to enhance their operations, most prominently models developed by companies like OpenAI, Google and Meta. Notable LLMs include OpenAI's Generative Pre-trained Transformer 4 (GPT-4), Google's Pathways Language Model (PaLM), and Meta's Open Pre-trained Transformer (OPT-175B), alongside other models like Google's BERT and Meta's LLM, Meta AI (LLaMA) (OECD, 2023[1]). These models offer foundational capabilities that can be specifically tailored to the unique requirements of integrity actors through techniques like Retrieval-Augmented Generation (RAG), which enriches LLMs with information from additional databases, including their own data sources. See Box 2.2 for further explanation of RAG.

Box 2.2. Retrieval-Augmented Generation for LLMs

Retrieval-Augmented Generation (RAG) is a technique developed to improve how large language models (LLMs), like the ones behind chatbots and virtual assistants, handle information. For different reasons, including reliance on old data, LLMs can provide incorrect answers and it can be difficult to understand how they derived at a particular response. RAG can help to address these challenges by allowing LLMs to access additional databases that can keep information current, which is particularly useful when applied to specialised domains or knowledge areas. For integrity actors, RAG can be an effective means for fencing-in their internal data sources, while improving the accuracy, relevancy and trustworthiness of a model's output.

RAG begins with identifying pertinent documentation and extracting vital text from it. Then, it breaks this text down into smaller parts and transforms these parts into a format (i.e. embeddings) that the model can understand and store efficiently. These pieces of information are kept in a special database (i.e. vector databases). When someone asks the model a question, it can look through this database to find up-to-date and accurate information to add to what it already knows before giving an answer.

For situations where it is critical for a model to provide facts that are current and accurate, such as when dealing with confidential information or needing to keep a clear record of data sources, the U.K.'s Generative AI Framework recommends using RAG. This approach can help to ensure that the model's answers are based on reliable data, making it particularly valuable for organisations focused on maintaining high levels of accuracy and accountability.

Source: (UK Government, 2024[16]; Gao et al., 2023[17])

On a technical level, integrity actors that responded to the questionnaire are either using an existing turnkey model—a model which is available in a ready to use form—without fine-tuning (7 out of 17 respondents), or they are fine-tuning a foundation model (7 out of 17 respondents). They primarily deploy GPT-4 and its predecessor, GPT-3.5, alongside BERT and LLaMA-2 for their advanced text processing needs. Several integrity actors employ platforms like ChatGPT in their generic form for broader tasks. However, the dependency on commercial LLMs poses challenges, particularly in data usage transparency and the risk of biases, as explored below (OECD, 2023[1]). Several integrity actors highlighted the use of RAG to fine-tune models. For instance, in the questionnaire responses, several SAIs highlighted the use of RAG for incorporating their own repositories of data and documents into the model, thereby further enhancing the customisation of the LLM.

A few respondents highlighted the use of LLMs tailored to the national context, such as in Norway and France, where integrity actors are making use of bespoke open-source tools. For instance, Box 2.3 provides an example of how an entity in the French government fine-tuned Llama to create a tool aimed at improving the efficiency and efficacy of parliamentary sessions by generating summaries of legislative proposals. This tool offers inspiration in a number of areas. Oversight bodies could benefit from an LLM that summarises complex legislative texts into concise versions to gain a quicker understanding of issues

that are relevant for audit engagements and decision making. In addition, ACAs could use a similar approach to create summaries to detect risks of undue influence in legislative proposals, such as clauses that might be overly beneficial to a specific group without sufficient justification.

Box 2.3. France's LLaMandement for summarising legislative text

Creating concise summaries is crucial in managing the legislative process, where tens of thousands of amendments, each spanning roughly two pages, are processed annually. These summaries are vital for a wide range of stakeholders—government officials, ministers, commission members, deputies, senators, administrative agents, journalists, and citizens—to quickly understand and discuss amendment contents without revisiting the full texts. AI-supported tools, especially those using LLMs, can play a significant role in this context. In contrast to other techniques, LLMs have the potential to efficiently distil vast amounts of complex legal texts into easily understandable information, enhancing efficient communication and informed decision making.

Recognising this opportunity, the Digital Transformation Delegation of the French Directorate General of Public Finances launched the "LLaMandement" project to automate the handling of legislative amendments. This project uses LLMs to assign amendments to the appropriate ministerial departments, search for past similar cases, and synthesise amendments into clear, ideally neutral summaries. The tool is designed to enhance the efficiency and accuracy of administrative work, supporting individuals to analyse bills and process amendments, especially during peak legislative periods.

LLaMandement draws on data from the Inter-ministerial Digital Management System for Legislative Amendments (SIGNALE), and it uses ministers' bench memoranda for training the model to ensure comprehensive understanding across different ministerial contexts. The developers of LLaMandement were sensitive to the possibility that the model would create biased results or promote misinformation. To address this concern, they used the Bias in Open-ended Language Generation Dataset (BOLD), a dataset used for evaluating biases in LLMs, particularly in open-ended text generation. Using BOLD, the developers assessed LLaMandement for biases related to gender, ethnicity and political ideology. They concluded the model reliably exhibited very few errors and the results were unbiased and neutral for different groups of people and beliefs.

Source: (Gesnouin et al., 2024[18])

2.2.2. Overcoming language barriers inherent in using or fine-tuning many off-the-shelf LLMs is a key challenge for integrity actors

One challenge that many organisations highlighted was the lack of existing LLMs trained in their native language. A 2023 study found that 38% of NLP models, which include LLMs, on the open-source platform Hugging Face are trained in English, followed by Spanish, German, and French (all at around 5%) (OECD, 2023[1]). Very few LLMs are trained in languages other than English. As illustrated in feedback from integrity actors who participated in OECD workshops and who responded to the questionnaire, integrity actors have had to invest extra time and energy into training their LLMs in their national language(s), usually by feeding the model regulations and reports written in native languages. This issue undermines the ability of many institutions to rely on existing LLMs with limited fine-tuning, as most LLMs are trained in English and a handful of other common languages (e.g. Spanish).

To enhance linguistic accuracy in local contexts, some countries are investing in the development of native-language LLMs that will be open source. Examples include the Netherlands' GPT-NL and Sweden's GPT-SW3, which are designed to excel in processing national languages by training on locally relevant texts.

These initiatives not only reduce dependency on technology companies but also offer improved performance in handling sensitive integrity-related data. Box 2.4 highlights the approach of the Office of the Comptroller General (*Controladoria Geral da União*, CGU) of Brazil to overcoming this and other challenges it faced while piloting its own LLMs.

Box 2.4. The Office of the Comptroller General (CGU) of Brazil's approach to piloting LLMs

CGU is an anti-corruption body within the public administration that is responsible both for financial management and transparency measures. It plans to use generative AI to support a variety of tasks, including inference of risks from internal audit reports, analysis of management response to internal audit recommendations, drafting of audit engagement findings, responding to support requests related to the asset and conflict-of-interest declaration system, and querying internal audit reports. The institution does not foresee generative AI replacing auditors but rather as a "co-pilot" that can help improve their efficiency. To this end, CGU's Data Intelligence Unit has invested in fine-tuning Llama-2 into their own LLM called Llama-2 GOV BR.

CGU encountered several challenges in its attempts to incorporate generative AI in its work. These included challenges related to inference time, scalability, costs, data sensitivity, and the content policy. CGU found that one way to overcome several of these challenges was by investing in a comparatively smaller LLM. Such models can achieve similar performance to larger models if trained well to do specific tasks with the added benefits that they can be served by local infrastructure, which reduces costs and inference time, improves scalability, keeps sensitive data on the organisation's premises, and allows for local management of the content policy.

When considering how best to deploy generative AI in their institution CGU also encountered the problem that existing LLMs did not perform well in Portuguese. However, since developing a new LLM from scratch is extremely expensive, it opted to fine-tune the Llama-2 model for its purposes. By pre-training the model with 10 million lines of high-quality Portuguese text from sources including audit reports, federal legislation, and PhD theses, CGU was able to reach a point where its LLM performed well enough to be used in its work. The CGU has plans to develop further monitoring and evaluation activities to ensure the LLM's reliability before rolling it out for day-to-day use.

Source: Meeting of OECD's Community of Practice on Technology and Analytics for Public Integrity: "Generative AI for promoting integrity and accountability in the public sector" (8 November 2023)

2.3. Ensuring the responsible development and use of generative AI and LLMs by integrity actors

2.3.1. Integrity actors recognised the need for safeguards, but more can be done to ensure the responsible and ethical use of AI as initiatives mature

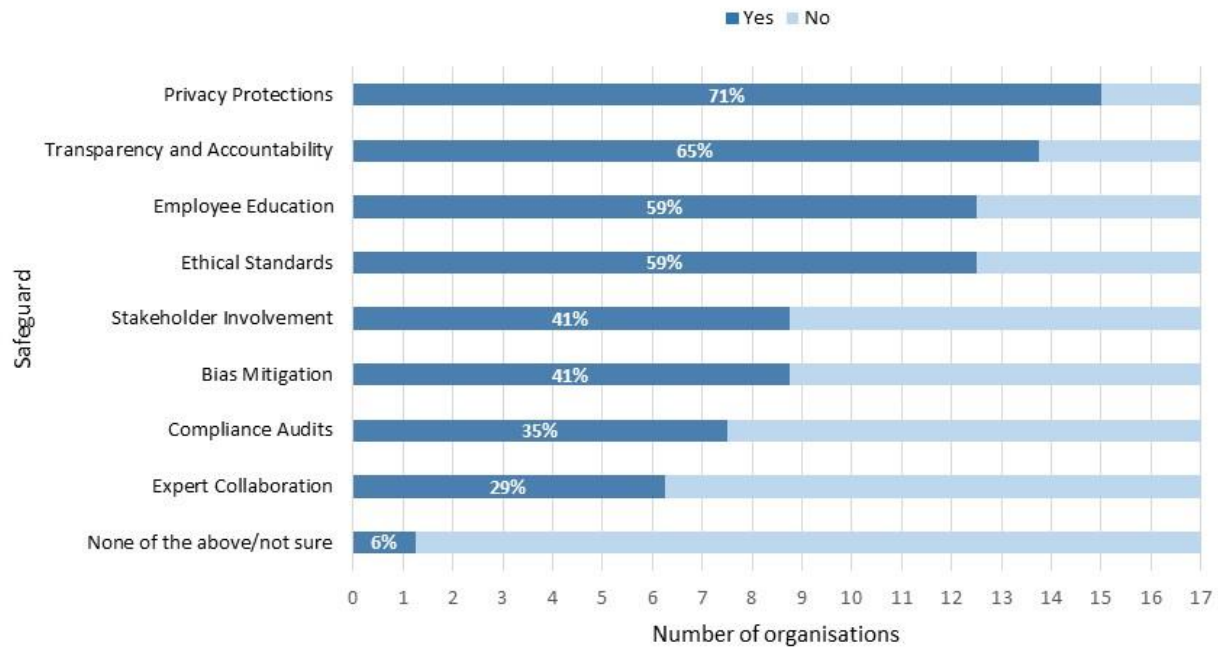
When asked to select from a range of challenges concerning the development and implementation of generative AI, including LLMs, and almost all integrity actors with relevant initiatives ranked issues surrounding compliance and ethics at the bottom of the list. This challenge involves navigating ethical, legal and privacy concerns, as well as regulatory compliance. Other challenges, such as technical development, resource management, and data management, ranked the highest (in that order) in terms of integrity actors' priorities for developing and implementing LLMs. This may reflect the current maturity of these integrity actors, most of which remain in the early stages of incubating ideas or ad hoc experimentation. Nonetheless, these challenges are important to consider during the design phase for the

reasons discussed below. Challenges will likely become more acute as public institutions solve more technical issues and are using LLMs more frequently and for more advanced tasks. Moreover, as regulations develop in this area it will also put pressure on these institutions to put measures in place to ensure compliance.

Integrity actors also offered insights about the measures their organisations employ to ensure the responsible use of AI as well as LLMs. Of the 17 organisations with LLM initiatives, the majority are employing the following measures: privacy protections, transparency and accountability measures, employee education and ethical standards. For instance, privacy protections include safeguards to adhere to data protection standards, and transparency and accountability broadly refers to measures to ensure open decision making about the use of AI, including redress mechanisms for citizens (OECD, 2022[19]) (see the note in Figure 2.5 for further explanation about safeguards).

Figure 2.5. Safeguards to ensure responsible use of AI and LLMs

What measures does your institution employ to ensure responsible AI and LLM usage?



Note: Possible responses included the following: 1) Ethical Standards: Implementation of ethical guidelines and policies; 2) Transparency and Accountability: Ensuring open AI decision making and maintaining accountability; 3) Bias Mitigation: Actively addressing biases to promote fairness; 4) Privacy Protections: Adhering to privacy and data protection standards; 5) Compliance Audits: Conducting regular ethical and legal compliance assessments; 6) Stakeholder Involvement: Engaging with stakeholders for input and addressing concerns; 7) Employee Education: Offering training and awareness programmes on responsible AI; 8) Expert Collaboration: Working with external experts for ethical and legal guidance; 9) None of the above/not sure; 10) Other. None of the respondents selected other.
Source: OECD questionnaire

Fewer institutions identified measures in place for bias mitigation, stakeholder involvement, compliance audits or expert collaboration as mechanisms to ensure responsible use of LLMs. The ranking of bias mitigation is notable. While it is unclear whether surveyed institutions do not see this as an issue or do not know what measures they should put in place, the issue of bias and hallucinations is a critical area for concern as LLMs become increasingly mainstreamed. Not only does this issue present policy, regulatory

and technical challenges, but it also poses political and reputational risks for those organisations that are experimenting with generative AI.

As integrity actors develop generative AI tools, including LLMs, they will need to contend with the issue of bias. Sampling bias is one form of bias that can be difficult to detect. This type of bias occurs when the data that underly a model are not actually representative of the population that they are meant to represent (Berryhill et al., 2019[20]). Sampling bias can be further broken down into historical bias related to pre-existing patterns in training data, representation bias arising from missing variables or an inadequate sample size, and measurement bias related the erroneous omission or inclusion of certain variables. For example, AI models designed to assign a corruption score to specific individuals based on previous conviction data could reflect biases related to higher wrongful conviction rates for racial minorities (Köbis, Starke and Rahwan, 2021[21]), thereby perpetuating the discrimination, marginalisation or exclusion of large segments of the population.

Similarly, when training algorithms, there is also the possibility of statistical bias. Statistical bias occurs when a model consistently makes the same error in prediction based on the expected outcome (Berryhill et al., 2019[20]). It is comparatively easy to detect. If a model consistently overestimates a value by the same amount, for example, the model simply requires more fine-tuning. This is fundamentally a problem with the model itself that those training it will need to resolve, and therefore may be less relevant for integrity actors that rely on LLMs of private companies and have less control over the design of the model.

Furthermore, if an LLM is trained disproportionately on texts produced by—or reflecting the experience of—certain categories of individuals, the LLM may eventually display more favourable views towards these categories of individuals or more unfavourable views towards other categories of individuals. Measures to mitigate this type of bias can include taking stock of training data for underrepresented groups, curation or semi-automatic curation of datasets to reach fairer results, as well as explainability and interpretability research and applying auditing processes (Lorenz, Perset and Berryhill, 2023[3]). In general, including more parameters when training a model reduces bias, but this can have other negative effects, such as increasing energy requirements or infringing more on personal privacy (OECD, 2023[1]), so integrity actors should carefully weigh these concerns when training LLMs. Another innovative approach is “red teaming” whereby researchers use one LLM to identify biases in another (Lorenz, Perset and Berryhill, 2023[3]).

For bias mitigation measures to be successful, there must first be a recognition of both the threat and consequences of biases. However, existing research on initiatives for AI as an anti-corruption tool uncovered a general lack of concern about bias mitigation, as well as a lack of accountability and transparency mechanisms to ensure the necessary bias mitigation was taking place (Odilla, 2023[22]). Examples that illustrate potential consequences of ignoring such issues can be found in different countries and sectors. For instance, the “Toeslagenaffaire” was a child benefits scandal in the Netherlands where the use of an algorithm resulted in tens of thousands of often-vulnerable families being wrongfully accused of fraud, as well as hundreds of children being separated from their families. This extreme case led to the collapse of the government. In Australia, in what became known as the “Robodebt scheme,” a data-matching algorithm calculated overpayments to welfare recipients that resulted in 470 000 incorrect debt notices and the sending of EUR 775 million in undue debt payments by welfare recipients, leading to a national scandal and a Royal Commission (OECD, 2023[23]).

While these examples are about AI in general, going beyond generative AI or LLMs, they illustrate the potential for severe political and social consequences that are relevant for integrity actors to consider as they embark on the use of generative AI and LLMs. It is critical for integrity actors to ensure they are taking the necessary steps to mitigate bias—including by ensuring compliance with national non-discrimination legislation—and sufficiently documenting how they have done so to establish trust in the tools that they have developed. For more advanced use cases, when appropriate, integrity actors can also promote redress mechanisms for citizens affected by algorithm-driven decisions (OECD, 2022[19]). Finally, maintaining a focus on a human-in-the-loop system, whereby trained humans play a central role in the

development of models and creating a continuous feedback loop, can help to mitigate the risk of machine biases that manifest into harmful decisions and actions.

The use of AI in general, whether generative AI or other forms of AI, to promote integrity and combat corruption poses a unique set of ethical concerns. Integrity actors already process large amounts of personal data, such as in mandatory interest or asset declarations. This means they must be careful that any use of generative AI to process this data protects individuals' privacy and that the entity does not disseminate data that would not otherwise be publicly available. For more detail on how the Corruption Prevention Commission (*ՀՀ Կոռուպցիայի կանխարգելման հանձնաժողով*, CPC) of Armenia has worked to address ethical concerns in this area see Box 2.5. Ethical issues can also arise when working with crowdsourced data, such as whistleblower complaints. AI models may have difficulty distinguishing founded complaints from unfounded ones, which could lead to wrongful denunciation of public officials or wrongful decisions on cases concerning citizens. Research has shown that individuals are generally against algorithms making ethical decisions (Köbis, Starke and Rahwan, 2021[21]), which therefore requires that AI and generative AI models utilised in integrity bodies still have some level of human oversight.

Box 2.5. The Corruption Prevention Commission (CPC) of Armenia's use of AI to verify asset declarations

Armenia established the CPC in 2019 as part of a wider package of anti-corruption reforms, and upon its creation it assumed the responsibility for overseeing the electronic register of asset declarations. However, given the large number of officials required to submit these declarations and the CPC's limited resources, it was difficult to perform any meaningful checks of the content of these declarations. Initially, the electronic submissions were not even machine readable. The CPC therefore decided to build a data platform that would provide the necessary structure for data analysis and link to the databases of other state bodies. The CPC recently introduced an automated verification system that conducts an initial screening of asset declarations and identifies red flags. It is currently piloting the introduction of an AI component that would enable this system to learn from this process and identify new patterns in corrupt behaviour.

CPC noted that stakeholder engagement played a key role in developing a tool that would handle this sensitive data properly. Consultations with private sector actors both domestically and internationally helped the CPC gain a better understanding of the technical infrastructure that would be necessary for this tool to function and effectively and responsibly. Ultimately, these consultations led the CPC to invest in improving the quality of the underlying data first before experimenting with machine learning algorithms.

In other areas, different ethical considerations conflicted with each other, making finding a solution more difficult. For example, the desire to promote transparency by publishing the algorithm used to verify the asset declarations conflicted with the need to respect the privacy of those declaring, particularly given that Armenia has aligned its legal framework with the GDPR. In the end, the CPC determined that while publishing the asset declarations themselves was in the public interest and therefore permissible under the GDPR, publishing the algorithm was not.

It nonetheless remains important to be transparent about how the system is flagging declarations to maintain public trust, and the need to balance transparency and privacy will persist as development of this tool continues. In neighbouring Georgia, a lack of transparency about how the government is using AI has been undermining trust in the public institutions using these tools.

Source: (Izdebski, Turashvili and Harutyunyan, 2023[24])

2.3.2. Integrity actors can put a greater emphasis on monitoring and evaluating their AI activities, including consideration of model interpretability

High-level principles, standards and national normative frameworks offer a starting point for integrity actors to ensure they a prioritising the responsible use of AI as their initiatives mature. The OECD Recommendation on Artificial Intelligence (OECD, 2023[25]) identifies five value-based principles for the responsible use of AI (see Box 2.6). Many countries have adopted similar principles within their national frameworks for regulating AI. For example, Switzerland's Guidelines on Artificial Intelligence for the Confederation mirror the OECD Principles and add principles on regulatory compliance, stakeholder engagement, and actively shaping global AI governance (Federal Council of Switzerland, 2020[26]). They also contain more detail on complying with specific legal principles. The same is true of the Government-Wide Vision on Generative AI of the Netherlands, which outlines six specific areas of action to support the principles (Government of the Netherlands, 2024[6]). The Netherlands has also taken the approach of establishing a Government AI Validation Team to review pilot projects and ensure compliance with the principles, which can help mitigate risks related to irresponsible use. In Denmark, the Agency for Digitalisation (*Digitaliseringsstyrelsen*) has issued targeted guidelines for managers in public authorities to ensure responsible use of generative AI in their institutions (Danish Agency for Digitalisation, 2024[27]). Such ethical frameworks for AI generally also play an important role in mitigating risks related to generative AI specifically (Lorenz, Perset and Berryhill, 2023[3]). Integrity actors can consider these broad responsible use issues when developing generative AI tools and put the necessary safeguards in place to ensure responsible use.

Box 2.6. The OECD Principles on Artificial Intelligence

The OECD Principles on Artificial Intelligence, which are laid out in the OECD Council Recommendation on Artificial Intelligence, are divided into values-based principles and recommendations for policymakers. The five value-based principles that aim to encourage responsible use of AI in line with key values of OECD member states are as follows:

- Inclusive growth, sustainable development and well-being. Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.
- Human-centred values and fairness. AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights. To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.
- Transparency and explainability. AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:
  - to foster a general understanding of AI systems,
  - to make stakeholders aware of their interactions with AI systems, including in the workplace,
  - to enable those affected by an AI system to understand the outcome, and,

    - to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.
- Robustness, security and safety. AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk. To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art. AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.
- Accountability. AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.

When exploring ways of incorporating generative AI into their work, integrity actors should therefore make sure they are adhering to these principles. Given the sensitive nature of the data that these organisations hold and process, respecting human rights, transparency, and security are particularly important. The OECD Council Recommendation on Artificial Intelligence is under revision and is expected for adoption at the Ministerial Council Meeting in May 2024.

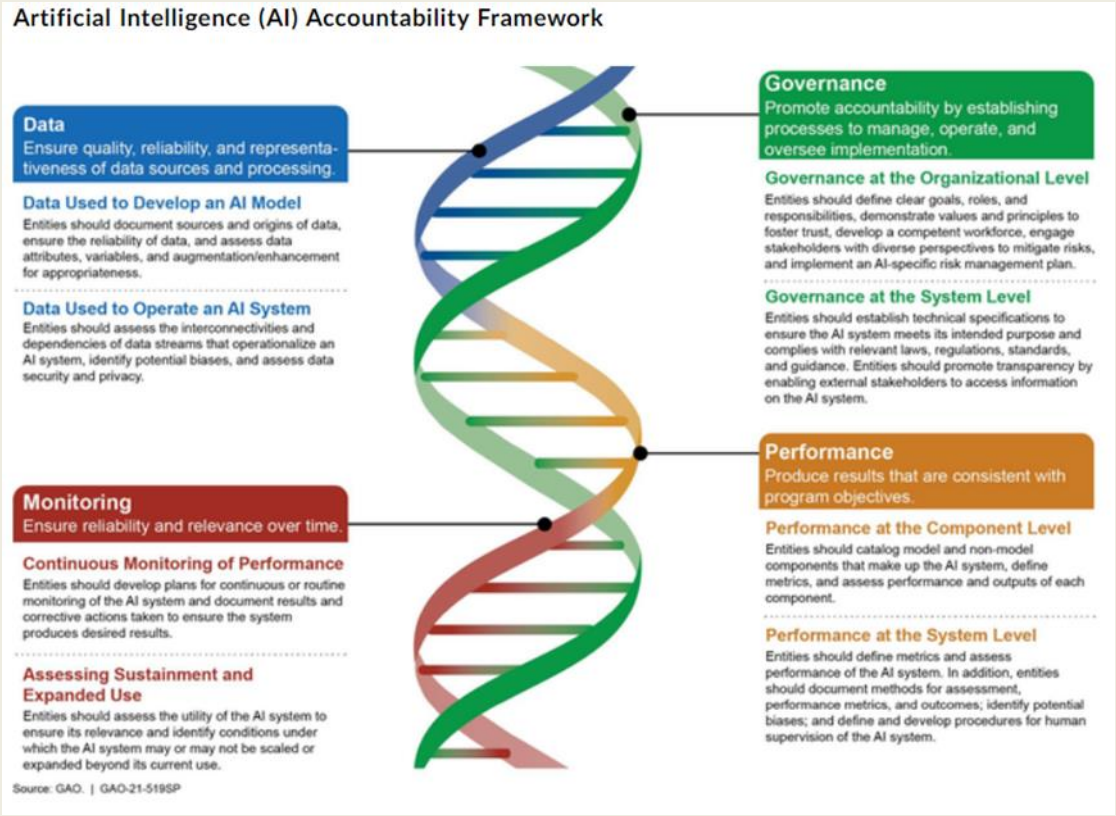
Source: (OECD, 2023[25]) (OECD, 2019[28])

Monitoring and evaluation play an important role in ensuring that AI use is indeed responsible. Monitoring during the implementation stage is necessary to ensure that risks are being mitigated and unintended consequences are identified. Public institutions can also take a risk-based approach to monitoring that involves higher scrutiny for processes with the potential for more severe negative consequences (Berryhill et al., 2019[20]), such as those related to anti-corruption. Box 2.7 provides more details on good practices for monitoring and evaluation within the US Government Accountability Office (GAO)'s AI Accountability Framework. Additional examples of how integrity actors, including SAIs and other oversight bodies, are promoting algorithmic can be seen around the world. This includes the development of a *General Instruction on Algorithmic Transparency* for public entities developed by the Chilean Transparency Council as well as a cross-border collaboration between the SAIs of Finland, Germany, the Netherlands Norway and the UK to develop a white paper on auditing machine learning algorithms (OECD, 2023[23]).

Box 2.7. The AI Accountability Framework of the US Government Accountability Office (GAO)

In order to ensure effective use of AI, GAO has developed a framework to evaluate the performance of AI systems to make sure they deliver value and remain fit for purpose over time. Pillars 3 and 4 of the framework on performance and monitoring, respectively, provide examples of best practices for monitoring and evaluation of data-driven tools (see Figure 2.6 below). These procedures for monitoring and evaluation are not only useful in guiding entities in their use of AI, but they could also be applied to other data-driven tools and systems, including those that do not employ AI.

Figure 2.6. GAO's Artificial Intelligence Accountability Framework



Regarding performance, at both the component level and the system level, AI models should be documented and assessed against predetermined performance metrics that are precise, consistent, and reproducible. Documentation should aim to address the following groups of questions:

- How are components and models solving defined problems? What is their intended use?
- How are the specifications and parameters are selected, evaluated, and optimised?
- How suitable are components and models to available data and operating conditions?
- How are components and models tested and what are the results?
- What ethical considerations exist? What biases and unintended consequences have been identified?
- What degree of human supervision is required and how was this determined?

When it comes to model evaluation, the selected performance metrics should be accurate and useful, and the justification for their selection and the person(s) responsible for their development should also be documented.

Once AI systems are put in place, a plan for continuous or routine monitoring should be developed. This helps ensure that AI models remain reliable and relevant. The plan should define acceptable levels of data and model drift that are based on a risk assessment and require documentation of monitoring activities and any corrective actions taken. As part of the monitoring of AI systems, their continued utility and any potential opportunities for scaling should also be assessed. Any decisions to retire or scale models or systems should be based on predefined performance metrics, and any updates that take place and their impact should be documented. Finally, throughout the process of monitoring and evaluation, it is important that entities keep in mind AI systems' consistency with their objectives and values in order to foster and maintain public trust.

Source: (US Government Accountability Office, 2021[29])

From a methodological perspective, one of the main challenges integrity actors face, like other organisations, relates to the limitations of LLMs in terms of interpretability, explainability and transparency. The breadth and variety of data that feed into LLMs, which are fundamental to their usefulness, present major challenges in tracing the connection between outputs and inputs. The complexity of the underlying architecture and decision-making mechanisms exacerbate this challenge (Shabsigh and Boukherouaa, 2023[2]) and can make it more difficult for citizens to understand how their government is making decisions or make appeals to protect their own rights and interests. In the integrity context, overcoming this challenge can be the difference between limited use (e.g. LLMs for summarising text) and more extensive integration of LLMs across core audit or investigative processes. For instance, preserving an audit or investigative trail, or providing justification for prioritising risks, are core tenets of the work of anti-corruption and oversight bodies alike. Their legal obligations and reputations rely on understanding the provenance of data that informs key decisions.

Challenges concerning interpretability, explainability and transparency of LLMs further highlight the importance of auditors, investigators and analysts maintaining professional scepticism and ensuring human-centred checks remain throughout the training and deployment of LLMs. There are no easy solutions to address this challenge. Government agencies have explored the use of decision trees to help illustrate the link between the results from AI systems and an explanation of how they came about (Berryhill et al., 2019[20]), and they have issued explainable AI toolkits to help assist in this area.[6] Academia offers additional ideas and insights. For instance, a group of researchers introduced a taxonomy of explainability techniques for LLMs, as well as metrics for evaluating generated explanations to improve model performance (Zhao, 2023[30]). Other research explores transparency within the unique context of LLMs and poses priorities and questions that can be helpful for integrity actors as they assess their own LLMs (see Box 2.8).

Box 2.8. Human-centred considerations for promoting transparency when evaluating LLMs

Integrity actors in government have a wide range of internal and external stakeholders to account for when considering the interpretability, explainability and transparency of the LLMs they develop. These stakeholders have different needs in terms of the what, when, and how of an AI initiative, given their different roles, responsibilities and levels of technical expertise in what researchers call the LLM ecosystem. Transparency in this context implies that relevant stakeholders can "form an appropriate understanding of a model or system's capabilities, limitations, how it works, and how to use or control its outputs."

There are several key areas and questions for consideration that are broadly applicable to any organisation that is developing LLMs, but they are especially useful for integrity actors in government given their responsibilities to the general public and other stakeholders. These areas and questions draw inspiration from the machine learning and human-computer interaction literature: and can provide integrity actors with a starting point for thinking about their own measures to promote transparency in the context of LLMs:

1. **Model reporting**
   - What information is needed to characterise the functional behaviour of an LLM?
   - What do different (and new) types of stakeholders need from model reporting frameworks?
   - What is needed beyond static documentation?
2. **Publishing evaluation results**
   - Who is the evaluation targeted at and for what purpose?
   - At what level should the evaluation take place?
   - How should LLM limitations and risks be evaluated?
3. **Providing explanations**
   - How can the organisation provide faithful explanations for the LLM, knowing that it is the ultimate black box?
   - What explanations are appropriate for LLM-infused applications?
4. **Communicating uncertainty**
   - What is a useful notion of uncertainty for LLMs?
   - What are the most effective ways to communicate uncertainty?

These questions are meant to guide future research, but they can also provide integrity actors with a starting point for thinking about their own measures to promote transparency in the context of LLMs, as well as ways to strengthen evaluation mechanisms with an approach that is tailored to the unique LLM context.

Source: (Liao and Vaughan, 2023[31])

2.4. Mitigating the risk of generative AI as a tool to undermine integrity

2.4.1. Generative AI can enhance the work of integrity actors, but it also creates the need for greater vigilance of evolving integrity risks

Generative AI poses unique risks to the work of anti-corruption and integrity bodies specifically. For instance, one category of risks is adversarial attacks. Broadly, adversarial attacks represent a cybersecurity vulnerability where attackers design inputs to evade detection. Generative AI can be used to create advanced phishing communications or enable actors with malicious intent to convincingly mimic individuals or entities, thus heightening the risk of identity theft, fraud and social engineering. Moreover, the spread of deepfakes (highly realistic videos, audios, or images) amplifies this threat (Shabsigh and Boukherouaa, 2023[2]).

Other risks range from LLMs making it easier for public officials to commit fraud to making it more difficult for integrity actors to detect corruption (Independent Commission Against Corruption, 2023[32]). The politically sensitive nature of anti-corruption work also means that leaning too heavily on automated decision-making can lead to the undermining public trust or augmenting ethical concerns. As it incorporates AI broadly and generative AI specifically into its work, the Independent Commission Against Corruption (ICAC) of New South Wales in Australia has considered how these and other threats could manifest. For more detail see Box 2.9.

Box 2.9. Insights from the Independent Commission Against Corruption (ICAC) of New South Wales on AI's potential threats to anti-corruption work

ICAC is an anti-corruption agency (ACA) in the Australian federal state of New South Wales that is responsible for a number of anti-corruption activities, including both the promotion of corruption prevention measures and investigation of corruption allegations. In response to a legislative inquiry on the use of AI in New South Wales in 2023, ICAC produced a report outlining both the opportunities and threats that AI poses to its work. Potential opportunities included the ability of AI to enhance intelligence through filtering, sorting and analysing large data sets; pattern recognition; forecasting and modelling; sentiment analysis; detection of anomalies in data; and data integration and multi-source analysis. ICAC also noted the ability of AI to reduce opportunities for corruption by limiting the degree of human discretion in decision making.

However, ICAC also noted that AI risks frustrating its anti-corruption efforts in several ways. These include:

1. The ability of AI to produce deepfakes
2. AI enhanced cybercrime
3. Exploitation of AI by public officials
4. Deference to AI
5. The use of AI to forge government documents
6. Threats to democracy and public discourse
7. Risks related to outsourcing

Points 3 and 5 are particularly noteworthy for the work of integrity actors in government, especially given the role of LLMs. First, ICAC notes that it has already investigated public officials who have tampered with IT systems to cover up corrupt conduct. LLMs could exacerbate this problem. An individual with enough technical expertise could poison data or manipulate models in order to alter system outputs.

They could also take advantage of known system vulnerabilities for personal gain or even sell information on these vulnerabilities to other corrupt actors.

ICAC also notes that many of its investigations relate to fraudulent documents, including procurement information, recruitment information, grant applications, building certificates, applications for business licenses, and conflict-of-interest declarations. While forging or altering these documents is difficult for humans to do in a convincing manner, it would be relatively easy for generative AI. There is also the risk that generative AI produces fraudulent documents without being prompted, resulting in fraud investigations against individuals with no malintent. In cases where fraud was premeditated, it is nonetheless easy for individuals to leverage the "black box" nature of advanced technology to feign ignorance.

These are just some of the threats that generative AI may pose to anti-corruption and integrity work. More broadly, ICAC notes that AI systems can reduce public trust in government decision making or may detach decision makers from those affected by their decisions to such an extent that they no longer consider moral ramifications. It is important that integrity actors keep these threats in mind as generative AI becomes more widely used and understood.

Source: (Independent Commission Against Corruption, 2023[32])

Going beyond LLMs, the OECD has previously raised issues concerning the risk that generative AI can amplify misinformation (i.e. the unintended spread of false information) and the deliberate spread of disinformation by malicious actors (Lorenz, Perset and Berryhill, 2023[3]). For instance, the widespread dissemination of false information during the COVID-19 pandemic highlighted the severe consequences disinformation can have on the execution of policies, as well as on trust and unity within society (Matasick, Alfonsi and Bellantoni, 2020[33]).

Generative AI also poses risks in the context of lobbying. As illustrated by previous examples, generative AI, particularly LLMs, can help entities to quickly process and provide inputs on draft legislation. Yet, generative AI as a form of AI also has the potential to completely overwhelm government consultation platforms by spamming them with fake or repetitive comments in order to amplify certain processes or stymy the policymaking process altogether (Smith and Harris, 2023[34]). The solutions are beyond the mandate of many integrity actors, but this issue is worth bearing in mind as they advance with relevant AI initiatives. Countries have done little to amend lobbying legislation to account for this threat, and opportunities remain to adjust the scope of lobbying legislation and the information that lobbyists are required to disclose in light of it. Countries can also invest in the necessary IT tools to manage increased comment volume and distinguish between legitimate and AI-generated comments.

While LLMs can amplify or create new risks that have implications for the work of integrity actors and their external environment, other risks are more internal in nature and can be considered when developing control and risk mitigation measures. Bias in AI can occur unintentionally, yet deliberate attempts to undermine or exploit LLMs for nefarious purposes present an evolving challenge. The OECD has reported previously on these issues in different contexts. For instance, based on the framework of the Surveillance Commission of the Financial Sector (*Commission de Surveillance du Secteur Financier*, CSSF) in Luxembourg, the OECD highlighted risk related to data poisoning and model theft, among others (Berryhill et al., 2019[20]).

- Data poisoning. This includes tampering with the training data, leading the AI to learn incorrect patterns. This is particularly problematic for types of AI that rely on continuously updated online data sources. For instance, individuals might create misleading content on social media to disrupt an AI's ability to accurately perform sentiment analysis. Similarly, subtle alterations to images, indiscernible to humans, can trick an AI into misidentifying new images.

- Model theft. This refers to the risk of unauthorised replication of a model, whereby attackers reverse-engineer an LLM or breach security measures to access proprietary or sensitive information. Examples include the hijacking of AI-powered chatbots for public services to create misleading or fraudulent services, or stealing of models used for predictive policing in order to circumvent law enforcement strategies.

While this paper concentrates on how integrity actors use generative AI, particularly LLMs in their operations, it is crucial to acknowledge that these technological advancements must be accompanied by a deeper awareness of the potential integrity risks they pose. The very risk assessments that can be enhanced by generative AI may, in a variety of government spheres, need to account for how the same technology can exacerbate integrity risks.

References

AI Sweden (2024), *A common digital assistant for the public sector*. [10]

Berryhill, J. et al. (2019), "Hello, World: Artificial intelligence and its use in the public sector", *OECD Working Papers on Public Governance*, No. 36, OECD Publishing, Paris, https://doi.org/10.1787/726fd39d-en. [20]

Bumann, J. and M. Peter (2019), *Action Fields of Digital Transformation - A Review and Comparative Analysis of Digital Transformation Maturity Models and Frameworks*, Edition Gesowip, https://www.researchgate.net/publication/337167323_Action_Fields_of_Digital_Transformation_-_A_Review_and_Comparative_Analysis_of_Digital_Transformation_Maturity_Models_and_Frameworks. [13]

Danish Agency for Digitalisation (2024), *Guide for public authorities on the responsible use of generative artificial intelligence*, Danish Agency for Digitalisation, Copenhagen. [27]

Emett, S. (2023), *Leveraging ChatGPT for Enhancing the Internal Audit Process – A Real-World Example from a Large Multinational Company*, https://ssrn.com/abstract=4514238 or http://dx.doi.org/10.2139/ssrn.4514238. [12]

European Parliament (2024), "Artificial Intelligence Act: MEPs adopt landmark law", *European Parliament News*, Press Release on 13 March 2024, https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law (accessed on 18 March 2024). [15]

Federal Council of Switzerland (2020), *Guidelines on Artificial Intelligence for the Confederation*, Federal Council of Switzerland, Bern. [26]

Gao, Y. et al. (2023), "Retrieval-augmented generation for large language models: A survey.", *arXiv preprint*, https://arxiv.org/abs/2312.10997. [17]

Gesnouin, J. et al. (2024), *LLaMandement: Large Language Models for Summarization of French Legislative Proposals*, https://arxiv.org/abs/2401.16182. [18]

Government of the Netherlands (2024), *The government-wide vision on Generative AI of the Netherlands*, Government of the Netherlands, The Hague. [6]

Huang, A. and H. Yi Yang (2023), "FinBERT: A Large Language Model for Extracting Information from Financial Text", *Contemporary Accounting Research*, Vol. 40/2, https://doi.org/10.1111/1911-3846.12832. [11]

Independent Commission Against Corruption (2023), *SUBMISSION TO THE LEGISLATIVE COUNCIL INQUIRY INTO ARTIFICIAL INTELLIGENCE IN NEW SOUTH WALES*, Independent Commission Against Corruption (ICAC), Sydney. [32]

Izdebski, K., T. Turashvili and H. Harutyunyan (2023), *The Digitalization of Democracy: How Technology is Changing Government Accountability*, National Endowment for Democracy, Washington. [24]

Köbis, N., C. Starke and I. Rahwan (2021), *Artificial Intelligence as an Anti-Corruption Tool (AI-ACT): Potentials and Pitfalls for Top-down and Bottom-up Approaches*, Max-Planck-Institute for Human Development, Center for Humans and Machines. [21]

Liao, V. and J. Vaughan (2023), *AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap*, https://arxiv.org/pdf/2306.01941.pdf. [31]

Lorenz, P., K. Perset and J. Berryhill (2023), "Initial policy considerations for generative artificial intelligence"*, OECD Artificial Intelligence Papers*, No. 1, OECD Publishing, Paris, https://doi.org/10.1787/fae2d1e6-en. [3]

Matasick, C., C. Alfonsi and A. Bellantoni (2020), "Governance responses to disinformation: How open government principles can inform policy options"*, OECD Working Papers on Public Governance*, No. 39, OECD Publishing, Paris, https://doi.org/10.1787/d6237c85-en. [33]

Odilla, F. (2023), "Bots against corruption: Exploring the benefits and limitations of AI-based anti-corruption technology", *Crime, Law and Social Change*, Vol. 80/4, pp. 353-396, https://doi.org/10.1007/s10611-023-10091-0. [22]

OECD (2023), "AI language models: Technological, socio-economic and policy considerations"*, OECD Digital Economy Papers*, No. 352, OECD Publishing, Paris, https://doi.org/10.1787/13d38f92-en. [1]

OECD (2023), *Global Trends in Government Innovation 2023*, OECD Public Governance Reviews, OECD Publishing, Paris, https://doi.org/10.1787/0655b570-en. [23]

OECD (2023), "Recommendation of the Council on Artificial Intelligence"*, OECD Legal Instruments*, OECD/LEGAL/0449, OECD, Paris, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. [25]

OECD (2022), "Facilitating citizen and stakeholder participation through the protection of civic freedoms", in *The Protection and Promotion of Civic Space: Strengthening Alignment with International Standards and Guidance*, OECD Publishing, Paris, https://doi.org/10.1787/9ca8987d-en. [19]

OECD (2022), *Strengthening Analytics in Mexico's Supreme Audit Institution: Considerations and Priorities for Assessing Integrity Risks*, OECD Public Governance Reviews, OECD Publishing, Paris, https://doi.org/10.1787/d4f685b7-en. [4]

OECD (2021), *Countering Public Grant Fraud in Spain: Machine Learning for Assessing Risks and Targeting Control Activities*, OECD Public Governance Reviews, OECD Publishing, Paris, https://doi.org/10.1787/0ea22484-en. [8]

OECD (2020), *Good Practice Principles for Data Ethics in the Public Sector*, OECD, Paris, https://www.oecd.org/gov/digital-government/good-practice-principles-for-data-ethics-in-the-public-sector.pdf. [14]

OECD (2020), "The OECD Digital Government Policy Framework: Six dimensions of a Digital Government", *OECD Public Governance Policy Papers*, No. 02, OECD Publishing, Paris, https://doi.org/10.1787/f64fed2a-en. [35]

OECD (2019), "Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449", *OECD Legal Instruments*. [28]

OECD (2014), "Recommendation of the Council on Digital Government Strategies", *OECD Legal Instruments*, OECD/LEGAL/0406, OECD, Paris, https://www.oecd.org/gov/digital-government/Recommendation-digital-government-strategies.pdf. [37]

Office of the Auditor General of Norway (2018), *Auditing to benefit the society of tomorrow: Strategic plan 2018–2024*, Office of the Auditor General of Norway, Oslo. [5]

Otia, J. and E. Bracci (2022), "Digital transformation and the public sector auditing: The SAI's perspective", *Financial Accountability & Management*, Vol. 38/2, pp. 252-280, https://doi.org/10.1111/faam.12317. [36]

Shabsigh, G. and E. Boukherouaa (2023), "Generative Artificial Intelligence in Finance: Risk Considerations", *Fintech Notes*, Vol. 2023/006, https://doi.org/10.5089/9798400251092.063. [2]

Smith, A. and M. Harris (2023), *How artificial intelligence and large language models may impact transparency*, Westminster Foundation for Democracy. [34]

U.S. Government Accountability Office (2024), *Artificial Intelligence Use Cases*, https://www.gao.gov/science-technology/artificial-intelligence-use-cases. [9]

UK Government (2024), *Guidance: Generative AI Framework for His Majesty's Government*, https://www.gov.uk/government/publications/generative-ai-framework-for-hmg/generative-ai-framework-for-hmg-html. [16]

US Government Accountability Office (2021), *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities*, US Government Accountability Office, Washington. [29]

World Bank (2023), *The Governance Risk Assessment System(GRAS) Advanced Data Analytics for Detecting Fraud, Corruption, and Collusion in Public Expenditures*, https://openknowledge.worldbank.org/handle/10986/40640. [7]

Zhao, H. (2023), "Explainability for Large Language Models: A Survey", *ACM Transactions on Intelligent Systems and Technology*, https://doi.org/10.1145/3639372. [30]

Notes

[1] This designation is without prejudice to positions on status and is in line with United Nations Security Council Resolution 1244/99 and the Advisory Opinion of the International Court of Justice on Kosovo's declaration of independence.

[2] We estimate that over 150 organisations received the questionnaire, but we did not attempt to track the total number of recipients given the qualitative purpose of our research. The aim of our questionnaire was to collect insights and use cases from a targeted group of government entities (i.e. integrity actors) rather than achieving statistical representativeness. Without additional information as to why some recipients decided not to complete the questionnaire, reporting on the total number of recipients does not materially contribute to the qualitative nature of our findings.

[3] LangChain is a framework designed to build applications powered by language models, facilitating the creation of context-aware applications that connect to various sources for context—such as prompt instructions, examples, and content grounding—and utilise language models for reasoning, including determining responses based on context and deciding on actions to take. See https://www.langchain.com/.

[4] For instance, see: OECD (OECD, 2023[25]), "Recommendation of the Council on Artificial Intelligence" https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449; OECD (2014[37]), "Recommendation of the Council on Digital Government Strategies", https://www.oecd.org/gov/digital-government/Recommendation-digital-government-strategies.pdf; OECD (2020[35]), The OECD Digital Government Policy Framework: Six dimensions of a Digital Government; OECD (2022[4]), *Strengthening Analytics in Mexico's Supreme Audit Institution: Considerations and Priorities for Assessing Integrity Risks*; OECD (2021[8]), *Countering Public Grant Fraud in Spain: Machine Learning for Assessing Risks and Targeting Control Activities*; Otia and Bracci (2022[36]), Digital transformation and the public sector auditing: The SAI's perspective; and Bumann and Peter (2019[13]), *Action Fields of Digital Transformation - A Review and Comparative Analysis of Digital Transformation Maturity Models and Frameworks*.

[5] Software, Tools, and Compliance, as a response option, covered essential software, development tools, and adherence to legal/regulatory standards. System Scalability and Integration pertains to the ability to scale IT resources and integrate the LLM with existing technology stacks.

[6] See, for instance, XAITK, an open-source explainable AI toolkit built with the support of the Defense Advanced Research Projects Agency (https://xaitk.org/ and https://www.darpa.mil/program/explainable-artificial-intelligence).

GOVERNANCE

GOVERNANCE

EQUITABLE GROWTH, FINANCE & INSTITUTIONS INSIGHT

Artificial Intelligence in the Public Sector

Maximizing Opportunities, Managing Risks

Supported by the GovTech Global Partnership: www.worldbank.org/govtech

GovTech Putting people first

Federal Ministry Republic of Austria Finance

Ministry of Economy and Finance Republic of Korea

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Economic Affairs SECO

WORLD BANK GROUP

1818 H Street NW, Washington DC 20433
Telephone: 202-473-1000; Internet: www.worldbank.org

Cover design and layout: Diego Catto / www.diegocatto.com

Contents

Boxes

Figures

Tables

>>>

Foreword

Disruptive technologies like artificial intelligence (AI), mobile apps, Internet of Things, blockchain, cloud computing, and data analytics have the potential to transform governments by enhancing personalized service delivery experience, improve back-end process efficiencies, and strengthening policy compliance. One of the most promising disruptive technologies, AI is already being adopted by the digitally advanced governments to maximize its potential benefits. And this trend is catching up with other governments as well. More than 50 governments have issued or are in the process of issuing AI strategies in recent years.

However, in many of our client countries, the public sector's ability to adopt AI is hampered by low access to digital skills, insufficient foundational digital technologies, and inadequate digital data as well as a lack of awareness of the potential of AI. These differences in the pace of AI adoption in the public sector could further exacerbate inequalities between the rich and the poor countries. To promote wider AI adoption in our client governments, this paper provides a preliminary synthesis of the existing opportunities, risks, and building blocks required for implementing and integrating AI in their operations. The paper also highlights policy, governance and people aspects necessary for AI implementation, as there are no shortcuts to technology adoption. The use of technology cannot be fast-tracked as many of the analog complements needed for adoption are not yet in place (World Bank 2016).

To better understand the role AI can play in public sector transformation, the World Bank produced this paper in partnership with the Swiss State Secretariat for Economic Affairs. This paper aims to distill the existing knowledge on the use of AI in the public sector and summarize the lessons learned from early adopters. It draws on the accumulated literature, case studies, and emerging trends to provide guidance to our teams working in this field. The World Bank's technical team benefited from a panel of experts from inside the World Bank and from the industry who shared their insights and enriched the paper. The goal is to alert our staff and clients to the opportunities, risks, and the potential to foster AI for public sector transformation.

Edward Olowo-Okere
Global Director
Governance and Institutions

>>>

Acknowledgments

This paper was prepared by a World Bank team consisting of Khuram Farooq (Senior Governance Specialist in the Global Governance Practice (GGP) and Bartosz Sołowiej (Consultant). The team received valuable guidance from Edward Olowo-Okere (Global Director, GGP) who is leading the GovTech agenda in the Bank; Tracey Marie Lane and Adenike Sherifat Oyeyiola (Practice Managers, GGP); Kimberly Johns (Senior Governance Specialist); and Cem Dener (Lead Governance Specialist, GGP).

The team benefited from the comments of external peer reviewers: Aaron Moffatt (Chief Technology Officer, Immersion Analytics) and Brittan Heller (Global Head of AI, Foley Hoag LLP, Harvard Tech, and Human Rights AI Fellow 2019). The team is also grateful for contributions from the World Bank's reviewers: Aki Ilari Enkenberg (Senior Digital Development Specialist, IDD02), David Santos (Senior Public Sector Specialist, ELCG2), Jana Kunicova (Senior Public Sector Specialist, EA1G2), Parminder P.S. Brar (Lead Governance Specialist, GGP), and Trevor Monroe (Senior Operations Officer, DECAT).

The team also wishes to express its thanks to Barbara Joan Rice (Consultant, World Bank) and Mary A. Kent (Working Copy Editor) for their editorial support; to Jasmine N. Brown (Intern, Foley Hoag LLP) for her research support, and Angela Hawkins (Team Assistant) for her formatting expertise. Finally, our thanks to Richard Crabbe for his editorial work and communications advice.

This paper was financed by the State Secretariat for Economic Affairs of Switzerland (SECO). We gratefully acknowledge this excellent partnership with SECO to promote the GovTech agenda.

>>>

Executive Summary

Disruptive technologies like Artificial Intelligence (AI) offer new opportunities to governments facing development challenges, especially now as fiscal stress is causing many governments to find new solutions to improving services without increasing the costs. Artificial Intelligence can be defined as the ability of the software systems to carry out tasks that usually require human intelligence: vision, speech, language, knowledge, and search.

Many governments view AI as a strategic resource for competitiveness and growth and are embracing it with speed and priority. According to Bughin et al. (2018), AI can potentially contribute $13 trillion to the global economy by 2030.[1] At least 50 governments have developed or are in the process of developing an AI strategy. However, the pace of AI adoption is uneven, and most countries are not ready for AI adoption. There is no country from Africa or Latin America in the list of the top 20 countries on the AI Readiness Index developed by Oxford Insights.[2] Except for four economies, Asia-Pacific is also one of the worst-performing regions on this list. Slower adoption of AI in our client countries may lead to further inequality between the rich and the poor nations

To reduce these inequalities, opportunities should be explored through the initiation of AI projects in areas of strategic impact and priority. Chapter 2 on AI opportunities provides examples of AI use from around the world. Moreover, it provides operational guidance on AI implementation on fundamental questions relating to developing country contexts. It broadens the perspective to explore opportunities for implementation. Government AI deployments exist in every sector. A common pattern of use cases includes citizen engagement, compliance and risk management, fraud and anti-corruption, business process automation, service delivery, asset management, and analytics for decision-making and policy design.

While AI should be explored to solve complex problems, associated adverse consequences in client contexts should also be fully understood and managed as AI comes with additional risks that could exacerbate the problems facing the public sector. Chapter 3 summarizes these risks. The ethical use of AI is fundamental to managing the adverse consequences of AI use in public policy. The ethical use of AI means that these systems should not harm humans. Rather, they are used to enhance overall human wellbeing. For example, an AI system that renders people jobless on a wide-scale, makes a biased decision against an ethnic minority applicant on eligibility for government welfare assistance or is used to propagate fake news on social media would be unethical. On the other hand, however, an AI system that improves anti-fraud measures through the reconciliation of multiple large data sets, facilitates medical diagnosis through image recognition, or enhances learning outcomes through tailored access to learning material, would be considered ethical and therefore, human-centered AI.

National level public policy response is needed to address these ethical issues. Inequality could rise due to unemployment, the lowering of wages for low-skilled workers, and the vulnerability of some communities to bias in AI-based automatic decisions. Control could increase due to

1. Bughin et al., 2018.
2. Government AI Readiness Index 2019.

state surveillance of citizens, robot-induced propaganda and fake news on social media and use of AI-enabled weapons like drones. The concentration of wealth could accentuate monopolies as a few firms with the AI resources could dominate the market and lead to net resource flows from the developing to the developed countries where these firms are based.

To manage the risks and maximize the opportunities of adopting AI in the public sector, the government should prepare an AI policy and governance frameworks to help guide the ethical use of AI and to provide clarity about AI principles and priorities. Following the adoption of AI policy and the development of a roadmap, an operating framework will anchor the principles as the use of AI is rolled out. Chapter 4 on governance and operations provides more details of the models and AI compliance frameworks currently in use. The models are fundamentally important to help guide the government in protecting the sanctity of human life throughout phases of AI adoption in the public sector. Governments also adopt basic principles to promote human-centered use of AI. These principles include personal data privacy, accountability, cybersecurity, transparency and explainability, fairness and non-discrimination, human control of technology, and human values.

A central innovation hub for AI could help pool scarce resources to support the initiatives of line ministries. In the use cases, most governments have set-up the main hub for AI that serves as a central authority over decentralized projects among line agencies. The AI hub helps them in several ways. It centralizes talent that guides and supports the line agency, connects industry expertise to the line agency, promotes research, and builds alliances with academic institutions and the private sector. It also helps connect with AI organizations internationally to exchange knowledge and resources. Neighboring countries that have a forum for coordination at the political level can develop regional AI innovation hubs suitable to many of the World Bank's client countries.

Innovation procurement frameworks provide agility for experimentation. In digitally advanced governments, a problem-driven request for proposal (RFP), rather than a tender with solution specifications, is developed and launched under innovative procurement methods. These methods allow an initial award of a small scope proof-of-concept contract to more than one vendor to compare a range of solution options and decide the best option for further scale-up. World Bank task teams could adopt these approaches in consultation with procurement colleagues and other available technical resources in the Bank, such as the GovTech team, Innovation Lab, Digital Development teams, Innovations in Big Data and Analytics for Development Program, and other sector colleagues.

Most early adopters are embracing a design thinking framework and agile methodology. These include a staged iterative approach to implementation—ideation (problem definition), conceptualization, proposal, procurement, prototype, testing, deployment, and scaling up. A feedback learning loop is built into the design at every stage.

Adopting a government-wide data fabric architecture will help governments leverage cutting-edge technologies to address data silos in a cost-efficient manner. The initial focus should be on foundational technologies, interoperability, open data, and standardization of data across government. Chapter 6 on AI building blocks illustrates the technology foundations for this architecture. The data fabric architecture will serve as the common denominator for standardized data interchange among the multitudes of subject-area specific applications such as an integrated financial management information system, payroll, tax administration systems, e-procurement, health management system, population census, and geographical information systems, among others. This architecture should be built on agile principles, evolve organically, and engender trust. Cloud computing offers immense opportunities to harness the power of such an architecture with agility. Inadequate foundational digital technologies, quality of data, and digital skills are the major barriers to AI adoption in developing countries and constitute critical elements of the digital divide.

AI threats during implementation need to be carefully assessed and mitigation actions planned. Threats include performance and bias, cybersecurity, control, and privacy. These risks should be managed at the implementation agency level, while broader ethical issues need policy action at higher levels. Chapter 3 offers steps toward risk mitigation. Involving stakeholders is crucial to mitigating risk, especially among groups most vulnerable to bias. Additionally, transparency and explainability could strengthen accountability. Compliance with privacy and data protection regulations is critical.

Governments and world leaders are instrumental in guiding the transition to automation and AI. They can provide leadership to influence the trajectory of AI adoption among citizens at national and international levels. This will help avoid adverse consequences and reap productivity gains. National governments could choose global guiding principles that will inevitably shape the acceptance or rejection of AI. Since AI will have a profound influence on service delivery, citizen engagement, and core operations, it is imperative to help formulate a cohesive governance model that supports the process of ethical implementation.

Priorities Going Forward

Based on the issues highlighted in the discussion, several priorities could be considered by policymakers.

- **Governments must adopt policies and governance frameworks that promote human-centric AI while maximizing opportunities.** A few aspects of the policy framework are mentioned below:

  - » **AI policy anchored in ethical principles would be essential.** It could be tailored to specific settings but should be approved at the policy level to provide the authorizing environment. Governments in many settings have issued AI strategies approved by the parliament, president, prime minister, or the cabinet. These policies should be based on ethical principles. Governance and operational framework are essential to specify broad guidelines and institutional arrangements. An innovation hub could be established to pool talent, establish partnerships with academia and the private sector, promote research, and facilitate experimentation by line ministries. The innovation hub should source the best talent through adequate incentives. Innovative procurement approaches should be adopted to leverage private sector skills with agility to allow iterative, problem-driven approaches to the RFP. The implementation teams should also manage the risks associated with AI, including bias, security, and unintended consequences, among others.

  - » **Promote transparency and accountability through inclusion and multi-stakeholder engagement at every step of the AI policy design and implementation.** Affected communities and populations should be informed and provided with avenues for contesting AI logic without delays and hurdles.

  - » **Adverse ethical implications of AI could be managed through broader economic policies.** These could include industrial policy, tax policy, competition policy, human capital policy, among others. These policies should aim to develop human capital, ensure fair competition, incentivize human-enhancing AI solutions, among others.

  - » **These policies should also promote digital skills, and broader education in science, technology, engineering, and mathematics (STEM) to support people as they adjust to the shifting nature of work in the coming decades.** Unskilled people and disadvantaged groups should be given special attention.

  - » **The regulatory framework to fight online propaganda, misinformation, libel, and cybercrimes should be given priority.** Also, governments could establish agency mandates to monitor policy compliance and track, prevent, and investigate disinformation to protect their citizens. Engagement with social media Big Tech—Facebook, Instagram, and Twitter—should aim at encouraging the deployment of AI tools and professional fact-check partnerships to take down content that is malicious, hateful, propagandist, and false.

  - » **Strengthen privacy, data protection, and civil liberties and monitor compliance, which is typically weak in most settings.** Promoting full disclosure of information being tracked by AI and robots through transparency frameworks should also be strengthened. Civil liberties and privacy are at a particular risk of infringement, which should be addressed through these regulatory frameworks.

- **Investments should be made in human capital and digital infrastructure.** AI research, digital skills, AI entrepreneurship, and foundational digital technologies could be prioritized.
  - » **Investments should be directed to fund research, education, and digital skills development programs in general and in AI in particular.** They could include scholarships, apprenticeships, and research funding in AI, computer science, STEM education, and AI-related disciplines such as data science. Special emphasis could be given to disadvantaged groups such as women, minorities, and those at risk of being left behind.
  - » **Innovative entrepreneurship could be promoted.** This could be done through an innovation fund, loan programs through state development banks, income-contingent loans for students or others, and small business loan programs. Variations of these funding modalities are already used in Brazil, China, Denmark, the European Union, Finland, Germany, Israel, and the United States (Mazzucato, 2015). AI could be one of the areas to be incentivized through these programs.
  - » **The innovation hub should be staffed with the best talent on market-based salaries.** These skills are in high demand and could easily drain overseas if not attracted and retained with appropriate incentives.
  - » **Data fabric architecture, including interoperability, should be considered for investments.** This will overcome silos, and leverage data assets for decision-making, compliance monitoring, and analytics. The initial focus should be on interoperability, open data, and data standardization. A hybrid cloud option, which combines on-prem data and cloud computing in a hybrid envirionment, should be explored to leverage the computing power at much lesser costs to pilot AI solutions.
  - » **Proof-of-concept and pilot AI projects could be the starting point for exploring opportunities.** Many governments have deployed AI to solve problems. Key use cases include citizen engagement, service delivery, regulatory compliance, decision analytics, fraud, and anti-corruption. Hackathons promote emerging talents and start-ups as seen in Austria, Estonia, India, Pakistan, Poland, and the United States.
- **Risks should be identified and managed, rather than avoided.** Good algorithm impact assessment framework models exist, which can be tailored to suit a country's context. The details could vary from context to context, but fundamental principles of risk mitigation are common. These include self-assessments, peer reviews, inclusion, and transparency.

>>>

Abbreviations

ACL	Access Control Layers
AI	Artificial Intelligence
ANN	Artificial Neural Network
API	Application Programming Interface
COTS	Commercial Off-The-Shelf
CPU	Central Processing Unit
DLT	Distributed Ledger Technology
FedRAMP	Federal Risk and Authorization Management Program
FMIS	Financial Management Information System
FOSS	Free Open Source Solutions
GAN	Artificial Neural Network
ICT	Information and Communication Technology
IoT	Internet of Things
IPC	Inter-Process Communication
IRS	Internal Revenue Service
ITU	International Telecommunication Union
ML	Machine Learning
MoH	Ministry Of Health
NGFM	New Generation Fiscal Machines
NGO	Nongovernmental Organization
NLP	Natural Language Processing
OECD	Organisation For Economic Co-Operation And Development
RFP	Request For Proposal
RL	Reinforcement Learning
SRT	Solicitation Review Tool
STEM	Science, Technology, Engineering, And Mathematics
TCO	Total Cost of Ownership
UN	United Nations
UNCTAD	United Nations Conference on Trade And Development
UNESCO	United Nations Educational, Scientific, and Cultural Organization
VPC	Virtual Private Cloud

>>>

Introduction

The World Bank launched the GovTech[3] Global Partnership in 2019 to support the modernization of client governments through the use of technology. To promote this effort, the Swiss State Secretariat for Economic Affairs partnered with the World Bank to produce a series of papers. This paper on artificial intelligence (AI) in the public sector, one in this initial series, offers insights drawn from the existing uses of AI in the public sector. The target audience is non-technical staff and policymakers who are developing and supporting the implementation of digital strategies for the public sector and drawn into conversations on the role of AI in modernizing the public sector. It refers to some fundamental technical concepts and provides more in-depth technical explanations in the appendices.

In recent years, governments have begun to investigate ways of leveraging artificial intelligence (AI) in public policy to better serve citizens, enhance compliance, and reduce fraud. The development of an appropriate policy and legal environment for AI could help countries stay ahead in commercial innovation, competitiveness, and international trade. The academic and professional research on AI ethics, policy, and regulatory reforms provides empirical and quantitative evidence on the opportunities and risks of AI adoption in the public sector The objective of this paper is to help World Bank's client governments understand the ethical issues and policy options associated with AI to promote ethical AI and to elaborate on the opportunities for AI adoption in the public sector.

3. GovTech is a whole-of-government approach to public sector modernization that promotes simple, accessible, and efficient government. It aims to promote the use of technology to transform the public sector, improve service delivery to citizens and businesses, and increase efficiency, transparency and accountability.

Advanced digital economies are increasingly adopting AI in both the private and the public sectors as part of their digital strategies. According to Bughin et al. (2018), AI can potentially contribute $13 trillion to the global economy by 2030. Use cases provide case studies for learning and also illustrate the potential for adopting AI in the public sector to enhance efficiency and quality of service delivery. The World Bank's client governments frequently request support on how to design digital transformation programs that can increase efficiency and quality of service delivery, improve citizen engagement, and modernize core government operations. One of the important areas of support is AI. With careful execution, AI programs can help a government to deliver services faster and more tailored to the needs of beneficiaries and citizens and the public administration charged with delivering them.

Public administrations that lack data collection capabilities, technical skills in the civil service and digital infrastructure are unlikely to be able to manage AI data requirements or benefit from the application of AI. But, generally, the volume of information produced and stored daily by people's movements, activities, and transactions is increasing, and combined with more computing power, such data can be used for effective analysis and policymaking. The speed of AI innovation and adoption has been fast; AI computation has been doubling every three months.[4] Governments could create readiness conditions to fully leverage the potential of AI as both the speed of government digitalization, store of data, and AI innovation evolve. The paper describes readiness conditions, such as governance arrangements, availability of digital data, local and international data source integrations, technical capacity, and infrastructure, for wider AI adoption and guides assessing these conditions. While this paper touches briefly on the policy as it relates to AI, a more detailed paper is forthcoming on AI policy aspects and elaborates on a comprehensive framework for policy domains. Also, this paper does not cover the re-engineering of business processes or project management aspects as they relate to AI adoption. Regulations and policies on data, privacy, security, transparency, and accountability, in addition to the business process review, must precede the actual implementation of AI.

The adoption of AI in government requires interagency oversight, coordination among interdisciplinary teams of policymakers, and requires the adoption of overarching policies to guide its use. In recent years, the public sector made impressive headway developing counsels and policies on AI applications, procurement, and adoption. The United Kingdom and Bahrain launched AI procurement guidelines across their governments (ANI 2019). The U.K. government published "A Guide to Using Artificial Intelligence in the Public Sector" (GDS and OAI 2019). Singapore issued the "Model AI Governance Framework" (PDPC 2020). The United Arab Emirates established the National Program for Artificial Intelligence.[5] The Organization for Economic Co-operation and Development (OECD) published the "Recommendation of the Council on Artificial Intelligence" (OECD 2019).

The use of AI poses substantial risks as models and data may be substandard or inaccurate leading to bias. Data privacy and security, and ethical use of AI pose major concerns in all contexts, but this is likely to be even more of a concern where there is a lack of transparency more generally, concerns over human rights, or what might be considered a "poor governance" environment. AI software is a "black box" that is opaque to policymakers. This means that algorithm opacity—the inability to detect design bias in constructing the algorithm—poses a major challenge for policymakers and auditors.

The adoption of digital solutions in government will require an investment in digital skills. The shift in the public sector needs from low-skilled to high-skilled workers will take place gradually over the long term, but it is a key consideration because building digital skills in the public sector and overcoming skills shortages more generally also takes time. The use of AI in the public sector may shift the characteristics of public sector employment and potentially result in job losses as more decision making becomes automated through the use of machine learning and models. However, the impact of adopting AI is likely to be less of a concern where the public sector wage bill is manageable, and the cost of labor is low. In some cases, demand for lower-skilled labor will decrease but whole scale substitution of professions with an AI program or machine is unlikely as the expert judgment will still be needed. The demand for high-skilled labor will likely increase. In some contexts, AI can automate systems of bureaucracy and create new job opportunities in, for example, policymaking, auditing, and resource management, jobs that require more analytical skills and judgment.

4. This is six times faster than the Moore's Law on processor speed doubling every two years – now closer to 18 months. See Artificial Intelligence Index 2019 Annual Report, 2019 Stanford Report, produced in partnership with McKinsey & Company, Google, PwC, OpenAI, Genpact and AI21Labs (Artificial Intelligence Index 2019 Annual Report, 2019).
5. For more information, visit the website of the National Program for Artificial Intelligence at https://ai.gov.ae/about-us/.

Methodology and Scope

This paper aims to provide some indications of the opportunities and risks of AI adoption in the public sector. It distills knowledge and guidance on the use of AI in the public sector. The AI use cases discussed in the paper demonstrate the potential to improve government services and create new opportunities to strengthen engagement with citizens.

The paper curates knowledge residing in public documents and aims to distill lessons learned on how to adopt and use AI as part of a public sector modernization strategy. The paper's primary scope is on governance-related aspects. Chapter 2 elaborates on the opportunities being availed by governments around the world through the use of AI. These opportunities should be availed while managing associated risks, which are discussed in Chapter 3. For maximizing opportunities and managing risks, governments need to adopt AI ethical principles and institutional arrangements, discussed in Chapter 4. Chapter 5 discusses the ethical dimensions that need a broader policy response at the national level. Chapter 6 enumerates the building blocks necessary for a successful long-term AI strategy.

The appendices contain information for practitioners. Appendix A provides technical information and additional resources for further support, and Appendix B highlights solutions that rely on AI for improvements in efficiency, scientific analysis, and prediction within the disciplines. To fully comprehend the impact that AI might have on governments, it is necessary to develop a solid understanding of key AI concepts. The paper does not offer in-depth coverage of work in specific sectors. The findings in the paper were validated through interviews with industry experts. Special efforts have been made to ensure the architectural design approaches discussed in the paper incorporate the best industry knowledge. The paper goes to great lengths to maintain a practical approach, with "hands-on" examples of architectures and applications.

The paper has limitations and AI adoption is not widespread. Actionable lessons in AI use are rare among client governments. Furthermore, there are limitations to the level of detailed, in-depth information, and availability of use cases from public resources.

Chapter 2 provides 14 use case examples of how AI has already been adopted in the public sector to address public sector issues such as how to control corruption. The associated risks of AI adoption are elaborated in Chapter 3. However, to harness the opportunities from AI governments need to develop the governance frameworks, address the ethical considerations and develop the building blocks of a government-wide AI architecture, issues discussed in Chapters 4, 5, and 6, respectively.

2

>>>

AI Opportunities

The public sector in advanced digital economies is rapidly adopting AI, notably in Austria, Brazil, China, Estonia, Israel, Mexico, Republic of Korea, Singapore, the United States, and the United Kingdom, among others. Noteworthy examples are also surfacing in Bank client countries. In this chapter, several AI use cases are provided to demonstrate the opportunities of AI already being harnessed in the public sector. Developing governments can also harness these opportunities to address some of the complex developmental challenges. However, governments in these countries need to address some of these challenges to maximize opportunities. The biggest bottlenecks in AI adoption are the availability of quality data, expertise, budget, and mindset for experimentation and problem-solving. Sectors or agencies that are more likely to adopt AI primarily have well-developed data infrastructures. These agencies are typically well resourced, experience compliance pressures, have a mission-critical need for analytical information for decision-making, or consider citizen engagement as an important element of the policy design. The role of leadership initiative is also important. Silos and closed systems with poor or inaccessible data impede AI development. Governments need to first evaluate the strengths and weaknesses of their data, procedures and AI policy framework before embarking on AI solutions.

Wider AI adoption in the public sector typically follows once prerequisites like sufficient digital infrastructure, adequate digital skills, enabling legal frameworks, and digital strategies are in place. The Oxford Insights' Government AI Readiness Index scores the governments of 194 countries according to their preparedness to use AI in the delivery of public services. The overall score is comprised of 11 metrics grouped under governance, infrastructure and data, skills and education, and government and public services. The data is derived from a variety of resources including desk research and the UN eGovernment Development Index. As presented in Table 1, the 2019 AI Readiness Index shows that Singapore comes first, with the rest of the top 20 mostly Western European countries.[6]

6. Government AI readiness indicators: https://www.oxfordinsights.com/ai-readiness2019.

> > >

TABLE 1 - AI Readiness Index

GOVERNMENT AI READINESS INDEX 2019	
Rank	**Government**
1	Singapore
2	United Kingdom
3	Germany
4	United States of America
5	Finland
6	Sweden
7	Canada
8	France
9	Denmark
10	Japan
11	Australia
12	Norway
13	New Zealand
14	Netherlands
15	Italy
16	Austria
17	India
18	Switzerland
19	United Arab Emirates
20	China

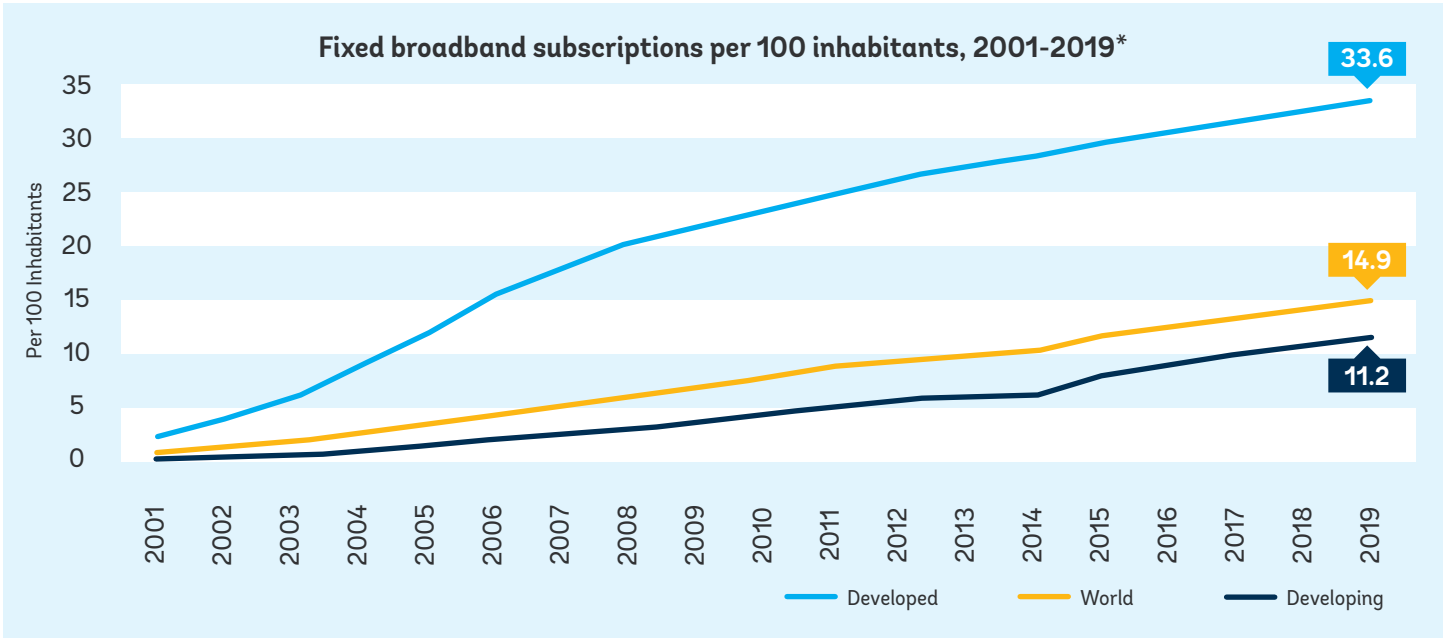
Source: Oxford Insights

Investments in data infrastructure, APIs, open standards, and data governance arrangements are all required for successful AI strategies in government, as discussed in the following chapters.

A digital divide exists across countries in terms of fulfilling the prerequisites for AI adoption. Most World Bank client countries are still far behind compared to the developed countries in terms of access to broadband, availability of digital skills, and adoption of relevant policies and legislation. Access to fixed broadband is significantly higher in more advanced economies, and the gap between the developed and developing countries has increased in the last 20 years according to the ITU (See Figure 1).

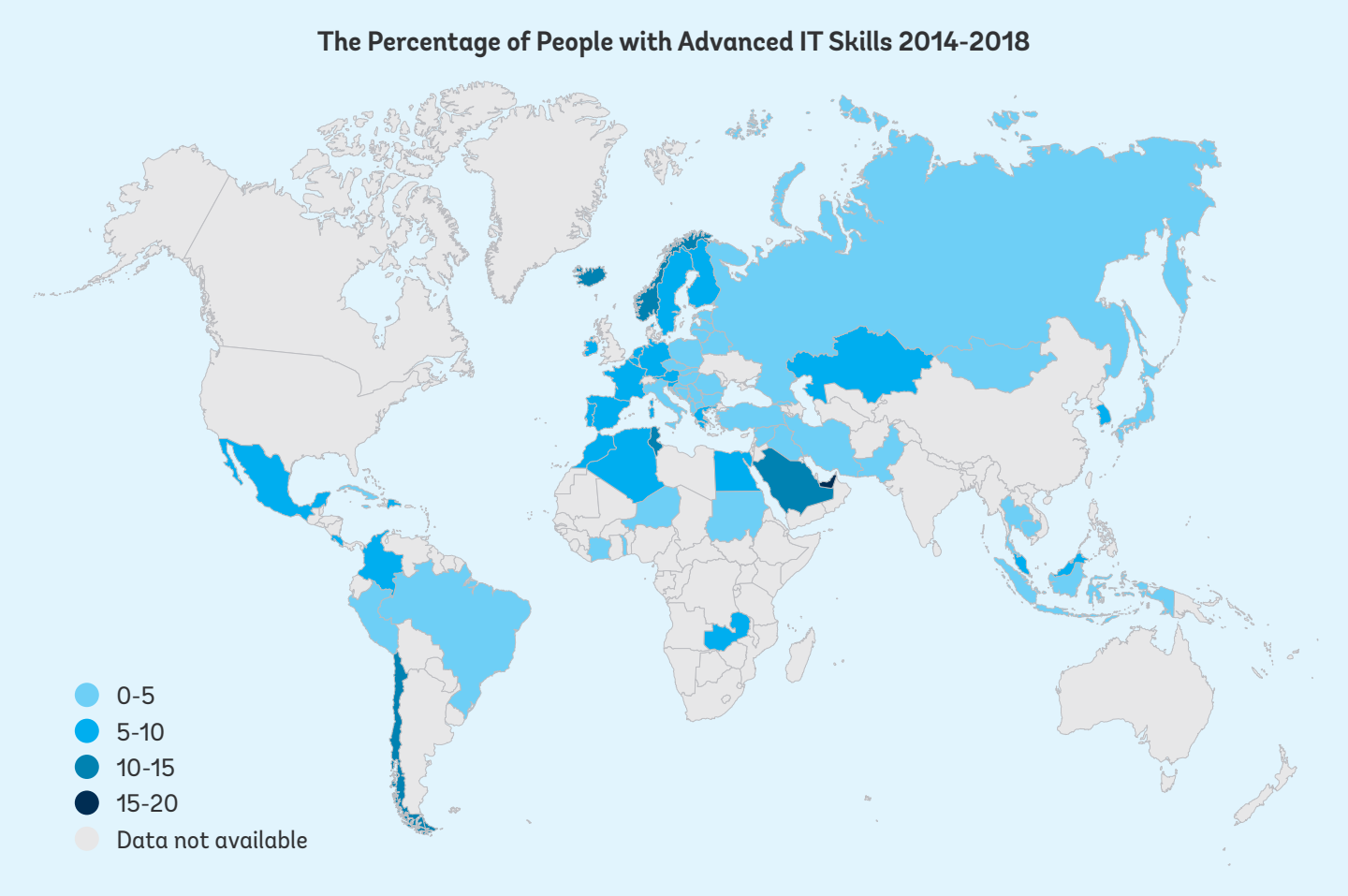
> > >

FIGURE 1 - Fixed Broadband Subscriptions per 100 Inhabitants, 2001–2019



Only 14.9 percent of inhabitants in developing countries have access to fixed broadband compared to 33.6 percent in developed countries (ITU 2019). Internet usage is limited to only 19 percent of the population in least-developed countries, compared to 87 percent in developed countries. There are only 67 data centers in 13 countries in Africa—of which 21 are in South Africa—compared to 1,237 data centers in 23 Western European countries. Advanced digital skills, such as writing software using a programming language, are also concentrated in a few rich countries. The disparity in information and communication technology (ICT) skills around the world is shown in Figure2. Europe is far ahead in terms of ICT skills, compared to Asia and the Pacific, Arab states, and Africa (ITU 2019). More generally, skills in data science and technology are scarce in low-income countries. Capacity constraint is an important issue. Those that have already adopted AI have promoted the adoption of additional AI capacity in government through sponsoring government officials to attend programs in academic institutions, introducing training programs in-country, or partnering with the private sector to provide expertise. Creating an innovation hub or a central AI unit as part of the centralized digital agency or as an independent agency helps maximize the use of scarce expertise.

FIGURE 2 - The disparity in ICT Skills across the Regions



Source: ITU, 2019

The legislative framework for data protection and privacy is relatively widely enacted, but policies that would allow accessibility to government-held data are mostly not in place in most developing countries. According to UNCTAD 2020, 132 of 194 countries, including 50 percent of the African countries and 57 percent countries in Asia-Pacific, have adopted data protection and privacy legislation. However, only seven governments out of 115 include a statement on open data by default in their current data management policies. Worldwide, only 7 percent of government-held data is fully open, and only one in every two datasets is machine-readable (Open Data Barometer 2020). There is also a significant lack of data sharing and interoperability within the government. Open, machine-readable, and interoperable data are some of the important preconditions for wider AI adoption in government.

AI use in government is therefore typically in a few advanced countries, and being taken up by digitally more advanced World Bank client countries. Some countries have adopted an AI strategy as a signal of the government's commitment to AI. At least 50 governments, in addition to the European Union, have developed or are in the process of developing a national AI strategy. Out of these, 37 have or plan to have either separate strategies in place for the public sector or a dedicated public sector focus embedded within a broader strategy (Berryhill et al. 2019). AI strategies are being adopted in some developing and emerging economies around the world including in India, Kenya, Malaysia, Mexico, Poland, Taiwan, and Tunisia (Dutton 2018).

AI patent applications (279,145) are also predominantly in the USA (55 percent), Europe, China, and Japan (Statistica, 2019). Similarly, AI research publications are dominated by developed countries (Microsoft Academic Graph, 2019).

Governments in some less advanced digital economies have started deploying AI to improve government effectiveness. While the scope and abundance of digital resources—talent, capital, infrastructure, and data—may be relatively limited, some developing governments have started piloting AI to address their development challenges.

The remainder of the chapter reviews several use cases where AI has been used in the public sector to address specific challenges: corruption, citizen engagement, customs compliance, health pandemic response, consistent judicial decisions, procurement compliance, taxation compliance, and policy, and audit efficiencies. Regardless of the stage of development, countries can develop AI initiatives based on their most immediate needs, but it is also recommended that the approach to AI should be part of the planning and accounting for future digital initiatives with a whole-of-government approach to infrastructure, standardization, governance, and execution.[7] The pattern of government adoption typically follows this typology of use cases:

- **CITIZEN ENGAGEMENT.** The introduction of AI tools such as chatbots that answer citizen queries. For example: Where is my ballot? Where is the nearest emergency department? How can I apply for social welfare benefits? Additionally, aggregation and pattern determination can be used to collect feedback from millions of citizens on a draft policy published online.

- **COMPLIANCE AND RISK MANAGEMENT.** AI systems are used to cross-reference and reconcile terabytes of data from multiple sources to create alerts for noncompliance. For example, financial intelligence units and central banks use AI to track illicit fund flow and beneficial ownership as well as terrorism financing to comply with the Financial Action Task Force. Tax authorities can use AI to track tax filers who use duplicate profiles to avoid taxation.

- **FRAUD DETECTION, PREVENTION, AND INVESTIGATION.** Closely related to compliance AI can be used to detect and prevent fraud for example by procurement agencies, anti-corruption units, or audit agencies.

- **BUSINESS PROCESS AUTOMATION.** AI automation tools can scan websites to get currency exchange rates and present information.

- **PERSONALIZED SERVICE DELIVERY.** Based on a profile, AI sends automatic alerts such as when to renew a driving license.

- **ASSET MANAGEMENT.** AI can be used to tracking asset movements across multitudes of systems, aggregating data from the Internet of Things devices.

- **ANALYTICS AND DECISION-MAKING.** AI or machine learning helps aggregate and cross-reference data such as household survey data with information on school enrollment, address changes, satellite images of floods, mosquito swamps, and pandemics to produce policy insights and identify areas needing greatest attention for targeted policy actions.

Use Cases

The following use cases illustrate real-world applications and opportunities for AI in the public sector in a range of contexts. The use cases provide a summary of AI initiatives to tackle corruption in China and Brazil, to engage citizens in Nigeria and Uganda, to improve efficiency and compliance in the United States customs administration, to tackle the COVID-19 response in Singapore and China, to improve public procurement in South Korea and the United States, to improve the effectiveness of the justice sector in China and the UK, the tax administration in Armenia, Mexico, and the UK, and audit in Canada and UK. The use case of health pandemic, COVID19, is developed in-depth to elaborate the concepts and the details of AI logic. Using the typology developed by Oxford Insights, each use case brief states the role of humans on the level of AI adoption in each application area. Table 2 describes the five levels of AI adoption. For examples of additional AI use cases, see Appendix B.

> > >

TABLE 2 - Role of humans - Five Levels of AI Adoption

Level	Description
Level 5	A fully automated system that never requires human intervention.
Level 4	Automation: A public service runs itself unless it hits an extreme case where it requires human intervention.
Level 3	Semi-autonomous: Computers monitoring and running (e.g., a regulatory sys-tem).
Level 2	Close supervision: Routine administration of systems (e.g., energy networks with difficult decisions referred to a human).
Level 1	Simple augmentation: Entering data, processing, identifying clusters of activity, and profiling, among others (e.g., in fraud detection).
Level 0	No automation: People-powered public services.

Source: Oxford Insights.

7. Governments can adopt an incremental approach when making investments to avoid the huge costs of full-scale infrastructure before development begins. Cloud solutions can provide any opportunity to reduce the total cost of ownership (TCO) for nascent projects. Cloud solutions enable incremental growth because they offer on-demand services at scale, without upfront investments or buying any on-premise servers. More information about cloud solutions and infrastructure is available in Appendix A.

In most instances, multiple AI methods and techniques, described below, are in use. A more detailed description of these techniques is provided in Appendix A on the AI Technical Primer.

Natural Language Processing (NLP): processing large amounts of natural language data by the AI systems. For example, NLP refers to the ability of an AI algorithm to read a text, convert speech into text, or vice versa. Specific use of NLP is chatbots, applications used to support online chat conversation using text or text-speech, typically used as customer support.

Data Mining: The ability of the AI algorithm to examine large amounts of raw data to determine patterns. For example, analyzing millions of comments from citizen feedback on an online policy document, and converting these comments into patterns of suggestions, approval, disapproval, etc. Common uses are:

- **CLUSTER ANALYSIS:** Clusters of similar objects or information are grouped to find patterns. For example, cluster analysis of tax filings to identify the same warehouse or same names of employees used by the same firm but using different registration numbers and titles to avoid or evade taxes.

- **FEATURE ENGINEERING:** Features are extracted from raw data to recognize patterns and classify information. For example, drone pictures of community rooftops could be used by the AI to identify types of roofs – thatch, corrugated, cement – and determine patterns of poverty for more targeted policy interventions.

Artificial Neural Network (ANN): AI algorithms that recognize relationships between different data sets similar to how the human brain analyzes such information. For example, reconciliation of two or more data sets to detect fraud patterns, medical image analysis to relate a specific feature in an image to a diagnosis and improving the diagnosis through adaptive learning. ANN techniques are also used in data mining.

Convolutional neural networks: A type of deep neural network, most commonly applied in visual imagery.

Generative Adversarial Network (GAN): Use of two or more neural networks to produce a realistic output. For example, fake videos can be made about some celebrity or popular figure by synthesizing two videos to misinform and manipulate public opinion.

AI in Corruption

Brazil

Governance Risk Assessment System

Use Case Brief	World Bank Artificial Intelligence Governance Risk Assessment System
Level 5	It is estimated that Brazil might be losing between 3 to 5 percent of GDP annually due to corruption. Over 48,000 companies tendered in public bidding processes between 2016-2018 in the State of São Paulo alone. Brazilian Government agencies can systematically identify public expenditure risks at this scale only through advanced digital technolo-gies.
Level 4	Government agencies do not have the tools or capacity to conduct sys-tematic fraud risk assessments. The current approach, which depends on manual input to a large extent, is time consuming, inefficient, and ineffective.
Level 3	Graph theory, clusterization, regression analysis, and supervised ma-chine learning.
Level 2	
Level 1	Level 3-4: Users must interpret the evidence concerning high-risk firms and agencies. The System analyzes complex networks of potential fraud with minimal effort.
Level 0	

Source: World Bank.

The World Bank Team in Brazil, with funding from the Disruptive Technologies for Development (DT4D) Trust Fund, developed an AI System that identifies 225 red flags of potential fraud in public procurement processes and can help improve expenditures. The World Bank partnered with the City of Sao Paulo, the States of Rio de Janeiro and Mato Grosso, and the Federal Ministry of Health to leverage the vast amounts of unused data to build a system to help improve their investigative and expenditure capabilities.

As part of the project, the World Bank created one of the world's largest data lakes, which currently includes 27 datasets with over 250 million data points and more than R$500 billion in public expenditure (approximately US $100 billion). This includes numerous sources and types of data: expenditure databases; electoral databases; beneficiaries of social

programs databases; blacklisted firms' databases; and electronic invoices. Overall, the system builds on:

- Analysis of over R$500 billion in public procurement in Brazil from 12 States and Federal Level.
- Analysis of over 15 million electronic invoices.
- Analyzed and geo-referenced over 750,000 firms and a Public Registry Dataset containing details about 30 million firms – HQ address, partners, data of incorporation, economic sector.
- Incorporated over 30,000 news feeds about corruption.
- Data on 20 million social program beneficiaries.
- Data on 30,000+ blacklisted firms.
- Data from 20 million politicians and 800,000 political donations.

The system optimizes the process of detecting fraud in public expenditure substantially, saving valuable resources – time and money – and increasing the effectiveness of audits and investigations. The system has, so far, led to the exposure of numerous high-risk cases, including:

- Identified over 420 firms that won bids against companies that have a high likelihood of being shell companies and reflecting potential bid-rigging. The winning firms have more than R$ 600 million in public contracts.
- Identified 857 companies that won bidding processes against firms that share at least one partner in common. These firms have executed at least R$ 800 million in contracts.
- 450 firms whose partners are beneficiaries of the conditional cash transfer program, Bolsa Família, which indicates that these individuals are potentially strawmen. These companies have more than R$ 600 million in contracts.
- Identified more than 500 firms owned by public servants working at the same government agency that has executed the contract. These cases amount to over R$ 4.5 billion in contracts.

The technology has a high potential for scalability across Brazil and beyond through the implementation of Scalable Data Unification, which drastically reduces the marginal cost of replicating the implementation of the algorithms and the system. This approach reduces the cost of replication by building a global public expenditure database schema upfront, based on identifying and converting local schemas—State's public procurement dataset—into that global schema. Therefore, instead of adapting all the complex algorithms necessary for extracting the 225+ red flags to match the schema of a single public expenditure dataset, the team did the opposite and now every new public expenditure dataset is converted to the global schema and the risk detection algorithms are implemented directly.

China

Zero Trust

Use Case Brief	World Bank Artificial Intelligence Governance Risk Assessment System
Strategic context	President Xi Jinping's policy of promoting technological innovations such as Big Data and AI in government reform. China has faced enor-mous challenges of controlling corruption and has 50 million employ-ees on the government payroll.
Problem statement	The extent of operational corruption among public officials.
AI methods	Natural language processing; Big Data; data mining; anomaly detection.
Role of humans	Level 2-3

Source: World Bank.

Zero Trust was developed by the Chinese Academy of Sciences and the Chinese Communist Party's internal control institutions to monitor, evaluate, and scrutinize the work and lives of public servants. Zero Trust can cross-reference more than 150 databases in central and local government systems. The system detects an individual's property transfers, infrastructure, construction, land purchases, and house demolitions. Zero Trust also detects unusual increases in a civil servant's bank savings, new car purchases, and if an official is bidding for government contracts or is doing so under the name of family members or friends. The system then calculates a probability that those actions are corrupt and alerts officials to highly probable cases of corruption.

Zero Trust was rolled out in 30 counties and cities and identified 8,721 government officials suspected of engaging in embezzlement, abuse of power, misuse of government funds, and nepotism. Some of these cases resulted in a prison sentence, most were allowed to keep their jobs after receiving a warning or minor punishment (Chen 2019). The future of Zero Trust is uncertain; the system faces backlash from public officials, and it may be decommissioned (Chen 2019).

AI for Citizen Engagement

Nigeria

DataCrowd

Use Case Brief	Public Spending Observatory AI Apps
Strategic context	Citizen engagement and feedback are helpful tools to complement formal mechanisms of accountability as they offer compelling insights for monitoring and evaluation of policies, project designs, and imple-mentation.
Problem description	The limited capacity of the agencies to receive, analyze, and respond to citizen feedback.
AI design	Natural language processing text matching.
Role of humans	Level 1-2

Source: World Bank.

A World Bank team is working in Edo State, Nigeria with Data Science Nigeria (DSN) to pilot an AI solution for citizen feedback to monitor project progress in sample locations. DSN has a mobile app called DataCrowd that is based on AI. The pilot was done over four weeks in May 2020. Its scope covered 77 locations in the state and collected citizen's feedback for the project, State Employment, and Expenditure for Results (SEEFOR). After initial positive results, the project is planned to scale up to cover three more states and about 350 locations. The AI solution has several features; the following were included in the Edo pilot:

AI-powered tag cloud. DataCrowd can summarize text and sentences, such as citizens' feedback through mobile phones, and instantly shows the keywords and their relevance. This AI feature was used on citizen feedback received during the Edo pilot.[8]

AI-powered geofencing. This feature instantly rejects a submission made outside of a geofenced location. This feature was used during the Edo pilot.

AI-powered image classifier. This feature can classify the contents of a picture. For instance, if a picture of a person is taken, the AI model can tell if it is a male or a female. Unlike the tag cloud and sentiment analyzer features, the image classifier feature is custom-trained based on the image data collected for a particular project, which always requires a lot of images to train. In the case of SEEFOR, the Research and Development Team is working on training the image classifier model to classify some of the SEEFOR images, especially under the public works category. This feature is particularly useful for quality assurance checks and when many images are being collected.

AI-powered image matching. This feature will allow DataCrowd to instantly match an existing image with a new image and report if they are the same or not. This feature is in development. It is expected to be useful as first-level data verification and validation when many images are being submitted by data collectors.

AI-powered opinion mining and sentiment analyzer. The sentiment analyzer feature can measure the sentiment pulse of text data, such as citizen feedback, and categorize sentences into negative, neutral, and positive sentiments. Although this feature exists on DataCrowd, it was not included in the pilot. It is useful for understanding the sentiment expressed in all citizen feedback and could be used potentially for scale-up. In the pilot, the project authorities were able to obtain citizen feedback on civil works and confirm various aspects of the project's implementation progress, including location, performance, quality, and completion.

8. The tag cloud is available at https://datasciencenigeria.github.io/DataCrowd/.

AI in Customs

United States

Northern Border Surveillance System

Use Case Brief	Northern Border Remote Video Surveillance System
Strategic context	The US Customs and Border Patrol is one of the world's largest law en-forcement organi-zations and is charged with keeping terror-ists and their weapons out of the U.S. while facilitating lawful international travel and trade There are 300 ports of entry into the United States that need to be secured without disrupting trade and transit.
Problem statement	Concerns of illegal trade, including drug smug-gling and human traffick-ing, and weapons entering the US under the mandate of the U.S. Cus-toms and Border Protection Agency.
AI methods	Convolutional neural network, computer vi-sion, pattern matching, anomaly detection, prediction.
Role of humans	Level 2

Source: World Bank.

Border patrols require vigilance to stem illicit trade including drug smuggling and human trafficking. The use of AI to combat illicit activities is on the rise. The U.S. Customs and Border Protection Agency uses the Northern Border Remote Video Surveillance System (NBRVSS). The NBRVSS can detect and monitor vessels from miles away and alert authorities when it recognizes unusual vessel movements. It commenced before 2016 and utilizes many radio towers equipped with computer vision that spot anomalies in vessel behavior and allow agents on the ground to intercept potential sources of contraband entering the United States from the Canadian border.

AI in Health

The unforeseen rise of unseen global threats to human health and safety has put AI on the frontlines of disaster response efforts. Furthermore, sudden changes in the behavior of the human population challenge existing models and stressed predictive AI systems to the breaking point (World Economic Forum 2018). The speed of response to disaster events substantially impacts the extent of economic losses and human suffering. Delays occur due to a lack of information, analytics, and predictive modeling of the best course of action. Data silos exacerbate the issue, leaving data locked up and inaccessible to communities.

AI can sort through regional data and identify which aspects of the overarching infrastructure have the greatest impact on resilience. It can simulate various disaster events in a region to uncover vulnerabilities and assist with the formulation of disaster recovery plans. A data fabric can hold data from silos and enhance disaster preparation by coordinating emergency information exchange capabilities. During a disaster, predefined use cases can equip first responders with better tools for understanding the local context to take more precise action. Reinforcement learning (RL) is a strong candidate for this type of future simulation.

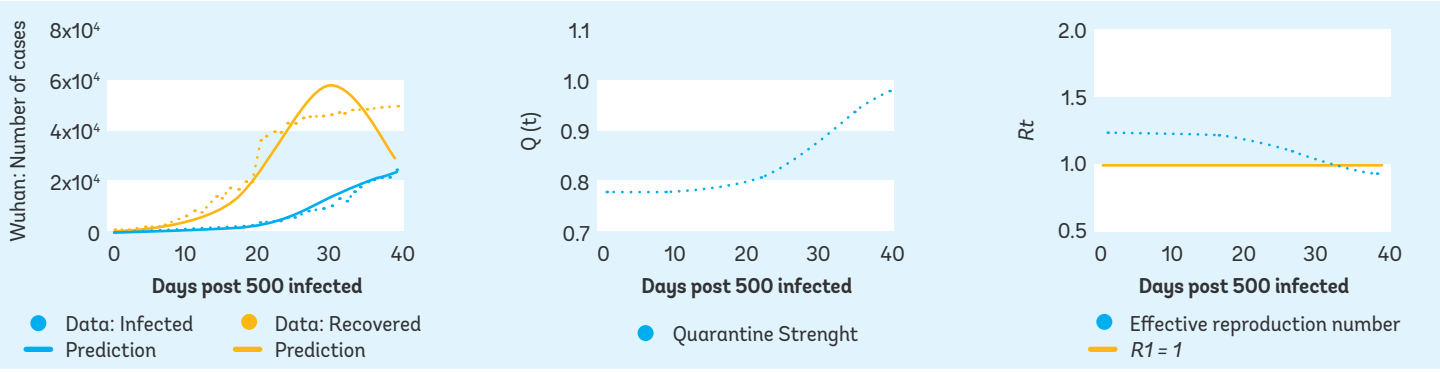
Use Case Brief	Contact Tracing and Temperature Detecting Camera Apps
(Sense-Time, Megvii, WeChat)	The US Customs and Border Patrol is one of the world's largest law en-forcement organi-zations and is charged with keeping terror-ists and their weapons out of the U.S. while facilitating lawful international travel and trade There are 300 ports of entry into the United States that need to be secured without disrupting trade and transit.
Strategic context	Contact tracing and screening to target policy response on quarantine for minimum disrup-tion on economic life and contain the spread of COVID-19.
Problem statement	The economic shutdown to contain COVID-19 has impacted jobs and growth and has trig-gered an unprecedented economic recession in many economies. Smarter and targeted response on quarantine and social distancing policy could save economies from economic disasters.
AI methods	Artificial neural network, reinforcement learn-ing, data mining, predic-tion.
Role of humans	Level 3

Source: World Bank.

On a more practical level, in light of the COVID-19 pandemic, AI methods are being employed in earnest to model potential effects of quarantine models and screen patients for potential infections using facial and thermal recognition models. Figures 4-7 demonstrate recent modeling of quarantine methods using artificial neural networks (ANNs) from several countries with varying degrees of quarantine policy. Predicted data are modeled in using solid lines, while actual observed data uses dots. Note the relative accuracy of the predictions for most sources and the detection of possible disparity of infection due to potential under-reporting (Dandekar and Barbastathis 2020).

> > >

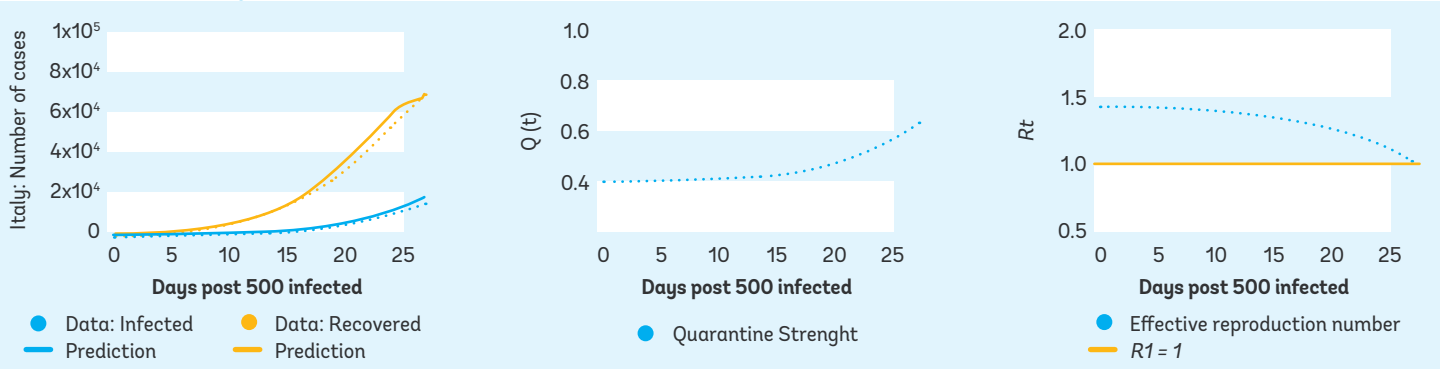
FIGURE 3 - Wuhan Neural Network Model with Quarantine Control



Source: Dandekar and Barbastathis 2020.

> > >

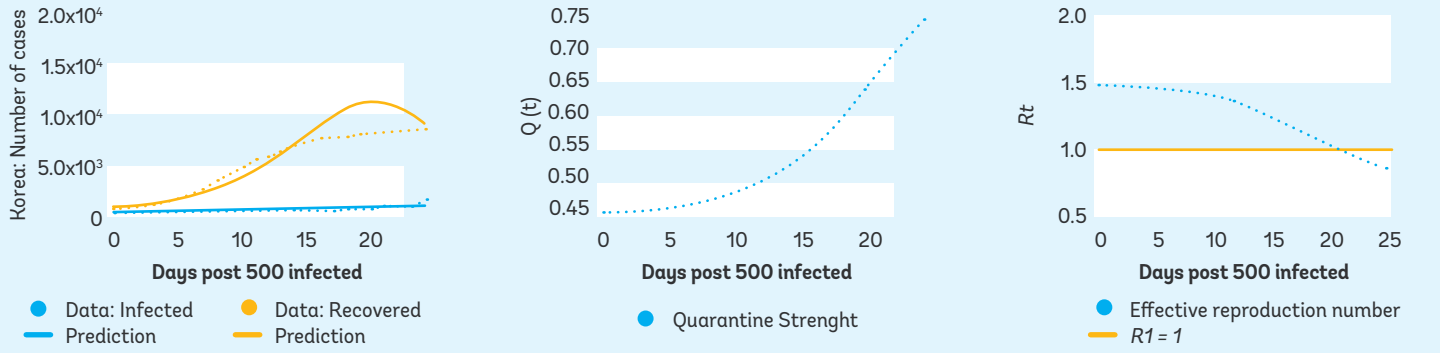
FIGURE 4 - Italy Neural Network Model with Quarantine Control



Source: Dandekar and Barbastathis 2020.

> > >

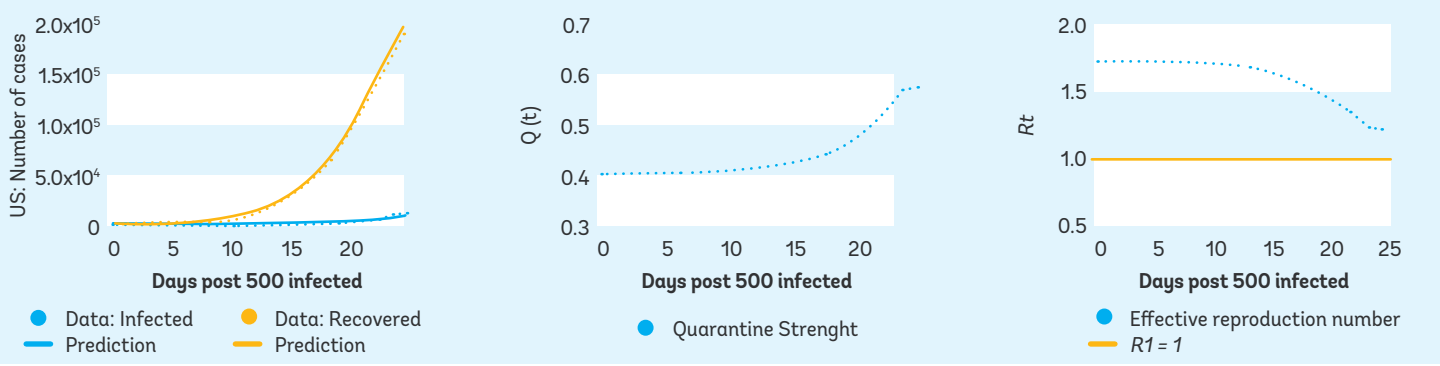
FIGURE 5 - South Korea Neural Network Model with Quarantine Control



Source: Dandekar and Barbastathis 2020.

> > >

FIGURE 6 - U.S. Neural Network Model with Quarantine Control

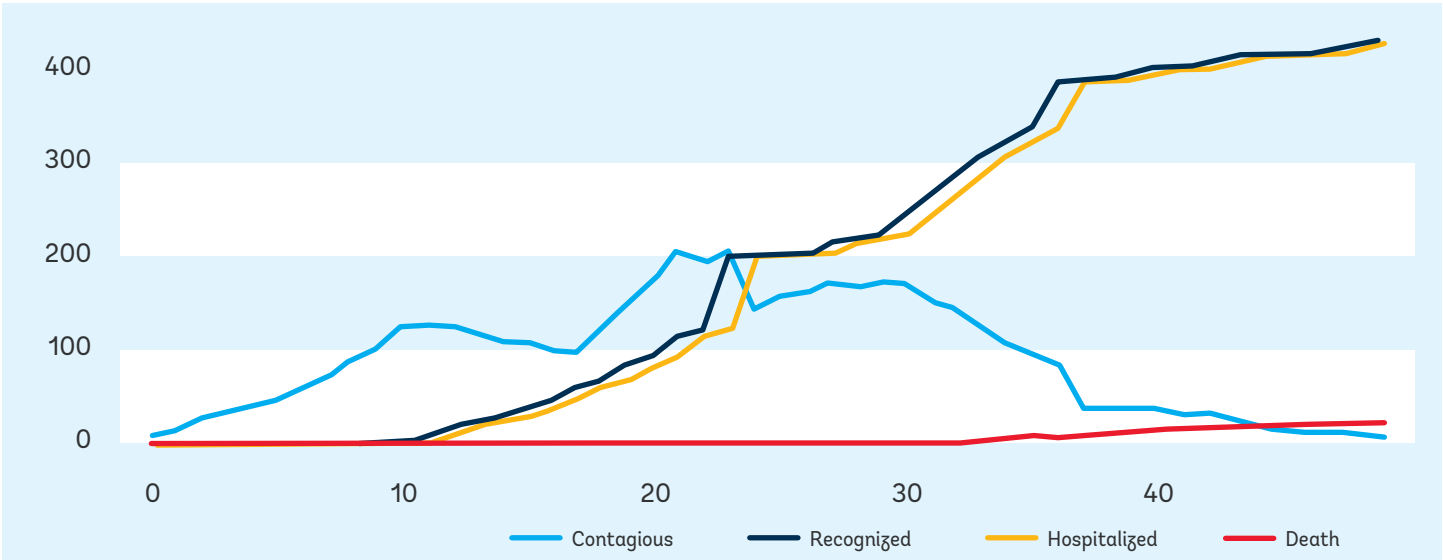


Source: Dandekar and Barbastathis 2020.

The impact that AI can have on pandemic mitigation continues with additional AI methods that are in place, beginning with facial recognition—it uses face scans to detect symptoms. Upon entering the Tampa General Hospital, patients are given an automatic face scan that determines signs of fever, including sweating and increased skin temperature within 0.3 degrees of variance over 1-3 seconds. In another modeling example, RL models learn to combat the illness using policies of quarantine and hospitalization to identify the most successful policy model (Chilamkurthy 2020). Figure 8 illustrates the results of the AI analysis, revealing the potential to thwart the progression of the pandemic within 50 days.

> > >

FIGURE 7 - Results of COVID-19 Analysis by AI



This is the bast action policy, where the agent brings down the whole virus in 50 days. At some point, this agent also allows increase in virus spread (25th to 29th day).
Distribution of action space {7: 3, 11: 4,5: 21,15: 4,13: 2,12: 2,0: 2,3: 1,1: 4, 10: 1,14: 1,9: 1,4: 2, 2: 2}

Source: Chilamkurthy, 2020.

Lastly, contact tracing applications are emerging on the front lines of halting the spread of infectious disease. One notable example taps into Bluetooth communication broadcasts from smartphone devices. In this system, data from a confirmed infected person's cell phone can be extracted to list the Bluetooth broadcast "chirps" detected within the phone's database. By uploading this information to an interoperable data platform, the signatures of the chirps can be cross-referenced with chirps from other reported infections. If the information is made available through an application interface on any smartphone, then the general public can determine whether they have come in contact with known sources of infection and can take measures to mitigate the risk of further exposure and potentially seek treatment, if the potential of infection is high due to repeated or multiple contacts. While this concept is possible to implement on a local basis and efforts to implement this technology are documented by both Google and Apple in partnership, no successful implementation exists for the general public at this time. The key problem is interoperability using a large-scale data fabric solution, though the two tech giants assure the public that a solution will exist in the coming months (Apple 2020).

Singapore

Bot MD

Use Case Brief	Hospital and health information app for doctors and front-line health workers
Strategic context	Doctors and front-line health workers need information on the latest health protocols, staff rosters, operational directives, and dosage to effectively manage the COVID-19 pandemic.
Problem statement	Health facilities are under immense pressure to respond to the un-precedentedly high volume of COVID-19 patients. An effective re-sponse needs timely information for a coordinated team effort.
AI methods	Natural language processing, data mining, chatbot, search.
Role of humans	Level 1

Source: Bot MD.

Bot MD is an AI Chatbot mobile app that acts like 'google' for hospital and clinical information on COVID19 for doctors and frontline health workers. Developed in Singapore, more than 13,000 doctors in 52 countries are now using the app. The doctors, front-line health workers, and Ministry of Health (MoH) officials can type a question and the app can provide information on staff rosters, health protocols, drug formulary information, disease guidelines, operational directives, and latest MoH circulars. The app was developed by Tan Tock Seng Hospital (TTSH), and the MoH's IT team in 2018. The system uses AI to predict situations before they occur, provide information for decision-making on resource allocation to deal with the pressures. These resources could include manpower, equipment, supplies, medicines, hospital beds, intake centers, etc. (The Straits Times, April 2020).

AI in the Judicial Sector

The inconsistent application of law and long pendency of cases due to excessive workloads plagues the judicial sector. AI has the potential to enhance consistency and efficiency in the judiciary.

China

Similar Cases Push System

Use Case Brief	Similar Cases Push System
Strategic context	China's Supreme People's Court is promoting the policy of "Similar Judgments in Similar Cases" to promote consistency in judicial deci-sions.
Problem statement	Inconsistent application of law during judicial decisions.
AI methods	Natural language processing, Big Data, data mining, and automation.
Role of humans	Level 1

Source: World Bank.

Before harnessing AI, the Chinese judiciary adopted policy measures that enforced the use of technology. China's Supreme People's Court (SPC) issued a policy of "Similar Judgments in Similar Cases" to promote consistency in the judicial decisions. Initially, a system called "Review and Approval of Judgement System" was implemented through which superior courts would review the judgments of the courts submitted online through the system. However, this led to inappropriate interference and delays. The use of this technology, therefore, was canceled. Under the new guidelines, the principle of self-accountability and independence was established, under which the final judgment is issued by the concerned judge without higher-level approval. However, this has led to the risk of inconsistent judgments across jurisdictions. SPC policies now require judges to research similar cases and cite these cases in judgments to ensure consistency.

To support this research, the Chinese judiciary is piloting AI in some provinces to improve consistency. Under this implementation, all prior judgments were digitized and stored in a database. Next, the SPC deployed NLP AI capabilities, through the Similar Cases Push System, to match key text relevant to pending cases using the database. The system presents relevant judgments before a judge using a pre-populated judgment template that the judge reviews and edits. The system reduces the time it takes to formulate a written judgment and all legal procedural documents by 70 percent and 90 percent, respectively (China Daily 2019).

Also, an AI pilot program records court proceedings. Some courts in China are now using AI speech recognition products to directly translate the court hearing recordings into texts in real-time and convert these into written court proceedings using Speech-to-Text NLP methods.

United Kingdom

Legal AI Tools and Bots

Use Case Brief	Robot Lawyer—DoNotPay App
Strategic context	Legal document processing in cases of litigation.
Problem statement	An AI legal assistant is necessary for improvements in the analysis of legal contracts; support of private legal bureaucracy among citizens; and guided legal advice.
AI methods	Natural language processing, chatbots.
Role of humans	Level 4

Source: World Bank.

Automated AI legal assistants and lawyers have surpassed human-level accuracy. An AI bot performed better than human lawyers in competitions for accuracy and efficiency held in London and Tokyo. In London, human lawyers from prominent law firms in the United Kingdom predicted whether the Financial Ombudsman would allow an insurance claim. Of the 775 total predictions, the AI "Case Cruncher" emerged on top with an 86.6 percent accuracy rate compared to 66.3 percent among 100 human lawyers (BBC News 2017).

DoNotPay, touted as the world's first robot lawyer, helps users dispute parking tickets. In one month, post-launch, DoNoPay.com helped people overturn 160,000 of 250,000 parking tickets—a success rate of 64 percent (King, n.d.). DoNotPay has now expanded its offerings to airline ticketing disputes and subscriptions. Other lawyer bots are also in operation. These include Ross (United States) for cash research powered by Watson AI APIs; *Billy Bot* (United States), which takes the role of a junior clerk to guide users to free online resources and to find legal representation; and *i-LIS*, South Korea's first intelligent legal assistant for legal research.

AI In Procurement

Central procurement agencies in governments face challenges when ensuring regulatory compliance of procurement among a large number of government entities. Central procurement agencies cannot manage the magnitude of procurement activities occurring across the government because the capacity of human agents is limited.

United States

Solicitation Review Tool

Use Case Brief	Solicitation Review Tool
Strategic context	Legal compliance in the tender documents with Section 508 of the Rehabilitation Act.
Problem statement	Reviewing hundreds of complex and voluminous bidding documents, issued by the federal agencies, to ensure compliance with regulations.
AI methods	Natural language processing, Big Data, data mining, feature engineer-ing, and automation.
Role of humans	Level 3

Source: World Bank.

The U.S. government is harnessing the power of AI to strengthen procurement compliance. The U.S. General Services Administration (GSA) has an Office of Government-wide Policy, which developed a new pilot using AI for scanning bidding documents to determine regulatory compliance. The tool is known as the Solicitation Review Tool (SRT).

The SRT AI platform uses NLP, text mining, and machine learning (ML) algorithms to scan and review whether federal solicitations posted on fbo.gov are compliant with Section 508 of the Rehabilitation Act. It alerts responsible parties of noncompliance and flags the need for corrective actions. Through the independent review, the predictions have an accuracy of 95 percent.

This innovation substantially alleviates the human resources needed to identify, audit, and enforce compliance. The SRT platform is innovative because it helps the GSA focus on limited available resources on noncompliant solicitations. The SRT AI platform has expanded to predict whether solicitations comply with other federal regulatory requirements, such as cybersecurity or sustainability (GSA 2018).

Korea

Bid Rigging Indicator Analysis System

Use Case Brief	Korea's Fair Trade Commission's Bid Rigging Indicator Analysis System
Strategic context	The Fair Trade Commission ensures fair competition in procurement practices in the government.
Problem statement	Unfair practices in procurement, using bid-rigging, to beat the compe-tition.
AI methods	Natural language processing, Big Data, data mining, feature engineer-ing, automation.
Role of humans	Level 2-3

Source: World Bank.

Korea is cracking down on bid-rigging through the use of AI. Officials converted a manual process that was in place since 2004 to detect bid-rigging cases using AI. The introduction of the AI system greatly increased speed and effectiveness.

Bid rigging refers to collusion between procurement officials and a pre-ordained vendor to award a contract using corrupt practices. Bid rigging can take various forms, including short bid submission windows, split procurements to capture funds below detectable thresholds, significant change

orders, and substitution of low priced items with high priced items after the award. Korea's Fair Trade Commission (KFTC) is leveraging an AI and analytics platform, the Bid Rigging Indicator Analysis System (BRIAS) to combat corrupt practices.

Before the introduction of the automated AI solution, the KFTC collected and manually analyzed hard copies of bid-related documents from major public organizations such as the Public Procurement Service, Korea Expressway Corporation, and Korea Electric Power Corporation, which issue large-scale public projects. Presently, the KFTC collects and analyzes this information electronically and flags cases of suspicious bid-rigging activities.

In total, 322 public organizations must report their bids to the KFTC. Construction projects over ₩5 billion and tenders for procurement of goods and services over ₩500 million must report to the KFTC. The affected public organizations must report related data into BRIAS within 30 days of selecting a bidder. The organizations that use internal bidding systems may transmit bid data to the KFTC in real-time using BRIAS APIs. The others must report bid information to the KFTC portal. The information submitted includes the following features:

- The organization's information on the executive agency and issuing agency.
- Procurement information: types and methods of tenders, the date and contents of tender notices, and the estimated price set out by issuing organizations before tender notice, which serves as a benchmark to determine the tender amount for the successful bidder.
- Bid evaluation information: the ratio of bidding price to the estimated price, the number of bidders, bidder-based tender details, company information for successful bidders, and the number of unsuccessful bids.
- Contract execution information: the number of estimated price increments and alterations to bids.

The KFTC weights the features according to a preset formula and uses the data to analyze the probability of bid-rigging quantitatively. An automated system calculates and assigns a score between 0 and 100 to the procurement item or contract. The higher the score, the more likely the concerned bid is rigged. The KFTC sends flagged bids to external departments for further investigation. In one example involving 12 construction companies for the Seoul subway, the KFTC detected bid-rigging, and the government imposed a surcharge amount of ₩5.108 billion.

AI in Tax Compliance

Tax administration authorities in governments consistently grapple with the challenge of ensuring reasonable tax compliance. Tax authorities are better positioned to pilot AI tools to strengthen their mandate on compliance for several reasons.

- Generally, tax agencies have more data assets than other agencies. The capacity is generally higher.
- Tax compliance and collections directly affect fiscal sustainability targets and the political agenda of most governments with an interest in funding capital and social projects as promised in their manifestos.
- Many tax agencies across the world deploy data warehouses, data analytics, and, lately, AI projects to leverage the power of technology to promote their mandate through a shift to risk-based auditing techniques for tax compliance.

More than 32 tax administrations worldwide have changed their strategies from a traditional data-oriented audit to a risk-based, cooperative compliance approach that relies heavily on analytics during the assessment process (Microsoft, PWC 2018). The huge data assets typically process a historical record of tax payments, electronic value-added tax (VAT) invoices, income tax returns, and personal and company information. By deploying a Big Data architecture, capturing all the relevant structured and unstructured information in one database, and running AI and analytics tools, tax administrations may significantly improve their effectiveness. These solutions offer a complete picture of businesses or individuals using risk-based compliance assessments. Examples highlight how tax authorities are deploying AI and analytics for common problems faced by most tax authorities.

Armenia

AI Use in Tax Administrationl

Use Case Brief	New Generation Fiscal Machines
Strategic context	Tax evasion among businesses and individuals.
Problem statement	Tax evasion practices remain undetected as evasive practices fail to cross-reference fiscal records that may reveal correlations resulting in the detection of tax reporting anomalies.
AI methods	Natural language processing, Big Data, data mining, and cluster analysis.
Role of humans	Level 2

Source: World Bank.

Tax evasion is carried out in many ways. One of the most common practices among small businesses and individuals of lower-income is to remain below revenue thresholds to benefit from lower tax rates. An existing business may open a new business when the existing firm reaches the threshold. Similarly, a business will split into several small businesses, often using the names of friends and relatives. The aim is to avail of the lower rate compared to the appropriate VAT. To combat this, tax administrators will analyze tax data and identify the interconnectedness of split entities. Armenia's tax authorities handled this issue using several techniques.

Single administrative document (SAD). Producing a SAD about importers of goods is one way of having a fuller view. Analytics detect whether a taxpayer is always importing the same goods from the same country and the same enterprises repeatedly. Moreover, electronic invoices help detect groups of taxpayers that use identical storage for imported goods. The tax administration investigates the anomalies.

Cross-matching of sales and invoices. Cross-referencing sales and invoice data provides important insights into various sellers' revenues. Armenia's tax authority collects data from the registration database—new generation fiscal machines (NGFM) or cash registers connected to the agency's servers—and invoice databases. The invoice database detects when a variety of entities are selling goods from the same warehouse. The NGFM data reveals when a group of taxpayers use a variety of fiscal machines at the same location. The registration database reveals when different enterprises have the same founders. Such suspicious anomalies are subject to detailed audits, which may not necessarily be tax fraud but need deeper scrutiny.

Taxpayer's employees. AI and analytics can detect suspicious cases in which different employers declare identical groups of employees in income tax filings or when a closed and reopened business hires identical employees. Employee information is obtained by linking a social security number and person identification database using Big Data infrastructure.

AI and analytics on sales data from the cash registers. The Monitoring Center leverages information received from NGFM. Some patterns call for further scrutiny. For example, if a fiscal machine does not work throughout the day, and the taxpayer prints 100 or 200 receipts within one or two hours, this flags a falsified fiscal amount with false receipts without actual sales. Some taxpayers print a single receipt with an unrealistic amount at the end of the day. All such cases are under control since the Monitoring Center automatically sends notifications and requires explanations. If no reasonable explanation is given, the case goes to audit.

Comparison of data from utility providers. The data from water, electricity, and gas, for example, reveals enterprise expenses, which demonstrate a logical correlation with the total amount of reported sales for a particular line of business. Again, this cross-referencing and correlation shows valuable insights.

The outcomes of Big Data analysis. Using targeted audits conducted during the first years of implementation of NGFM, the tax administration reduced the number of audit cases by about 2.5 times over recent years. The effectiveness, measured by the average amount of additional tax per audit, grew constantly over recent years. Also, the agency achieved substantial cost savings due to the rapid reduction of the number of local tax administration offices. Armenia reduced the number of local tax offices from 52 in 2009 to only two offices in 2017 (IOTA 2018). These are the departments for large, medium, and small taxpayers.

United States

Palantir Gotham Platform

Use Case Brief	Palantir Gotham Platform
Strategic context	Tax refund fraud, identity theft, and compli-ance.
Problem statement	AI is necessary in detecting tax evasion and conducting criminal inves-tigations in cases of tax fraud and identity theft.
AI methods	Cloud computing, Big Data, analytics, aggre-gation, and automation.
Role of humans	Level 1-2

Source: World Bank.

In 2011, the Internal Revenue Service (IRS) created the Office of Compliance Analytics (OCA) to construct analytics programs that could identify potential refund fraud, detect taxpayer identity theft, and handle noncompliance issues efficiently. OCA leverages an advanced analytics program that relies on the use of Big Data and predictive algorithms to reduce tax fraud. In 2016, significant organizational changes took place when the OCA and Research, Analysis, and Statistics merged to create the Research Applied Analytics and Statistics (RAAS) division. RAAS leads a data-driven culture through innovative and strategic research, analytics, statistics, and technology services in partnership with internal and external stakeholders. By combining AI and advanced analytics platforms, RAAS extracts value by leveraging vast amounts of proprietary data stored within the IRS legacy computers.

The IRS uses the Palantir Gotham platform to run its Lead and Case Analytics (LCA) service. Special agents and investigative analysts in IRS Criminal Investigations use LCA to "generate leads, identify schemes, uncover tax fraud, and conduct money laundering and forfeiture investigative activities" (Federico and Thompson 2019).

The various divisions of the IRS have access to several data mining applications. These include the Investigative Data Examination Application—formerly known as Investigative Data Analytics; LCA; Return Review Program (RRP); Financial Crimes Enforcement Network Query; and Compliance Data Warehouse. In 2016, RRP generated more than 693,000 identity theft leads, with a 62 percent accuracy rate and more than 103,000 other nonidentity fraud leads with a 49 percent accuracy rate (U.S. Department of the Treasury 2017).

AI in Tax Policy

United States

AI Economist

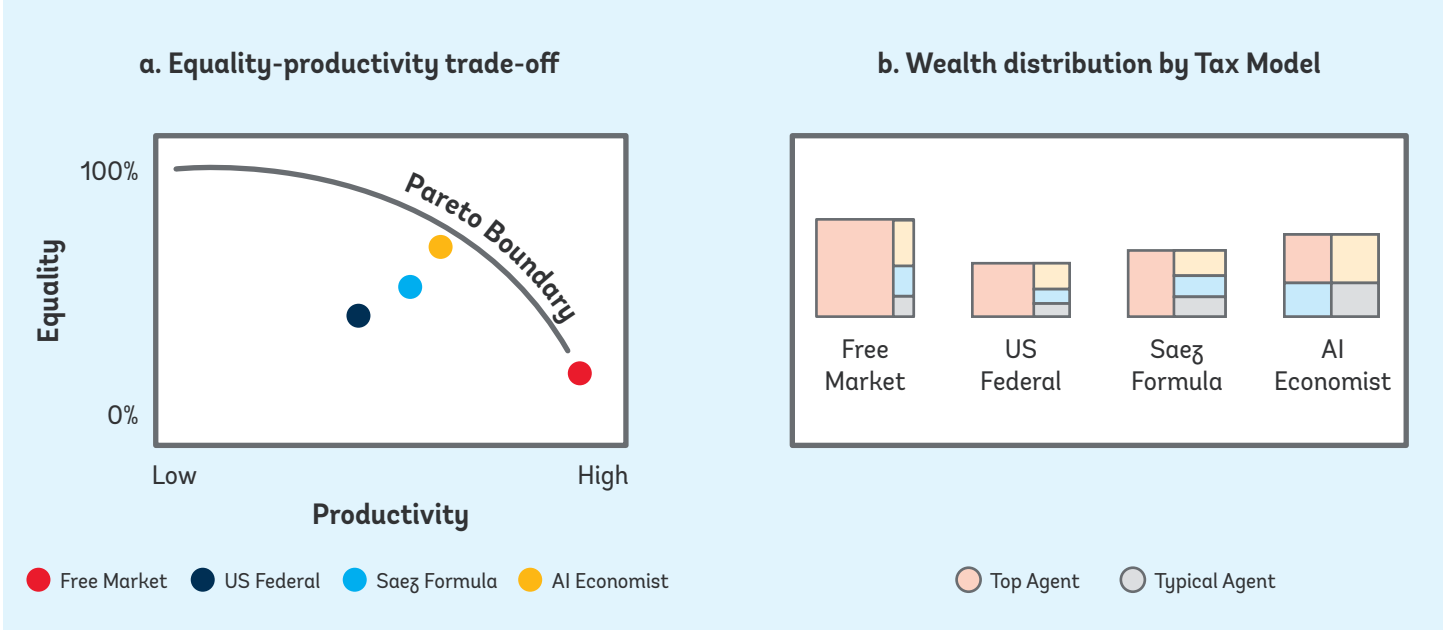
Use Case Brief	AI Economist
Strategic context	Macro-fiscal policymakers need better data and analytical reporting to design data-driven policies.
Problem statement	Data-driven macro-economic policies are hampered by a lack of data, skills, and robust models.
AI methods	Artificial neural networks, cloud computing, Mechanical Turk, and au-tomation.
Role of humans	Level 1

Source: World Bank.

Modeling data-driven tax policies in most developing countries is hampered by a lack of reliable data, forecasting skills, and robust models. These impediments could be overcome through the use of emerging AI tools if concomitant analog complements are in place. The challenge in most settings is devising a tax policy that optimizes equity and productivity. The AI Economist employs AI models based on RL algorithms to model and predict tax policy design through data-driven simulations using a two-level RL framework composed of agents (workers) and tax policy to model and learn the effects of dynamic tax policies in principled economic simulations. The framework does not use prior world knowledge or make any modeling assumptions. It can optimize for any socioeconomic objective. It learns from observable data alone. Though the framework is not yet deployed in government, results show that the AI Economist can improve opportunity costs and trade-offs between equality and productivity by 16 percent when compared to a prominent tax framework proposed by Emmanuel Saez, professor of economics and Director of the Center for Equitable Growth at the University of California at Berkeley. The framework captures even larger gains over an adaptation of U.S. federal income tax in the free market (Zheng et. al. 2020).

> > >

FIGURE 8 - An Optimal Tax Policy Optimizes a Balance between Equality and Productivity



Source: Zheng et al. (2020).

Notably, the AI Economist leveraged real-world human actors in the roles of workers competing with AI-driven policy models that evolved based on human interactions. Figure 9 compares the overall results of the study. They take into account the Pareto boundary, which is the event horizon where marginal benefit and cost trade-offs result in reduced productivity. Note the parity in wealth distribution among sectors of society and the overall gain in productivity due to the tax policies enacted by the AI Economist model. The AI Economist is in active development with plans for open-source distribution and government engagements shortly.

Canada, UK

MindBridge for AI Auditor

Use Case Brief	MindBridge AI Auditor
Strategic context	Oversight, assessment of the effectiveness of risk management, con-trols, and governance through external and internal audit.
Problem statement	AI models help maximize the efficiency of document analysis for legal infraction detection and policy audits.
AI methods	Natural language processing, Big Data, data mining, anomaly detection.
Role of humans	Level 1-2

Source: World Bank.

Private sector audit and assurance firms are the primary adopters of AI in the audit. Their goal is to maximize efficiency, minimize the costs of audit work, and enhance the coverage of audit procedures. Specifically, these procedures require two functions:

- Analyze contract documents—leases, rental agreements, etc.—for pre-identified keywords, such as key clauses, dates, persons, and relevant terms.
- Present potential anomalies for further human investigation.

Because these documents may be several thousand pages long, they are often reviewed on a sample basis due to limitations associated with manual labor.

However, AI allows document analysis at a fraction of the time cost. In some cases, it reduces the time cost by more than 90 percent. Furthermore, the quality of risk assessment is also vastly improved.

When detecting anomalies, AI produces a risk score using general ledger entries with financial features to meet compliance and assurance parameters. Some of these features are:

- Materiality levels
- All urgent payments
- Unbalanced debits and credits
- Rare flows
- Cash to bad-debt conversions
- All payments that went through multiple adjustments or reversals
- Journal entries beyond a threshold
- Open invoices beyond a period
- Sudden spikes in otherwise dormant vendors
- High-value transactions for a historically low-value vendor
- Duplicate entries
- End of the year or end of the period procedures
- Uncleared bank reconciliation entries
- Multiple changes to the bank account information of a vendor.

AI processes allow auditors to extract and load accounting and finance data directly from financial management information systems (FMIS) or underlying enterprise resource planning (ERP) systems. Human auditors use a dashboard to visualize the risk scores and investigate anomalies externally. Auditors can flag data and trigger ML algorithms to refine scores. In minutes, AI can do work that will otherwise cost several auditors for many weeks. Some tools are compliant with international audit standards like SAS 99, CAS 240, and ISA 240, such as the MindBridge AI Auditor, an application developed by a private Canadian firm. The UK and Canadian federal governments are testing the tool for wider applicability and adoption (MindBridge).

3

>>>

AI Risks

For all the potential benefits, there are also significant potential risks that will need to be mitigated for the adoption of AI as part of a government's digital transformation. The risks and their mitigating measures discussed here are primarily at the project-level, while policy-level ethical issues for society at large are discussed in Chapter 5.

Performance, Trust, and Bias

Negative bias is an inherent problem in AI that arises as a result of many factors, including incomplete, inaccurate, or corrupt data (statistical bias) which cause a predictive outcome that is in favor of or against one or more groups of people. There are well-known cases of how harmful such negative bias can be leading for example to unfair access to public services such as housing and social benefits or unfair incarceration. For example, analysis of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software used by U.S. courts and police to forecast which criminals are most likely to re-offend found it was biased against African Americans. The COMPAS algorithm provided information to police and judges to make decisions on defendants and convicts, for example setting bail amounts and sentences. The analysis found that the software was twice as likely to falsely label black defendants as future criminals than white defendants.[9]

9. Venkateswaran 2020

Some bias is inherent in AI models because data are finite, even when made available at scale. AI systems need to continuously be refined and improved AI systems as datasets and tools evolve or weaknesses emerge. Even with considerable preparation, sources of bias can be difficult to identify preemptively. As a result, AI results can be deceptively rational, even when biased (Ntoutsi et al. 2020). Sometimes, the AI team of developers or data scientists carries some inherent bias (cognitive bias), which should also be carefully monitored. Also, AI firms voluntarily manipulate data and algorithms to maximize profits (economic bias), which should also be addressed through policy action and public scrutiny.

To manage the risks of bias and the impact on access to services, a policy framework needs to address these issues. The full disclosure of the datasets and algorithms used in AI is the key to managing bias. Data and algorithm disclosure can aid in building trust and also aids the production, collection, and engineering of "good" data, which is defined as follows:

- Good data are available in abundance. The more data, the better.
- Good data have explainable features that relate to the problem statement. Raw unprocessed data contains simple, human-readable values.
- Good data are extensible. In other words, new features (data points or parameters to the layman) can be added to each record as models evolve. Feature engineering is possible with good data, which involves using existing features to derive additional information about each record of data (see Annex A).
- Good data are normally distributed. A normally distributed sample of the population is easily derived using random selection methods during training and testing. The values of data are not random; however, the selected members of a broader population are random.
- Good data are complete. Data are not missing key features that are critical to the problem statement.
- Good data can be traced back to the origins. Data can obfuscate sensitive personal information about people. Data come from official government sources.

Nonetheless, bias can emerge throughout the AI project life cycle, often unconsciously, through selective data gathering, requiring additional policies to oversee data selection processes. For example, data scientists may choose to collect data from groups that are perceived to be relevant, but these groups may be selected as a matter of personal preference. This is a classical polling technique that yields favorable results from a population-based selection of data around information on gender, race, ethnic origin, zip code, color, and disability.

Bias is best mitigated by policies and processes that ensure inclusion, conscientious oversight, transparency, disclosure, and contestability. Where models may influence public policy or mission-critical outcomes, the publication of data collection criteria as well as the release of open-source code for the implemented frameworks may mitigate the risk of producing nefarious outcomes. Even more so, the democratization of data and policymaking can improve the practical outcomes of AI frameworks and enhance trust in AI infrastructure in government.

Additionally, governments should develop competing AI systems that focus on the same problem statement. By employing multiple solutions on one problem statement, a practice adopted in Singapore and Israel, governments can significantly improve the likelihood of a positive outcome. Two systems with varying degrees of bias help reduce the likelihood of unintended outcomes by converging on results in different ways. Also, AI systems could be developed to identify bias – use the same tool to fight bias that caused the bias in the first place.

Human oversight could provide an additional safeguard against machine-invoked bias. Introducing human oversight can help detect skewed results from influences such as training data manipulation, forgery, and intentional bias.

9. Venkateswaran 2020

Implementing agencies could develop risk mitigation frameworks. Many governments have already developed model AI risk mitigation frameworks, which can be tailored to the local context. The Government of Canada developed an Algorithm Impact Assessment[10] for implementing agencies that consist of an online questionnaire and scoring scheme to assess the level of risk and mitigate the risk. The U.S. Department of Homeland Security has developed an AI risk assessment framework that is also useful in mitigating risks in AI performance. Key aspects of this framework are summarized below:

> > >

TABLE 3 - AI Risk Mitigation Framework

Types of Standard	How Applicable to AI	Where Standards Are Applied	How It Can Reduce AI Risk from an Adversary
Analytics and re-search	Standards that evaluate the quality of analysis and scrutability of algorithms	Back end: explainability and transparency	Identify faulty logic or reasoning, increase the difficulty of deceiving and/or manipulating analysis from AI Determine how much to trust system inputs and outputs
Legal and regulatory	Standards-based on govern-ance and regulatory over-sight into preserving privacy and consent	Front end: usability and personalization; back end: standardized architecture	Change understanding of liability for mistakes and enhance attribution Transform the notion of the jury of peers and evolve crime and punishment
Moral and ethical	Standards that prevent AI from performing actions that are contrary to a moral or ethical norm	Back end: fail-safes	Reduce the likelihood that AI will do the "wrong thing" (i.e., immoral or unethical behavior) if exploited or infiltrated by an adversary
Technical and indus-try	Standards to measure the performance of an algorithm on relevant tasks	Front end: performance	Meet appropriate tech-nical specifications (e.g., low number of false posi-tives) to be robust against adversary denial and deception activities
Data and Information security	Standards for the protection, sharing, or use of data relevant to a task	Front end: training; back end: data integrity and availability	Limiting access to and information about how an AI system works to appropriate people could help prevent exploitation by an adversary Preventing manipulation of training data

Source: Oxford Insights

10. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

Cybersecurity

Hacking poses a serious risk in AI systems. Forged data and bad actors can impair training algorithms to cause harm. One of the most common hacking techniques to exploit security vulnerabilities in AI is phishing.

Spear-phishing tactics include the practice of delivering malicious code or gaining unauthorized access through socially engineered messages. The best-known example of a general phishing attack is that of a digital hustler – a foreign prince offering unclaimed money in a foreign bank account in exchange for a small cash advance or a bank account number. In this example, broad stroke methods of AI message creation leverage socially desirable outcomes, which AI in spam filters have become adept at detecting. Propagation of malicious code that spreads itself across systems, networks, and even 'networkless' devices offers exponential reach in offensive cyber operations. The most notable example is the infiltration of a country's uranium enrichment program called Stuxnet, where a targeted propagation attack led a centrifuge to spoil its payload. Another is NotPetya, which relied on password theft and caused over $10 billion in damages across hundreds of thousands of computers in more than 100 countries. NotPetya was later repurposed by the National Security Agency of the U.S. Department of Defense to rip through targeted networks in seconds or minutes, making it one of the fastest-spreading pieces of malicious code in history. The adage goes, "by the second you saw it, your data center was already gone." NotPetya did not leverage even the slightest bit of AI.

Solid governance practices help mitigate the risks by imposing explainability, transparency, and validation in AI systems, in addition to the security best practices at the technical level. Governments can prevent adversarial attacks on data sources and computing resources with the use of security best practices, such as access-control lists (ACLs) and API tokens for inter-process communication (IPC) and human-facing endpoints. These practices are standard rules among corporations. Government systems are no exception to these rules of practice.

Prevent common patterns that kill critical processes. Proactive cybersecurity operations conceptualize the kill chain—the sequence of steps that hackers cycle through to achieve nefarious goals. Both hackers and defenders have a vested interest in finding vulnerabilities in AI systems; the former to exploit, the latter to remediate. AI is useful in vulnerability discovery.

Mitigate zero-day exploits—those with no patches—that are the targets of cyberattacks. Cybersecurity and AI teams also use tools known as fuzzers to discover errors and security loopholes by inputting massive amounts of data (called fuzz) to the system in an attempt to make it crash.

The implementing teams should also ensure back up data with redundant systems and enforce no single point of failure (SPOF). Wiping attacks that erase or overwrite otherwise benevolent files on computing systems are difficult to detect because their effect is known only after they propagate and execute. However, through the use of learning-enabled AI, engineers can develop defenses against these types of propagation attacks, though no known examples of such AI systems exist in the public domain. Obfuscation and anti-forensics employ methods of detection avoidance. AI can be quite beneficial in detecting obfuscation attacks as well as creating them. Destructive attacks are unlikely candidates for AI prevention.

AI holds great promise in cybersecurity defense. However, given the fact that destructive propagation attacks can proliferate and remain dormant for months, even years, the detection of these attacks may be limited in scope. Still, the effort to detect security breaches remains a key focus of AI systems in cybersecurity.

It stands to reason that if AI can learn to detect threats, it can also alter them to further delay their effects if not block them altogether. Furthermore, attribution mechanisms that detect external sources of threats through AI clustering techniques are proving themselves in identifying sources for threats in geographic regions. In some cases, NLP can detect grammatical nuances in source codes that allow defenders to home in on geographical regions for further investigation.

All told, the many methods of subterfuge and espionage employed by hackers and defenders are writhe with theory and are unclear to the general public. Sometimes, tools designed at the hands of government defense departments are responsible for the greatest defenses and offenses. It is in the domain of a government's responsibility to determine the degree of impact that its defensive strategies have on the safety of the general population.

Remain proactive. Not all offenses are the source of political cyberwarfare. Many still emerge from obscure corners of the internet, to "prove" that the vulnerabilities of government and commercial organizations are real. Although they are effective in advancing the evolution of cybersecurity best practices, they are more often than not isolated incidents that fall under the jurisdictions of international authorities and garner stern responses from enforcement officers and legislators.

Control

Because many AI systems operate autonomously and interact behind the scenes with one another using IPC, machine-centric feedback loops can cause unintended consequences. In 2010, stock exchanges that allow high-frequency trading experienced a flash crash caused by AI algorithms that went awry in competition with one another. This led to unintended artificial financial market inflation. Moreover, chatbots interacting with one another can create their language that humans cannot understand.

Proactive control, monitoring, testing, and validation are necessary to control the outcomes of rogue AI systems and prevent edge cases in software development from getting the best of humanity, if only on a rare occasion.

Privacy

The use of data fabrics and Big Data, growing reliance on automation and decision-making, and the gradual reduction of human involvement in human processes raise concerns about fairness, responsibility, and respect for human rights. Moreover, AI data policy raises concerns for privacy and individual identity. Group and community-driven AI has the potential to increase the risk of harm by what Carl Jung describes as the collective unconscious of humanity, a shadowy force or dark side of personality that collectively propels human digressions at a macro level. AI is no exception.

Protect privacy and human identity. Yet, despite all the foreboding ethical predictions, ethical influences begin with the protection of individual identity within large-scale datasets, access control, and policies. This prevents the arbitrary exploitation of identity recognition systems. In the United States, municipalities are enforcing policies to ban facial recognition technology altogether. The use of AI in some countries to detect fever from facial recognition software in cameras installed at public places carried the risk of human surveillance and infringement of privacy. In Singapore, the GovTech agency and the Ministry of Health (MoH) have co-developed an app, TraceTogether, that can trace individuals without infringing on privacy. Citizens download the app, turn on the Bluetooth, and allow push notifications and location services. The app can exchange signals in a short distance of 2-5 meters with other app users, exchange anonymized identifications (IDs), and store anonymized data locally of all the persons in the proximity of the app users. If the user allows on the app, the MoH will contact the user by sending a code. MoH will then be able to decrypt the random IDs of individuals with whom the user came into contact. The authorities comply with the privacy and data protection laws, as no personal details are collected except the phone number.

Furthermore, policies can enforce limitations on group inference models that lead to individual discrimination. For example, organizations are choosing to obfuscate individual identity to mitigate against the risk of fraud due to unauthorized access to data. Rather than use names and ID numbers, data systems are using salted cryptographic hash functions to “scramble” identifiable information. Because the use of a salted hash function is idempotent—it always yields the same result for a given input—systems can protect exploitable data and retain uniqueness for algorithmic purposes.

Privacy legislation and regulatory framework provide a solid legal basis for mitigation privacy risks. Governance frameworks that promote self-assessment, peer review, and public inclusion could strengthen compliance with these legal frameworks. The details could be adopted based on the context and existing mechanisms of transparency, citizen engagement, and accountability. However, the value of public inclusion is critical in this process.

4

>>>

AI Governance and Operations

Most advanced digital governments have issued governance frameworks, including ethical principles for the use of AI. An overview of these governance models is presented in this chapter, which discusses three aspects of governance models: ethical principles, the role of a central agency, and operational framework.

AI Ethical Principles

The risk mitigation for AI requires the adoption of some ethical principles and several of the key ethical considerations. Several advanced digital economies are adopting AI governance models and policies developed by an interagency team of policymakers and AI experts, this chapter summarizes those principles, identifies good practices for the institutional design for adopting AI in the public sector, and shares innovative procurement practices for acquiring AI implementation services. The models of AI governance typically include bias, privacy, algorithm opacity, limited data access, security, citizen consent, and inadequate supervision. National governments, including Australia, Canada, China, Japan, Singapore, United Arab Emirates, and the United States as well as international organizations including the European Commission (EC), the Institute of Electrical and Electronics Engineers (IEEE), International Organization for

Standardization (ISO), UN, and World Economic Forum, are actively proposing governance models for AI that emphasize common principles:

- **PRIVACY AND DATA PROTECTION.** AI solutions should respect an individual's right to privacy and civil liberties. Individuals should have control over their data. Individual consent is necessary for using and re-distributing their data. They should have the right to restrict the processing of their data, rectification, and erasure.

- **ACCOUNTABILITY.** Mechanisms must ensure accountable behavior during the life cycle of AI design and implementation. Impact assessment frameworks should be done to identify accountability at every step of the process. An agency or body should be responsible for monitoring accountability.

- **SAFETY AND SECURITY.** Cybersecurity is critical. AI solutions should have predictable behavior. Leaders must ensure the well-being of society at large and private individual humans.

- **TRANSPARENCY AND EXPLAINABILITY.** The algorithm, business case, data collection, design, and policy information must be transparent to stakeholders and those impacted. Open-source data algorithms could enhance transparency. Individuals should get notifications when interacting with AI or when AI decides for him or her. There should be regular Reporting requirements on transparency. The rights of citizens to information are important. Data should be of high quality and representative.

- **FAIRNESS.** AI solutions should minimize bias and identify and manage risk. Inclusiveness should be ensured in design and impact.

- **HUMAN CONTROL OF TECHNOLOGY.** The AI should be under human control. People should review automated decisions. Individuals should be allowed to opt-out of automated decisions.

- **PROFESSIONAL RESPONSIBILITY.** Multistakeholder collaboration, accuracy, and scientific integrity of the solution should be ensured.

- **PROMOTION OF HUMAN VALUES.** AI should be human-centric. It should promote human values and benefit society.

Some governance models and guidelines emphasize common program and project management practices like cost-benefit analysis, legal and regulatory compliance, risk management, flexibility, and the use of an agile approach.

Ensuring compliance with these principles would require a careful balance between oversight and agility.[11]

These principles are given a different level of emphasis in different settings. The Berkman Klein Center for Internet Society at Harvard University tracks and maps the global consensus on ethical principles for AI. Figure 10 is adapted from their work which shows the global adoption of these principles and the level of emphasis of each principle. Despite different levels of emphasis on different principles, there is a consensus that ultimate control of AI must remain with people. AI must not be a regulatory means unto itself.

11. To enforce policies, the European Union (EU) is considering establishing a standards body, similar in composition to the U.S. Food and Drug Administration, to assess the impact of algorithmic processes before release. There is a key problem here since algorithmic innovation occurs at such a speed that it outpaces the government's ability to evaluate every potential outcome. The agency may even become a bottleneck that developers simply bypass due to capital constraints. Instead, some propose that such validation should be part of a certification process that is executed through peer review.

> > >

FIGURE 9 - Global Consensus on the Principles of AI

Source: Fjeld et al. (2020).

Country Examples of AI Governance Systems

The **Australian** Government Department of Industry Innovation and Science funded research into the ethical principles of AI usage in government in 2018 and published a white paper on it in 2019. The core principles in its AI governance framework are:[12]

1. **GENERATES NET-BENEFITS.** The AI system must generate benefits for people that are greater than the costs.

2. **NOT HARM.** Civilian AI systems must not be designed to harm or deceive people and should be implemented in ways that minimize any negative outcomes.

3. **REGULATORY AND LEGAL COMPLIANCE.** The AI system must comply with all relevant international and Australian local, state, territory, and federal government obligations, regulations, and laws.

4. **PRIVACY PROTECTION.** Any system, including AI systems, must ensure people's private data are protected and kept confidential plus prevent data breaches which could cause reputational, psychological, financial, professional, or other types of harm.

5. **FAIRNESS.** The development or use of the AI system must not result in unfair discrimination against individuals, communities, or groups. This requires particular attention to ensure the "training data" is free from bias or characteristics which may cause the algorithm to behave unfairly.

6. **TRANSPARENCY AND EXPLAINABILITY.** People must be informed when an algorithm is being used that impacts them and they should be provided with information about what information the algorithm uses to make decisions.

7. **CONTESTABILITY.** When an algorithm impacts a person there must be an efficient process to allow that person to challenge the use or output of the algorithm.

8. **ACCOUNTABILITY.** People and organizations responsible for the creation and implementation of AI algorithms should be identifiable and accountable for the impacts of that algorithm, even if the impacts are unintended.

The Canadian government's 2019 Directive on Automated Decision-Making[13] guiding principles for the ethical application of AI governance are:

- Understand and measure the impact of using AI by developing and sharing tools and approaches.
- Be transparent about how and when to use AI, starting with a clear user need and public benefit.
- Provide meaningful explanations about AI decision making, while also offering opportunities to review results and challenge these decisions.
- Be as open as possible by sharing source code, training data, and other relevant information, all while protecting personal information, system integration, and national security and defense.
-
- Provide sufficient training so that government employees developing and using AI solutions have the responsible design, function, and implementation skills needed to make AI-based public services better.

Furthermore, the Canadian government formulated a comprehensive analysis and exposition of the key government processes in play across the entire government. The document includes objectives and expected results, definitions, and rules for semi-annual re-evaluation, which is crucial in light of the rapid pace of AI development. The government also developed an Algorithm Impact Assessment (AIA), which is a questionnaire designed to assist agencies in assessing and mitigating their risks.[14]

> > >

BOX 1 - Actionable Insight: Adopt Principles of AI and Issue an AI Governance Model

The central digital agency should adopt the common principles of ethical AI and prepare a governance model. The model should formulate operational arrangements, including an innovation hub, data governance, data standards, collaboration with the private sector, skills development, adoption in the public sector, and partnership with nonprofit and academia to promote AI research, among others.

12. https://www.industry.gov.au/news-media/towards-an-artificial-intelligence-ethics-framework.
13. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.
14. For more information, visit the "Responsible Use of Artificial Intelligence (AI)" on the website of the Government of Canada at https://www.canada.ca/en/government/system/digital-government/modern-emerging-technologies/responsible-use-ai.html#toc1.

Singapore

Most of the themes discussed have been incorporated into the Model Governance Framework, issued by the Government of **Singapore** (PDPC 2020).[15] Singapore maintains an active leading role in the strategic development of integrated government AI systems around the world. Singapore is actively investing in AI policy and process standards among partner nations to support global AI development in trade and commerce. Transparency affords opportunities for the successful development of systems impacting key strategic international partners. The willful commitment to long-term execution provides a global foundation that extends far beyond Singapore's borders. Singapore is also experimenting with policy enforcement using AI-powered robotics and contact tracing since the start of the COVID-19 pandemic. The government's governance model is driven by two fundamental guiding principles:

- **EXPLAINABLE, TRANSPARENT, AND FAIR PROCESS.** The organizations using AI should ensure the decision-making process is explainable, transparent, and fair.

- **HUMAN-CENTRIC AI. AI SOLUTIONS ARE HUMAN-CENTRIC.** AI helps amplify human capabilities and protects human interests.

The model advocates that organizations should embrace four key measures in their quest for AI adoption:

1. **INTERNAL GOVERNANCE STRUCTURE.** The involvement of top officials and their sponsorship of AI initiatives is critical. This ensures ethical considerations are introduced in the decision-making process and monitored regularly at the highest levels.

2. **DETERMINING THE LEVEL OF HUMAN INVOLVEMENT IN AI-AUGMENTED DECISION-MAKING.** AI algorithms can support processes with or without the involvement of humans. Any process that affects human beings must involve humans "in-the-loop."

3. **OPERATIONAL MANAGEMENT.** This aspect includes data management, talent, skills, and procurement. Organizations must ensure data governance arrangements are in place to ensure integrity, consistency, transparency, security, interoperability, and accountability for data. Also, organizations must strive to incorporate relevant talent through proactive partnerships between academia, private sector firms, and start-ups, which are fundamental pipelines leading to the success of AI initiatives. Procurement must provide flexible experimentation, produce proofs of concept over multiple iterations, and scale up with an acceptable risk of failure.

4. **STAKEHOLDER INTERACTION AND COMMUNICATION.** Strategies must ensure consistent and transparent communication with the key stakeholders and manage relationships with them. In the public sector, public scrutiny and transparency are critical aspects of AI initiatives.

The **U.K.** Office for Artificial Intelligence is responsible for overseeing the implementation of A I and has produced several Reports.[16] The agency is a joint effort of the Department for Business, Energy, and Industrial Strategy and the Department for Digital, Culture, Media, and Sport.

AI and the Multilaterals

The **EC** has formed a high-level expert group to prepare the ethics guidelines which were circulated for comments, testing, and assessment in 2019 and being vetted by many organizations (EC 2019). The EC envisions developing an AI ecosystem that brings benefits to citizens and businesses for improved service delivery, promotes new products and services, and emphasizes sustainability while ensuring safeguards, rights, and freedoms. EC is promoting a common European approach to reach scale and avoid fragmentation of the single market. According to these guidelines, trustworthy AI should be:

- Lawful, comply with all applicable laws and regulations.
- Ethical, respect ethical principles and values.
- Robust, both from the technical and social perspective.

The guidelines put forward seven key requirements that AI systems should meet. They incorporate the ethical principles promoted by the EC and include human agency and oversight, privacy and data governance, transparency, diversity, nondiscrimination and fairness, societal and environmental well-being, and accountability.

In 2019, the **United Nations** (UN) launched its Centre on Artificial Intelligence and Robotics, under the UN Interregional Crime and Justice Research Institute (UNICRI), to monitor developments in AI and robotics, with the support of the gov-

15. https://www.pdpc.gov.sg/Help-and-Resources/2020/01/Model-AI-Governance-Framework.
16. Understanding artificial intelligence—GOV.UK. This is an introduction to using AI in the public sector. The Data Ethics Framework. *A Guide to Using AI in the Public Sector* enables public bodies to adopt AI systems in a way that works for everyone in society (GDS and OAI 2019). Guidelines for AI procurement—GOV.UK. These new procurement guidelines will inform and empower buyers in the public sector, helping them to evaluate suppliers, then confidently and responsibly procure AI technologies for the benefit of citizens.

ernment of the Netherlands. The center, based in The Hague, helps focus expertise on AI throughout the UN in a single agency. UNICRI initiated its program on AI and robotics in 2015. One of the leading agencies of the UN, the United Nations Educational, Scientific, and Cultural Organization (UNESCO) recently appointed an international expert group to draft internationally applicable global recommendations on the ethics of AI (UNESCO 2020). This action follows the decision by UNESCO's 193 member states during its last General Conference in November 2019.

> > >

BOX 2 - Arivate Sector AI Principles

There is broad convergence on the adoption of AI principles in the public and the private sector. Several private organizations have adopted principles to enhance trust and transparency in the process of developing AI applications:

- IBM's principles of trust and transparency state that AI should augment human intelligence rather than replace it, trust is key to adoption, and data policies should be transparent (Dignan 2017).

- Google's principles on AI state that AI should protect the privacy of citizens and be socially beneficial, be fair, be safe, and accountable to people.

- The Asilomar AI Principles were outlined at the 2017 Conference on Beneficial AI organized by the Future of Life Institute and cover research, ethics, and values in AI. The 23 principles have been adopted and signed by 1,273 researchers and 2,541 other interested parties, including Elon Musk and the late Stephen Hawking.

- Organizations interested in joining the Partnership on AI must endeavor to uphold eight tenets and support the Partnership's purpose. They include calls for an open and collaborative environment to discuss AI best practices, social responsibility on the part of companies delivering AI, explainability, and a culture of trust, cooperation, and openness among scientists and engineers.

- The AI4PEOPLE principles and recommendations are concrete recommendations for European policymakers to facilitate the advance of AI in Europe (Floridi et al. 2018).

- The World Economic Forum's five principles for ethical AI cover the purpose of AI, its fairness and intelligibility, data protection, the right for all to exploit AI for their well-being, and the opposition to autonomous weapons (O'Brien et al. 2020).

- The IEEE's set of principles place AI within a human rights framework with references to well-being, accountability, corporate responsibility, value by design, and ethical AI (IEEE 2019, 17–35).

The Institute for Ethical AI & Machine Learning adopted eight principles of responsible ML development to provide a practical framework to support technologists when designing, developing, or maintaining systems that learn from data.[17]

17. https://ethical.institute/principles.html.

Role of a Central Government Agency or AI Hub

The use of AI in many advanced digital governments is seen as a broader effort for the citizen-centric digital transformation of public services. A central coordinating agency is typically established and responsible for issuing the ethical principles and guidelines for trustworthy AI. It develops government-wide data strategies and policies to harness the power of AI. This is the essential first step in making advances in this domain to ensure commitment, governance, line-of-sight, and monitoring for the acceptable use of AI in the public sector. The AI policy should address key policy domains: research, talent, entrepreneur ecosystem, ethical standards, data access, AI in government, AI in sectors, and governance capabilities (World Bank 2020).

> > >

TABLE 4 - The Role of a Central Agency in AI

Country	Agency or Program	Role
Canada	CIFAR (formerly the Canadi-an Institute for Advanced Research)	Leads the strategy in close partnership with the Canadian government and three new AI institutes: the Alberta Machine Intelligence Institute in Edmonton, the Vector Institute in Toronto, and Mila in Montreal. It is primarily a research and talent promoting institute, while the implementation of AI in the government is decentralized.
Finland	Aurora AI National AI Program	The program seeks to provide a holistic set of personalized AI-driven government services for citizens and businesses in a way that is human-centric and works toward their well-being as its ultimate goal, instead of being driven by the needs of the public authorities.
	Finnish Center for AI	A joint partnership by Aalto and Helsinki Uni-versities to promote AI research, talent, and industry collaboration. It also supports an AI accelerator pilot program and the integration of AI in the public service.
France	Joint Center of Excellence for AI	State-level agency to help recruit AI talent and to serve as an advisor and lab for public policy design.
	Inter-ministerial coordina-tor	The coordinator's role is to implement France's AI strategy, including public sector AI transformation efforts, and serving as an interface between the public and private sectors.
Germany	German Research Center for AI	A major actor in this pursuit and provides funding for application-oriented research.
	Plattform Lernende Sys-teme	Brings together experts from science, indus-try, politics, and civic organizations to devel-op practical recommendations for the gov-ernment.
India	National Institution for Transforming India—Aayog program	Aayog adopted a three-pronged approach: (a) undertaking exploratory proof-of-concept AI projects in various areas; (b) crafting a na-tional strategy for building a vibrant AI eco-system in India; and (c) collaborating with various experts and stakeholders.
Saudi Arabia	Saudi Data and Artificial In-telligence Authority	Strategy approved in 2019 provides a core mandate to drive and own the national data and AI agenda to help achieve the govern-ment's Vision 2030's goals. To fulfill this mandate, the Authority and its sub-entities—National Information Center, National Data Management Office, and National Center for AI—will deliver on the promise to create a data-driven and AI-supported government.
Singapore	Digital Government Office	One of the leading agencies on AI, which also brings together research institutions and the private sector.
United States	The White House	Issued a memorandum to the agencies providing guidance on ethical principles and operating framework.
	U.S. Commerce Depart-ment's National Technical Information Service	Delivers a fed-to-fed framework for data sci-ence innovation through partnerships with industry, universities, and nonprofits at the velocity of the government's needs.

Source: World Bank.

The institutional arrangements for the implementation of AI could be centralized or decentralized. In several jurisdictions such as Canada and the USA, implementation of AI is delegated to the agency level, while the central agency issues the AI ethical principles, AI data strategy, and operating framework. The central agency may partner with the private sector and academia to bring in talent and do research. Table 1 presents an overview of the role of a central agency in several countries. Under centralized arrangements, governments create a hub within a central digitization agency to implement the AI strategy. The central hub pools scarce talent, partners with the line agencies, provides an AI lab, and develops alliances with academia, the private sector, and start-ups. Governments typically view themselves not only as the service providers for citizens and businesses but also as an orchestrator of public services through expanding public-private partnerships. This model is adopted by many economies such as Austria, Estonia, Israel, Saudi Arabia, Singapore, United Arab Emirates, and the United Kingdom.

In the United States, AI is both centralized under the federal government and decentralized among state governments. Centralization is enabled through the National Technical Information Service (NTIS) under the U.S. Commerce Department and the Federal Risk and Authorization Management Program (FedRAMP). The former is responsible for helping federal agencies rapidly analyze, manage, and implement scalable data solutions by leveraging an extensive NTIS network of technical talent from private industry, which is often difficult to locate in today's competitive information technology landscape. FedRAMP's mission is to promote the adoption of secure cloud services across the federal government by providing a standardized approach to security and risk assessment.

The central agency encourages and promotes agency-, ministry-, and department-level initiatives. U.S. agencies such as the IRS, Treasury, and General Services Administration (GSA), have their centers of excellence focused on agency-specific AI solutions. The National Security Commission on AI Strategy focuses on defense, security, and war. Regardless, state and municipal levels aggressively pursue independent AI initiatives, primarily for land management, tax revenue management, and fraud detection.

The Canadian government tapped CIFAR (formerly the Canadian Institute for Advanced Research), a global research organization based in Canada, to lead the development of its Pan-Canadian AI Strategy. CIFAR is focused on ethics, research, and talent promotion, while implementation is done at the government agency level.

AI Operations Framework

The central agency responsible for leading the AI initiatives generally provides an operating framework. It guides agencies and departments through steps for operationalizing: defining the idea with a problem statement; conceptualizing the problem with experts; proposing a solution to the problem; developing a proof of concept; and implementing this idea through iterative stages. The framework focuses on integrating AI into operations to produce efficiencies, enhance the quality or augment data-driven policy capabilities. It also accounts for ways in which the solution will augment human decision-making capabilities by increasing the breadth of data beyond human comprehension. A key example is using NLP to analyze millions of policy documents from citizen sources and public records. The operating framework typically guides key implementation steps. Governments may customize the framework contextually, but overall, it could include six components as presented below in Table 5).

> > >

TABLE 5 - Operating Framework

Component	Description
Ideate	The problem statement is produced in detail. The statement is agnostic to technology.
Conceptualize	The project manager coordinates discussions between small and medi-um enterprises and AI experts.
Propose	A detailed proposal is prepared. It contains the problem statement, po-tential solution options, and a checklist with a brief description of each to ensure alignment with legal, policy, and ethics risks, mitigation ac-tion, and expected results. A separate section on data sources is critical. Management approves.
Develop a prototype	The project manager ensures technology teams work together with Small and Medium Enterprises (SMEs) seamlessly to develop a proof of concept. A prototype visualizes the solution with or without code.
Test	SMEs and technical teams test the system.
Develop and deploy	The system is developed full scale, tested again, and deployed for oper-ational use. It is also integrated with the environment.

Source: World Bank.

The implementation steps are summarized as follows:

- **IDEATE.** The detailed problem statement involves leveraging subject matter experts. The problem statement should be technology agnostic. It captures sufficient detail and contextualizes the overall strategy and vision to maintain a clear line-of-sight.

- **CONCEPTUALIZE.** Domain, subject, and technology experts enter into discussion and conceptualize technical components to the problem statement. These experts are either from the center of excellence in the government or the private sector. The output of this stage is a conceptual Report that details how the solution will address the problem statement.

- **PROPOSE.** In this stage, the team formulates a proposal for the implementation. Typically, implementation partners are private sector firms, including start-ups, nongovernmental organizations (NGOs), legislative, and human rights experts with experience and knowledge of these solutions. The procurement framework engages these firms with flexibility; without detailed specifications, but rather based on problem statements and a high-level solution concept, amenable to change based on market response.

- **DEVELOP A PROTOTYPE.** The team selects an implementation partner and requests a working proof of concept. This software demonstrates how the solution will work as a vertical, without pursuing full-scale production deployment, customization, and data migration.

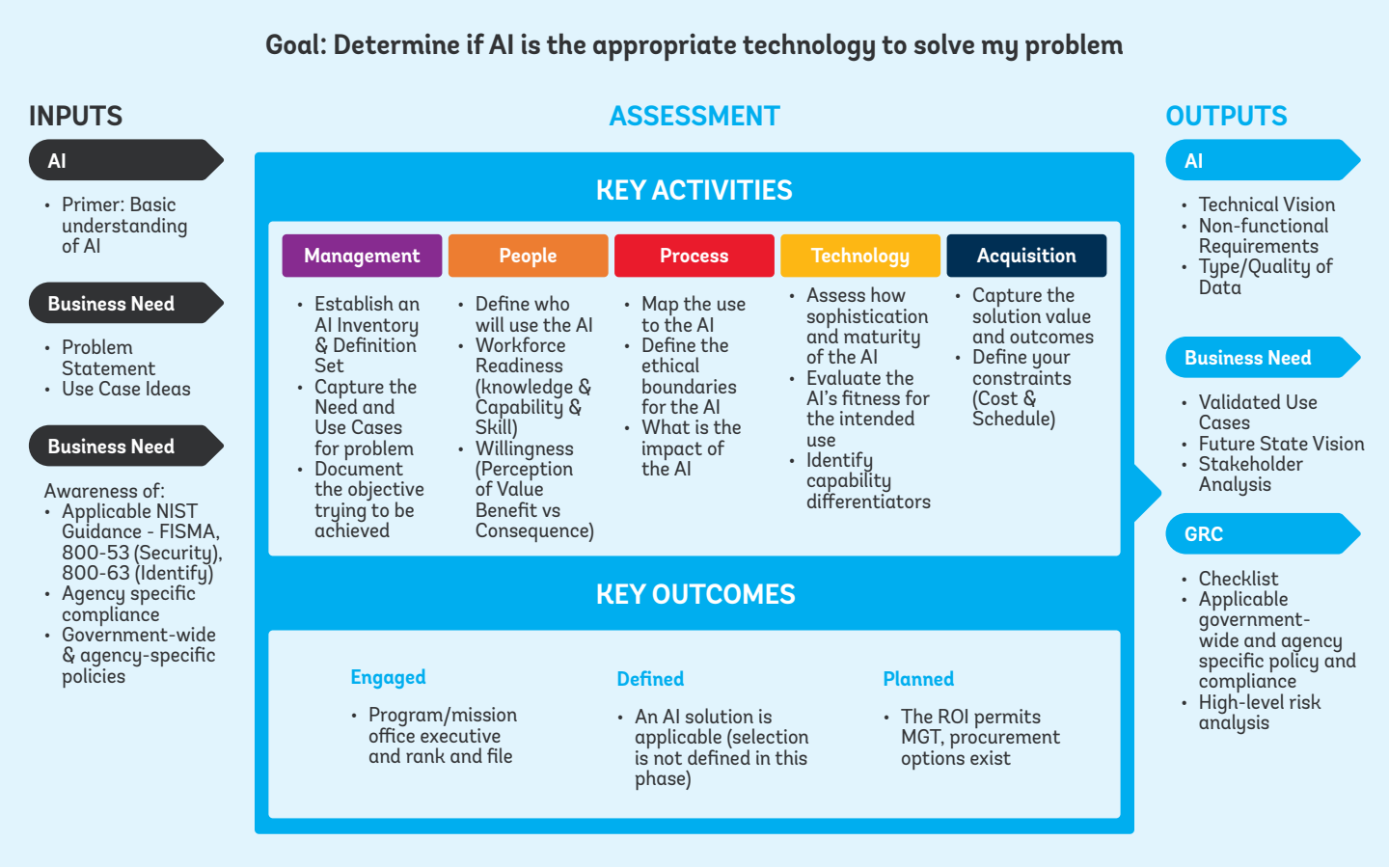
- **TEST, DEVELOP, AND DEPLOY.** Proof of concept typically goes through several iterations, leading up to implementation, based on working feedback from the subject matter and domain experts participating since the early planning stages. Upon maturation, the solution is ready for go-live production as a pilot capable of scaling horizontally. An important operational issue in procuring AI is experimenting at the big-data scale. Traditional approaches to linear solution silos require detailed specifications that interfere with AI innovation, which involves many iterations, much experimentation, optimization, and iterative learning from performance tuning based on unprecedented results due to the immense scale of AI modeling beyond the scope of human capabilities.

Justification at the Conceptualization Stage

AI is not the solution to every problem. How should an organization evaluate the scope and needs of a problem statement to determine whether AI fits the playbill or is little more than a theater act? The American Council for Technology and Industry Advisory Council (ACT-IAC) AI playbook for the U.S. government offers a questionnaire for assessing the necessity and fitness of AI solutions. Figure 11 illustrates the full scope of the playbook consisting of five phases. “Phase 1, Problem Assessment” stipulates that a government must “[d]develop a vision and business objectives through various assessments to ensure the AI solution addresses a specific use case and delivers results that optimize services and operational delivery” (ACT-IAC 2020). In more detail, the inputs and outputs of this assessment are shown in Figure 11.

> > >

FIGURE 10 - AI Business Case Assessment



Source: ACT-IAC (2020).

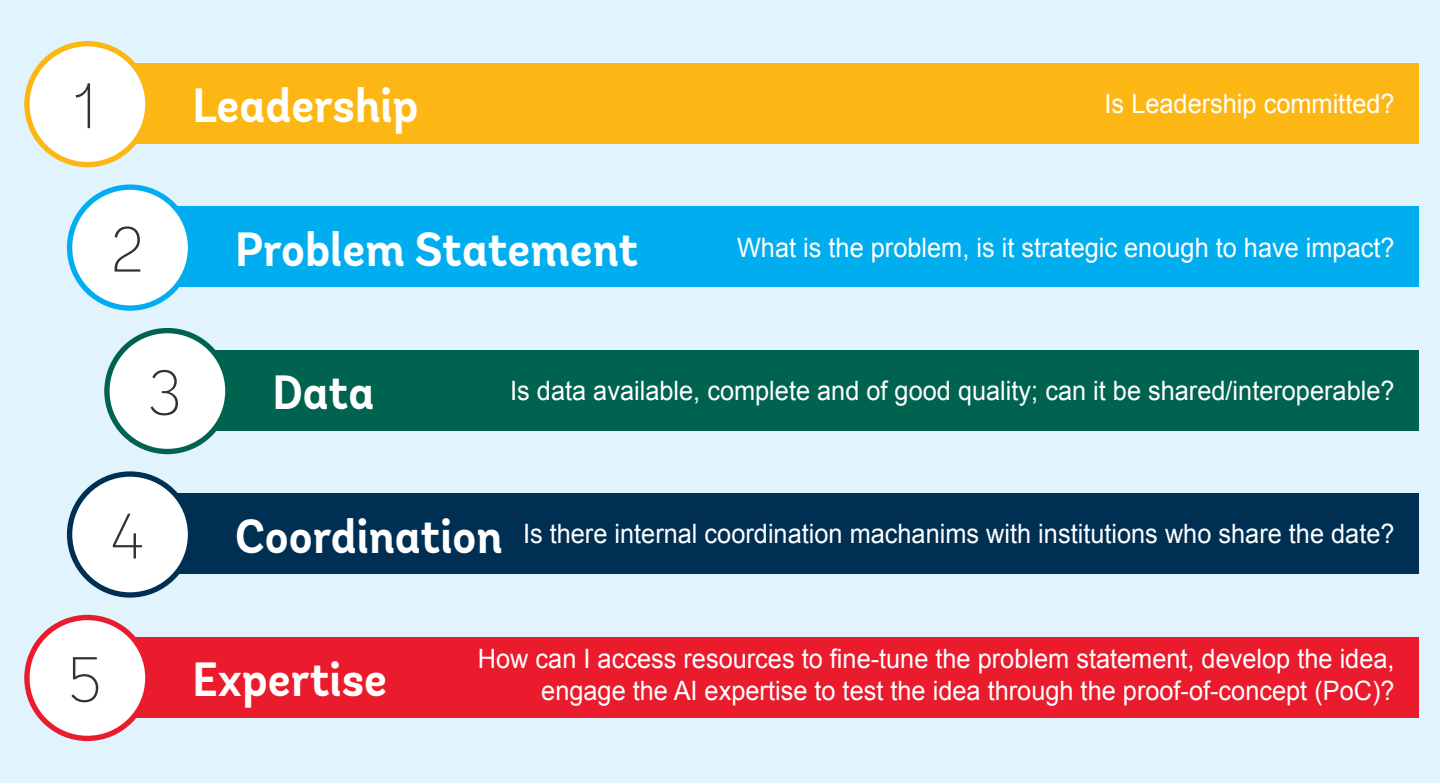
On a granular level, a 14-point questionnaire accompanies the assessment phase, which asks questions of stakeholders and key decision-makers. Answers fall on a scale of zero (not at all) to five (critical). A score of 18 or less indicates limited applicability and low return on investment; 19 to 40 indicates that AI could be applicable, but not without more in-depth analysis, and over 41 represents compelling applicability and significant benefits from a potential AI solution. The questions are:

- Does the use case clearly and accurately describe the problem to be solved?
- Does the use case accurately outline current processes in place?
- Does the use case align the goals and objectives with desired outcomes?
- Does the use case identify what data are required and available, accessible, and accurate?
- Does the use case need greater insight from the data?
- Has sufficient data been identified for the use case?
- Are the data from the use of case annotated and curated? (Does the data contain meta-information?)
- Does your use case largely need manual process automation? (That is to determine if only RPA [robotic process automation] is needed)
- Is there a predictive element to the use case? (Assumptions and testing made based on prior data)
- Have other technologies successfully been applied to address elements of the use case? (Could you somewhat solve your use case with an existing solution?)
- Does the data fit for purpose (descriptive modeling), and is it operationally relevant (predictive modeling)?
- Are the authoritative data sources of the use case organized, structured, deconflicted, and matriculated?
- Could the result of the use case change how conformance requirements need to be applied—for example, personally identifiable information (PII), classified, etc?
- Does the use case contain ethical considerations, and is there a potential for bias, for example in the data, algorithms, or aggregation process?

The implementation agency should assess the high-level governance conditions in Figure 12.

> > >

FIGURE 11 - Operationalizing AI



Source: The World Bank.

The operating framework should also address the issues of organizational roles and responsibilities. Entities implementing AI must identify key roles and responsibilities when designing the internal organization for managing AI, suitable to their context. At a minimum, these roles include:

- **EXECUTIVE SPONSOR.** Depending on the context, this role is the head of an agency, chief information officer, or department director. This role ensures compliance and alignment with the broader legal framework, policy objectives, strategies, and ethical considerations for AI. Also, this role develops coordination mechanisms with involved agencies.

- **WORKING GROUP.** Stakeholders from different departments whose data will be used, or who will be impacted by AI or have a stake in the solution, should be consulted at every step.

- **SUBJECT MATTER EXPERT.** Someone that understands the business process and its data, core nature of the qualitative objectives, and key results required for the successful implementation of an AI solution. This person does not need a background in AI to fulfill this role.

- **DEVELOPER, AI ENGINEER, AND DATA ARCHITECT.** An engineer with a mind for understanding the practical implementation of the AI infrastructure and engineering requirements. This person needs a background in AI software systems engineering.

- **DATA SCIENTIST.** A quantitative engineer that understands the data requirements for the project based on both qualitative and quantitative best practices that leverage statistical methods for assessing inbound and outbound data for bias and qualitative excellence. This person needs a background in AI modeling and should be a champion for data interoperability.

- **PROJECT MANAGER.** A project manager who manages teams, resources, results, and procurement in project planning at all stages of the project life cycle. He or she needs to be versed in AI systems engineering at a level of competency that will allow for the proper scoping of team objectives and key results. This person must also take the overall responsibility of aligning expectations from the subject matter expert, developer, and data scientist so the policy liaison can properly construct a policy plan that ensures on-time delivery and overall project integrity.

These roles and responsibilities can be tailored to the context, but essentially, they should cover the contextual areas of activity:

- Oversight of the various stages of AI planning, budgeting, design, development, legislation, and operations.
- Integration of roles and responsibilities defined by an internal risk management framework.
- Procedures for data governance, transparency, and disclosure.
- Policies for information governance, which enforce security, interoperability, and access control among stakeholders.
- Oversight of data science and AI modeling procedures that emphasize documentation and explainability to stakeholders.

Stages of Technical Solution Development

The following concepts are important components that the stages of AI implementation must address.

- **DATABASE COLLECTION.** Collected data must be cleaned and checked for bias.
- **SOFTWARE AND ALGORITHM DEVELOPMENT.** Multimodal data recognition must be implemented to reduce discrimination, bias, and unjust consequences. Algorithm transparency must disclose the steps taken to explain the results.
- **MODEL TRAINING AND EXCHANGE.** Standardization and consistency offer practitioners the opportunity to exchange trained models without revealing sensitive data, yet offering explainable disclosures for the practical purpose of understanding results.
- **TESTING AND VALIDATION.** Fairness and bias testing must be evaluated against standardized test sets created with oversight from representatives of affected populations and stakeholders.

Procurement

Most governments acquire expertise from the private sector through innovative procurement methods. The private sector, in particular start-ups, brings cutting-edge expertise to solve the complex public sector problems through AI. The implementation team must produce a broad overview of how they will customize procurement to these initiatives by using the procurement framework. Governments should consider adopting a set of guidelines and principles published by the World Economic Forum (see Table 6).

> > >

TABLE 6 - Innovative Procurement Guidelines

Guideline	Principles
1. 1. Prescribe a procurement process that defines the scope of problems and opportunities while allowing room for iteration.	**b.** Allow innovative procurement processes for AI systems development. **c.** Develop a clear focus with a specific problem statement. **d.** Avoid putting any energy toward defining the details of the solution. **e.** Support an iterative approach to product development.
2. Produce an RFP that publicly defines the benefits and costs associated with an AI solution while assessing risks.	**a.** Assess why AI is relevant to the problem. Be open to alternative technical solutions. **b.** Explain which public benefits are the main drivers in the decision-making process when assessing proposals. Consult with external experts if needed. **c.** Conduct an initial AI risk and impact assessment before starting the procurement process. Ensure that interim findings inform the RFP and revisit the initial assess-ment at key decision points.
3. Align procurement with relevant existing governmental strategies and contribute to their further improvement.	**a.** Consult relevant government AI initiatives on national, innovation, or industrial strategies. Review any guid-ance documents informing public policy about emerg-ing technologies. **b.** Collaborate with other relevant government bodies and institutions to share insights and knowledge.
4. Incorporate potentially relevant legislation, policies, and codes of practice in the RFP.	**a.** Conduct a review of relevant legislation, rights, admin-istrative rules, and other relevant norms that govern the types of data and kinds of applications in scope for the project. **b.** Consider the appropriate confidentiality, trade-secret protection, and data privacy best practices that may be relevant to AI systems deployment.
5. Articulate the technical and administrative feasibility of accessing relevant data.	**a.** Implement the proper data governance mechanisms at the start of the procurement process. **b.** Assess whether relevant data will be readily available for the project. **c.** Define data sharing policies for the vendor(s) during the procurement initiative and subsequent project.
6. Highlight the technical and ethical limitations of intended data uses to minimize issues with bias.	**a.** Consider the susceptibility of data and if the usage of the data is fair. **b.** Highlight known limitations (e.g., quality) of the data by consulting domain experts and require bidder(s) to describe strategies for addressing these shortcomings. **c.** Have a plan for addressing relevant limitations as they arise.
7. Work with a diverse, multidisciplinary team.	**a.** Develop ideas and make decisions throughout the procurement process in a multidisciplinary team. **b.** Require the successful bidder(s) to assemble a team with the right skillset and consult with the established domain experts.
8. Focus on mechanisms of algorithmic accountability and of transparency norms throughout the procurement process.	**a.** Promote a culture of accountability across AI-powered solutions. **b.** Ensure that AI decision-making is as transparent as possible. **c.** Explore mechanisms to enable the interpretability of the algorithms internally and externally as a means of establishing accountability and contestability.
9. Implement a process for the continued engagement of the AI provider with the acquiring entity for knowledge transfer and long-term risk assessment.	**a.** Consider that acquiring a tool that includes AI is not a one-time decision. Testing the application over its lifespan, adapting to new models, and extending to new datasets is crucial to success. **b.** Ask the AI provider to ensure that knowledge transfer and training are part of the engagement. **c.** Ask the AI provider for insights on how to manage the appropriate use of the application by nonspecialists.
10. Create the conditions for a level and fair playing field among AI solution providers.	**a.** Discover a wide variety of AI solution providers. **b.** Engage vendors early and frequently throughout the process. **c.** Ensure interoperability of AI solutions and require open licensing terms to avoid vendor lock-in.

Source: WEF (2019).

The procurement of AI expertise should be done within the procurement framework of the government, exploring flexibilities within the framework to ensure the best value for money. Practitioners should adopt an iterative and agile approach to developing a solution.

Innovative Procurement Examples

The U.S. government has launched two programs to facilitate the procurement of innovative solutions: FASt Lane and Startup Springboard. The FASt Lane program aims to facilitate and streamline the process for younger, innovative companies and suppliers to do business with the government. Under this program, the suppliers get shorter processing times for specified contract categories (e.g., IT Schedule 70 contracts) including a 48-hour turnaround for contract modifications and a turnaround as quickly as 45 days for new contract offers.

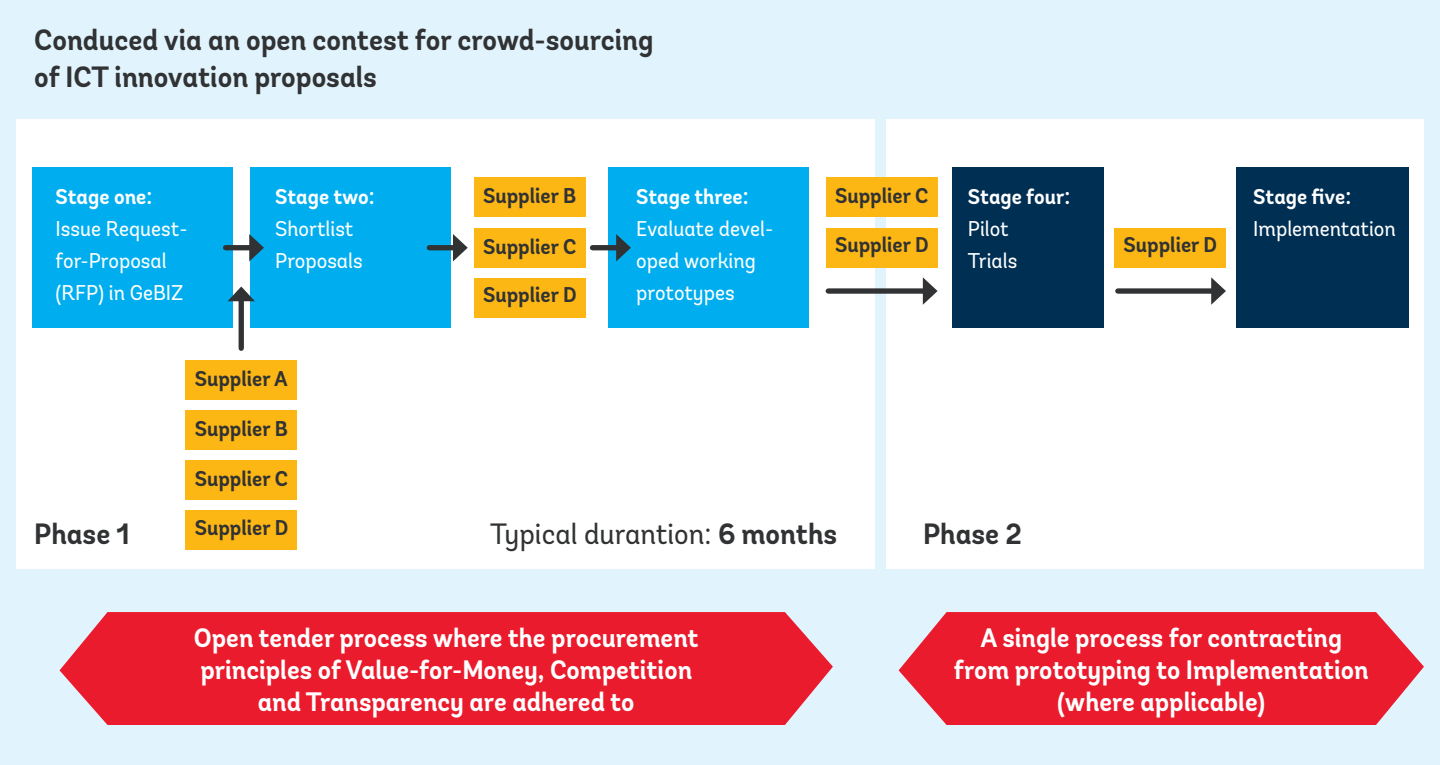
Under the Startup Springboard program, if a start-up does not have the required experience, it can use the experience of its executives and key professionals as a substitute for two years of corporate experience. Startup Springboard has one primary objective: helping federal agencies quickly gain access to the latest innovative technologies from fresh, vibrant private sector firms (Nakasone 2018).

Smaller digital economies also offer similar flexibility in their procurements. In Israel, the government issues challenge tenders that outline the problem statements, without the solution specifications.

The Government of Singapore launched a process called Call for Solutions. It entails sourcing ICT innovations through the evaluation of working prototypes and awarding contracts by stages to one or more suppliers. These multiple solutions are assessed in parallel through a series of pilot trials when the preceding stage or pilot proves successful. Facilitated by the Infocomm Development Authority of Singapore, this process will allow government agencies to collaborate more closely with the industry on ICT innovation needs. The EC adopted a similarly innovative approach.[18] Figure 13 depicts the process in Singapore.

> > >

FIGURE 12 - Singapore Procurement Model



Source: Reproduced from Annex A: Innovation Procurement for Singapore Government, Infocomm Development Authority of Singapore, available at https://www.imda.gov.sg/-/media/Imda/Files/Inner/Archive/News-and-Events/News_and_Events_Level2/20120531094015/AnnexA.pdf.

18. For more information, visit Shaping Europe's Digital Future on the website of the European Commission at https://ec.europa.eu/digital-single-market/en/innovation-procurement.

Role of the Public Sector in Society

The public sector can have a much wider role in governing AI outside the government for society at large. Its many facets include promoter of science, technology, and innovation culture to be a source of talent for AI; promoter of research in academic institutions; a regulatory body to regulate the AI developments in the private sector; and a promoter of AI by opening up its administrative and sectoral data to the private sector in machine-readable and downloadable formats to promote innovative use of these data. In this manner, the public sector can set the direction for the development of technology and set the rules for its application.

For example, in the United States, the White House issued Executive Order 13859—Maintaining American Leadership in Artificial Intelligence—to federal agencies to guide them on regulatory and nonregulatory oversight of AI applications developed and deployed outside of the federal government. The memo encourages the agencies to avoid regulatory and nonregulatory actions that needlessly hamper AI growth. It also provides guidelines on new regulations to ensure the principles of AI, as described in this paper, are adhered to in the private sector as well. It calls on agencies to facilitate the private sector innovation and growth by giving the public access to agency data. This access should be open, public, and electronic according to the Public, Electronic, and Necessary Government Data Act.

AI Operationalization in World Bank Projects

Task teams sometimes support Bank clients in experimentation and proof of concept for AI within the scope of World Bank projects. These engagements may be development policy operations, investment project financing, or advisory services, and analytics. The Bank's New Procurement Framework is flexible enough to allow experimentation with agile approaches, customized to the context; see Box 3.

> > >

BOX 3 - Procurement: Important Steps to Consider

A few steps for developing the RFP and designing the procurement process are given below:

- Outline the problem, not the solution specifications. The problem must be agnostic to technology. Special considerations should be given to sources of data and their quality.
- Define the benefits or results, which are of strategic importance and impact.
- Align with existing legal frameworks, public policies, and government strategies. Ethics and associated risks should be assessed together with mitigation strategies. Risks should be managed, as it is difficult to eliminate or avoid risks.
- Constitute a working group or multidisciplinary team.
- Establish mechanisms for transparency and accountability of AI systems.
- Ensure knowledge transfer from the AI vendor.
- Ensure value for money and fairness through competition, especially for scaling up AI that will involve large investments.
- Ensure code ownership. AI vendors could standardize the code, make it agnostic to client context, and resell the license, as with any technology, to create win-win. Consider opportunities for open-source code sharing.

Source: WEF (2020).

Ethical Considerations

Managing AI ethics is important and unavoidable for the productive use of AI in either the public or private sector. Failure to address ethical considerations, in government and private-sector AI solutions alike, leads to public mistrust and potential backlash. Most of the discourse on AI is dominated by the power of technology to process data faster, learn faster, and propose or take actions automatically to increase efficiencies and effectiveness. However, the societal implications of wider AI adoption have ethical dimensions that need to be understood and addressed at the outset. This chapter focuses on those ethical dimensions needing national-level policy response, while the technical risks to be mitigated at the implementing agencies level were discussed in Chapter 3 on AI risks.

AI harbors the inherent risks of automating poor decision-making and hiding complex decisions behind opaque algorithmic logic. AI can also do harm, for example, through AI-generated disinformation campaigns on social media. Malicious actors may leverage AI to further strengthen their influence over society.

Policy-level concerns on the ethical use of AI are can fall under the following three categories:[19]

- **INEQUALITY.** Bias in the use of algorithms, or as a result of a biased data pool may enhance negative bias toward vulnerable and weak communities and exacerbate inequalities; AI could lead to more demand for higher-skilled labor and exacerbate the returns to education which may not be equally accessed in the first place.

- **CONTROL.** AI could increase the misuse of information, surveillance, and use in defense systems.

- **CONCENTRATION.** The concentration of power and wealth in a few actors could be aggravated through the net flow of resources into a few firms, and success in achieving singularity when machines become equal or better than human general intelligence.

The detailed discussion on policy level ethical issues is given below.

19. See WDR World Bank 2016.

Inequality

AI may lead to specific job losses in both public and private sectors, more likely among lower-skilled workers, which has implications for the government to skill-up its workforce and to introduce policies that manage this transition. It is estimated that as much as 30 percent of today's jobs will be replaced by AI and automation by 2030, and up to 375 million workers in both the private and public sectors worldwide could be affected by emerging technologies (McKinsey Global Institute 2017). While the impact in the governments of the World Bank's client countries is likely to take place time, it will also take a longer period to prepare the workforce for the future. According to one study, 50 percent of the activities people do can be automated by adapting currently demonstrated technologies. This could have significant implications for the use of automation in the public sector. To manage this change, a distinction should be made between human-replacing AIs and human-assisting AIs. Government policies should promote human-assisting AIs, rather than human-replacing ones. To offset the effects of AI, unskilled labor should be progressively diverted to sectors needing personal attention and care, including health, education, and hospitality sectors.[20]

The potential threat to low skilled jobs in the private sector from AI is also a potential issue for the World Bank's client governments whose comparative advantage economically stems from a large unskilled and semi-skilled labor force. Unlike the innovations of the past, AI solutions could be more labor-replacing than human-enhancing. German robots have already begun replacing workers of garment factories in Bangladesh.[21] Chatbots are increasingly taking over call center work. It is estimated that 80 percent of customer interaction will be managed without human interaction. Autonomous vehicles could soon become a reality, with potential erosion of jobs for the taxi, bus, and Uber drivers in all countries. On the optimistic side, countries could potentially increase productivity in sectors like agriculture, health, education, and climate change through human-enhancing use of AI. For example, AI can improve diagnosis through image recognition, increase crop yield through monitoring soil and crop health using drone-generated data on farming, strengthen the fight against fraud and corruption through the reconciliation of data from multiple data sources.

To manage the labor market transition, the policy framework needs to be developed to show how investments in human capital and skills will be deployed and how equality of access to skills enhancement opportunities will be managed. Priority should be given to research, education, and skill development programs. Investing in such skills now for the future use of AI in the public sector is also important. Special emphasis should be given to managing equality of access and reaching groups vulnerable to missing these opportunities. This could include scholarships, apprenticeships, and research funding in computer science, STEM education, and AI-related disciplines such as data science for skill development. Governments could also create an innovation fund, loan programs through state development banks, and income-contingent student loans. Variations are already used in China and Brazil, and examples can be drawn from the experiences of Denmark, the European Union, Finland, Germany, Israel, and the United States (Mazzucato 2015). Governments could also initiate hackathons to promote opportunities for emerging talent and start-ups, as is being done in many countries, including Austria, Estonia, India, Poland, Pakistan, and the United States.

Control

One of the potential risks introduced by AI is who has control over the information and how it can be manipulated for certain outcomes. Developing policies early on to deal with the use of AI to misinform or mislead groups is an important issue. The use of fake news and targeted but distorted newsfeeds can have several consequences leading to polarization of ideas and groups in society and influencing political choices. AI-enabled social media bots can analyze millions of personality profiles by using cookies to track websites that people visit and deliver tailored news, including fake news, suitable to the profile. Fake or selected news can be used as a tool for manipulating political outcomes and discrediting a political opponent. Managing the development of policies and legislation to manage what is and is not acceptable, while at the same time balancing rights to form an opinion, is a complex endeavor.

Governments should develop or strengthen policies and agencies that cover the treatment of online propaganda, misinformation, libel, and cybercrimes. Agencies are required to monitor policy compliance and track, prevent, and investigate disinformation to protect its citizens, to enforce compliance and sanction lack policy violations. Governments

20. Stiglitz 2018.
21. Wall Street Journal.

need to regulate and influence social media Big Tech companies (such as Facebook, Instagram, and Twitter) to ensure the appropriate use of AI tools and to take down content that is malicious, hateful, propagandist, and false.

The government's use of AI to provide citizens with information about access to services also needs to be covered by a policy that governs the use and re-use of this information. This will mitigate the risk of misuse of this information. Handled correctly, AI has enormous potential to ensure appropriate targeting of information about, for example how to access certain government services, to the groups most likely to be beneficiaries of such programs.

AI can also be used as a tool that can be used to track and surveil people, something that may be very helpful, for example in managing public health outcomes or reducing traffic congestion, but which also has risks of excessive government surveillance that could infringe on human rights. The opacity around governmental use of AI as a surveillance tool makes it very difficult to assess the magnitude of the problem. According to Feldstein 2019, at least 75 out of 176 countries were using AI technologies for surveillance. Typical platforms for surveillance include smart cameras under the smart city initiatives, smart police projects, and facial recognition systems for contact tracing to quarantine COVID-19 carriers. AI can be used to track the movement of employees to monitor performance in the public sector (police rounds), the private sector (pizza delivery). Therefore, policies governing the privacy and rights of employees need to be developed to avoid misuse of AI.

Data privacy laws, transparency, and citizen's voice should be strengthened to manage risks that AI used for surveillance is in the public interest. Europe has adopted the General Data Protection Regulations and many governments have legislation covering personal rights to privacy, personal data protection, and civil liberties but compliance and enforcement remain challenges. Promoting full disclosure of information being tracked by AI and robots through existing transparency frameworks can be strengthened, and managing the risks of misuse of such measures will pave the way for the productive use of AI in this domain for the public good, for example, to trace and identify those at risk from contact with a contagious disease.

Weaponized AI systems have the potential to increase the use of autonomous weapons in conflicts, requiring a specific policy to address the ethical use of AI in warfare. The control and use of autonomous weapons systems may in turn destabilize regions and increase potential conflicts as human costs may be reduced. Global military spending on autonomous weapons systems and AI is projected to reach $16 billion and $18 billion respectively by 2025.[22] The cost of drones that can be advanced enough to defeat a U.S Air Force fighter pilot in combat simulations is as little as $35.[23] AI principles of adoption emphasize human control and AI use for human benefit. The application of these principles to the use of autonomous weapons is an issue of global importance and coordination. Global governance through multilateral forums and international cooperation is needed to address these issues. The role of civil society to influence the debate is also important.

Concentration

AI can also lead to increased concentration of wealth in the hands of a few individuals controlling the big firms. These big firms can finance expensive research and attract top talent through better financial incentives. These big firms not only control the AI research and talent but also the associated data center infrastructure through cloud computing. This concentration would provide even more resources at the disposal of these individuals to influence public policy through campaign financing, lobbying, corruption, and influence peddling. This will also lead to a net outflow of resources from the developing to the developed countries, as most of these big firms are based in the high-income countries.

22. Sander and Meldon, 2014; Research and Markets, 2018.
23. Cuthbertson, 2016.

>>>

Government's AI Building Blocks

A whole-of-government, data fabric AI architecture is central to the technology vision of the government and forms the building blocks for the use of AI. The government's approach needs to encompass interoperability and security and the importance of continuity of architecture among AI systems designed for use in a whole-of-government architecture. By understanding the components and building blocks of AI systems at a high level, common knowledge becomes a tool for exploring relevant entry points with technologists to guide the broad direction of possible solutions. Three key concepts that constitute the building blocks are (a) a whole-of-government architecture; (b) interoperability; and (c) data standardization.

Whole-of-Government Architecture

Most World Bank client countries are managing stand-alone legacy systems, often referred to as "silos." These systems are not interoperable or have problems with interoperability. Since AI models need large amounts of data to work well, the "ideal" architecture needs the silo systems to feed data into a large distributed data storage repository—often referred to as a data lake. The data lake is then made accessible to various AI applications. A government aspiring to greater digital transformation should adopt a whole-of-government architecture, which is the de-facto industry standard.

Siloed systems can be "stitched together" through a common data platform. A government has many ministries, departments, and divisions. Each one typically operates autonomously, but often reports to a central government agency. A data fabric is a similar concept (see Box 4 below). It has several data centers with many departmental computing resources. Each one operates autonomously, but each one reports to a central computing administration system using a standard set of rules or protocols for data storage, security, and processing. They are all "stitched together" using a common software platform that spans the whole-of-government. The data though remain separate and independent. This is a simple description of how the kind of system that can lead to incredibly powerful capabilities in AI and data processing.

> > >

BOX 4 - Data Fabric in Brief

The term “data fabric” refers to the very large-scale continuous Big Data architectures used by some of the largest organizations in the world. A data fabric provides storage, computation, and security for organizations with exceptionally large data pools, such as governments and multinational corporations. A data fabric also supports distributed computing between multiple data centers spanning entire countries. In 2019, Gartner identified the data fabric among the top 10 trends in data and analytics technology (Gartner 2019).

The *difference between Big Data and data fabric*. Big Data systems are more uniform and monolithic while data fabrics offer a common computing layer across a variety of systems that include these characteristics:

- One or more databases containing data from various sources (Big Data). The database and file system layers comprise the data lake as explained later

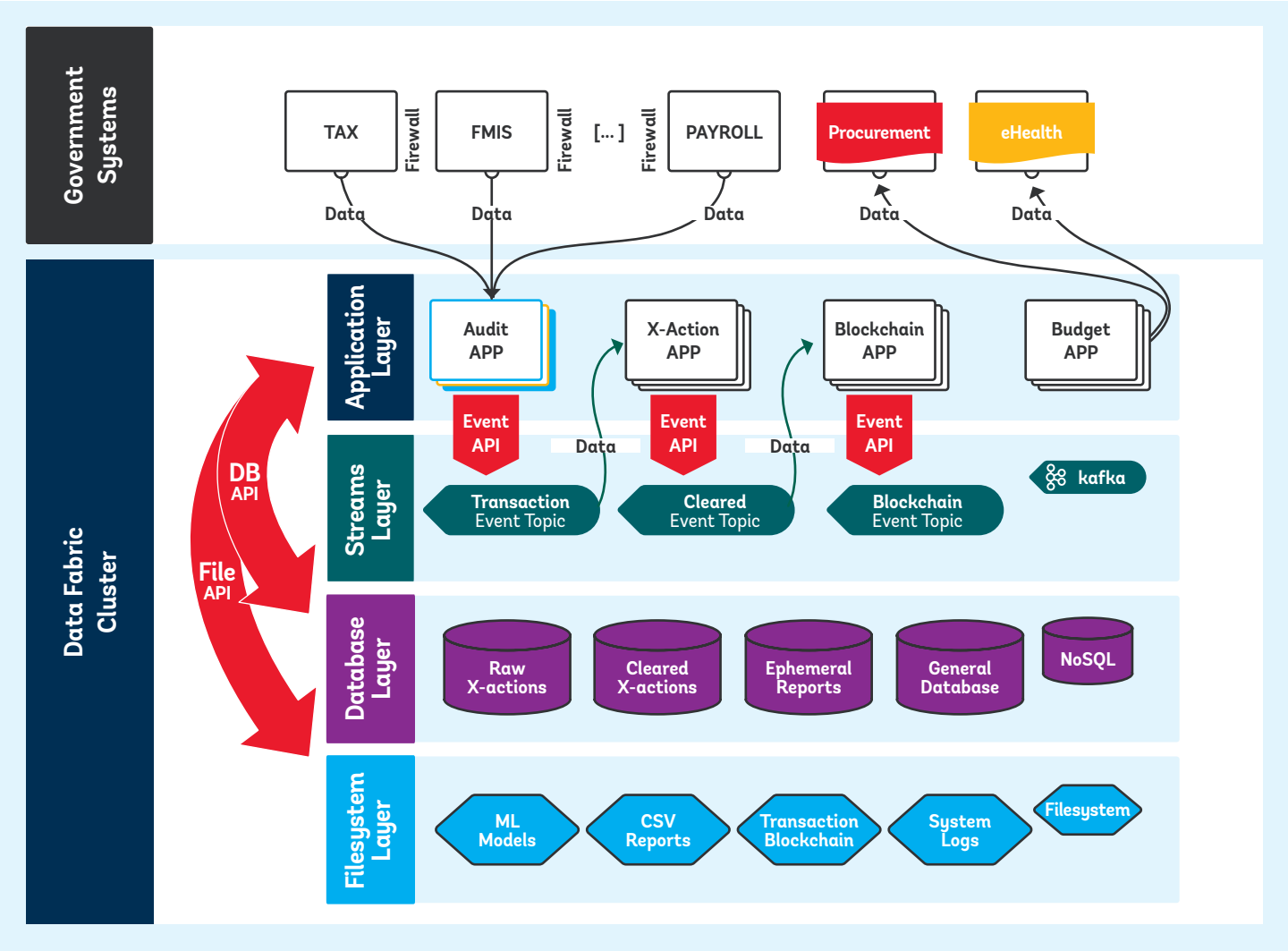
- Application Program Interfaces (APIs) to connect with external government systems such as financial management information systems, payroll, integrated tax administration systems, and e-procurement.

- Data and cluster management tools, including:
  - Storage APIs for real-time (or batch) data ingestion, updates, creation, and deletion.
  - Data tools such as streaming, machine learning, and preprocessing systems.
  - Administrative tools for data access control, monitoring, and provisioning.

The general purpose of data fabric architecture is to unify data storage and AI computation across many independent government departments while keeping data safe from loss and protected from unauthorized access. A data fabric does not replace an existing architecture in one iteration. A government can roll out a data fabric over time and incorporate all the existing data systems into the fabric architecture, slowly replacing “old” walled-off legacy systems with “new” interoperable systems at their discretion. Figure 14 presents architecture built atop a data fabric.

> > >

FIGURE 13 - General Data Fabric Architecture for Whole-of-Government Use



Source: The World Bank.

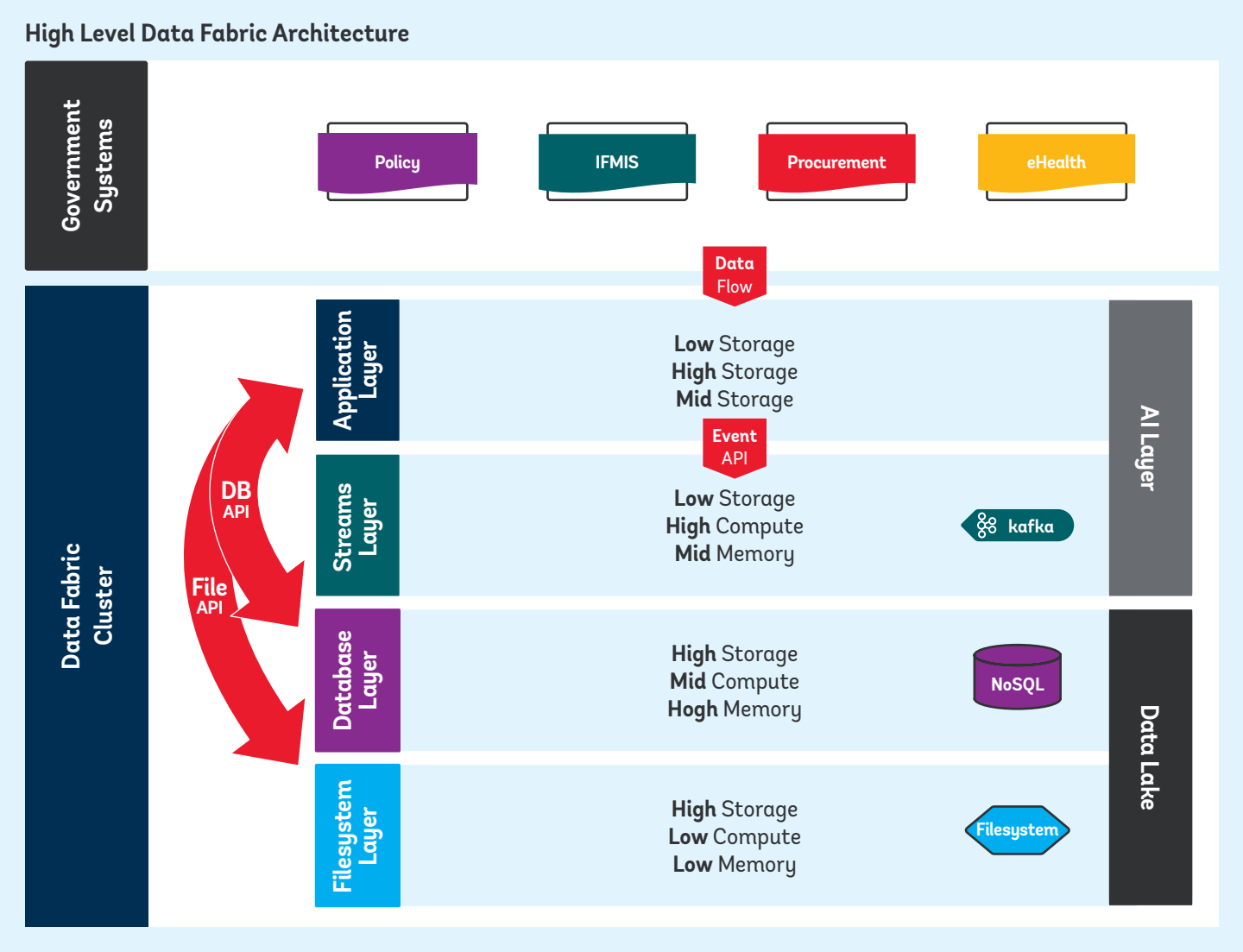
A data fabric architecture has two high-level layers: government systems and the data fabric cluster. Government systems are shown in the gray box at the top of Figure 14. Inside this layer are all the government's applications, which belong to the various departments. Each white box represents departments or divisions. The two applications on the right, Procurement and eHealth, represent commercial-off-the-shelf (COTS) solutions. All applications send and some receive data from the data fabric cluster. Existing applications, such as tax, FMIS, and payroll, share data with an AI application layer inside the data fabric cluster, which is at the top of the data fabric cluster portion of Figure 14.

The data fabric cluster layer is subdivided into four layers: application, streams, database (DB), and filesystem. The first layer, the AI application layer, holds all the custom AI applications inside the data fabric cluster. Underneath that, a stream layer ensures that data flows from one place to another inside the cluster in real-time. Beneath the streams layer, a database layer gives departmental AI applications a place to store their rapid-access data. Lastly, beneath the database layer, the filesystem layer stores archival data and even larger data structures for long-term storage in blockchains and flat file systems similar to the hard drive on a personal computer, but scaled to handle the data needs of an entire country and all its citizens. Appendix A provides a discussion of the potential role of blockchain technology for government systems.

The Standardized Application Programming Interfaces (APIs) are the threads that stitch the fabric together. Figure 14 represents these connections with bold arrows. They are labeled DB API, File API, and Event API. They are the core of interoperability for this architecture. The most successful large-scale operations, including India's system for issuing a unique digital ID (Aadhaar), use this design.

> > >

FIGURE 14 - High-Level Data Fabric Architecture



Source: The World Bank.

Figure 15 illustrates the resource utilization and performance requirements in a mature whole-of-government architecture. Resource requirements and utilization are major factors in determining the total costs of ownership (TCO). Here, the AI application layer and the streams layer comprise the whole of the primary AI layer. The database and file system layers comprise the data lake. The general takeaway from Figure 15 is the distinction between the AI layer and the data lake within a data fabric cluster. Any external AI solutions can leverage data APIs to access data within the cluster from anywhere within the dominion of the whole-of-government. Also, the relationship between resource consumption and broader layers of architecture is not uniform, which allows for lower TCO. Moreover, the data fabric cluster remains independent of top-level government systems. Each department may have its services built into silos. All the computers within the data fabric cluster can have different capabilities and distributed locations.

Sometimes referred to as a data lake, a data fabric – though it is more than a data lake as mentioned in Box 4 – is made up of commodity hardware systems at up to exabyte scale (1018 bytes) throughout a wide variety of architectural patterns; some cloud-based, some on-premise, and some in a hybrid configuration of both.

Data storage for AI is broken into three tiers: ephemeral, persistent, and archival. Each tier favors a particular subset of structured data, accessible through a standardized interface and abstracted into an accessible format through programmatic and algorithmic convention, which may be open source or proprietary. Storage in 2020 can be localized to one machine, one drive, or spread across geographies in sophisticated and redundant data topologies that distribute exabytes across global geographies while offering access-control layers (ACLs) for strict management. The storage tiers are considered part of the storage layer in the AI technology "stack."

> > >

BOX 5 - Blockchain: Distributed Ledger Technology

The data in the persistent file system tier may be grouped into blocks of information and hashed with an identifier linked to a previous block: a blockchain. This blockchain formulates an archival decentralized ledger. To further the utility of blockchain technology, also known as distributed ledger technology (DLT), applications relying on the use of a distributed ledger may prevent the completion of a transaction or asset transfer until enough computing nodes in the infrastructure reach consensus through Application Program Interfaces in real-time.

By distributing and requiring consensus among participating compute nodes, DLT essentially offers an immutable solution to asset tracking and transactional audit. Modern DLT solutions even offer safeguards against Byzantine attacks in which malicious agents attempt to gain majority influence within a network of computing nodes that are running DLT applications. Appendix A has more detailed information about how the blockchain plays an integral role in the long-term use of whole-of-government data architecture.

Besides the core AI architecture layers, a few other considerations play an important part in the technical design of the data fabric AI stack.

Interoperability Patterns

Data silos are the opposite of scalable, interchangeable, and interconnected computing systems. They are rigid, limited, and isolated from other systems. Imagine a government in which various departments, ministries, or divisions did not speak a native or common language and could not communicate. These are silos. Successful large-scale deployments rely on the following patterns to compensate for a lack of interoperability between silos:

- Data exchange standards and schemas.
- Secure APIs.
- Cohesively interconnected layers of services using IPC best practices.
- Geographically distributed data centers within the data dominion.
- Architectural redundancy and replication.
- ACLs.

A data silo is an architecture that is isolated due to the absence of a common API for IPC. Data silos can emerge from vendor lock-in, proprietary systems design, or poor planning. A system of silos lacks a common denominator to effectively allow for interoperability. Data are trapped in the silo. Over time, the silo will bloat and stagnate with information that could otherwise be utilized by AI systems.

Various agencies and departments tend to pursue entirely independent solutions to solve narrow problem statements specific to their short-term needs. This common practice creates complex pervasive fragmentation. As a result, interdependent organizational units end up with entirely independent systems that are isolated from one another in the long run.

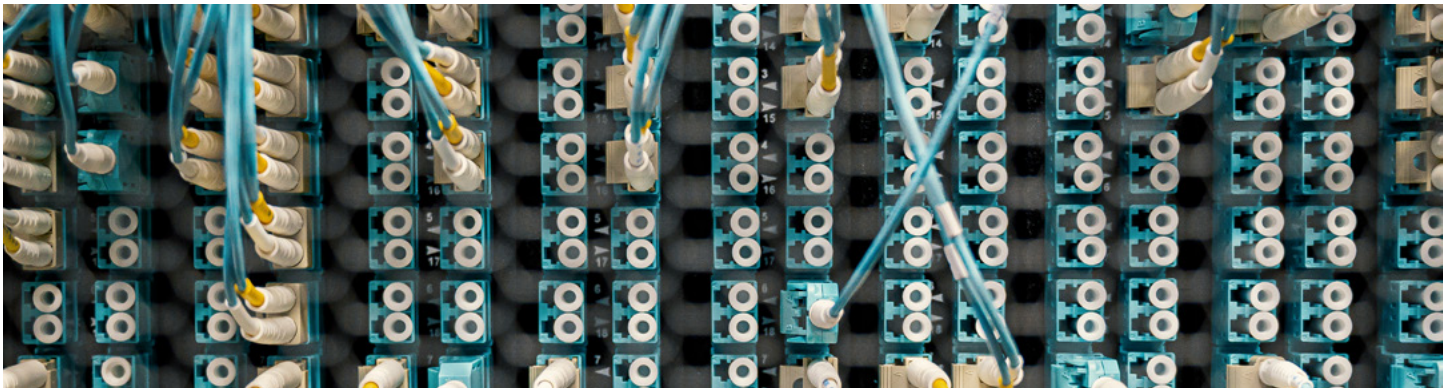
Siloed systems can potentially become bottlenecks for data sharing that prevent useful implementations of AI. As a result, to discover trends and patterns with AI, departments must export enormous volumes of data to a centralized storage location, which is extremely time-consuming and costly.

Data silos stifle whole-of-government AI development, although they are preventable. This pattern is consequential to siloed systems, reflective of turf sensitivities, and lack of interagency coordination mechanisms. Luckily, there are solutions to address the issue.

> > >

BOX 6 - Actionable Insight: Data Fabrics Can Overcome Silos

A data fabric architecture prevents and solves problems arising from data silos. A central agency, responsible for government-wide digitalization, could deploy data fabric architecture to overcome silos. It deserves much-needed consideration for governments wanting to harness the power of data by streamlining operations with a large-scale AI-ready infrastructure.



Data Standards

Data are the lifeblood of any AI architecture—it is the "gold." As a form of untold wealth, data are worth sharing among stakeholders within an organization. To overcome entrapment in silos, interoperability plays a crucial role in successful AI systems development. Although enterprise computing solutions—such as enterprise resource planning (ERP), data lakes, and databases—often use compressed binary streams internally, there is a high probability their data storage systems open a pathway to external applications through an API. Standards make this possible.

Instituting data governance arrangements promote standardization of data necessary for interoperability. Modern governments, like Estonia, create data governance councils and appoint data stewards in each agency to coordinate data standardization and interoperability. These arrangements are part of the data governance strategy that defines the authority and control over the data assets and includes policies, processes, standards, definitions, and data exchange arrangements.

Data standardization across agencies could also follow good practices of standardization internationally. These are more common in the private sector, though some models also exist in the public sector. In the private sector, several standards evolved using these practices. Programming, internet, and network protocols rely intensively on standards established by the Internet Engineering Task Force, International Standards Organization (ISO), and the Institute of Electrical and Electronics Engineers (IEEE). The processes and substructures within these organizations are oriented toward developing a very uniform agreement between ground-level engineers responsible for implementing the technologies that drive the internet's evolution. When a fundamental technological agreement must be reached, a consensus is reached through a Request for Comment (RFC), which contains guidelines for the implementation of and use of the technology needing standardization through peer review. A complete RFC must contain core tenets explaining and enumerating every behavior and function in technical detail and depth. RFC practice was used in several global standards: World Wide Web, JavaScript Object Notation (JSON),[24] and the Portable Operating System Interface, a family of standards developed by IEEE that provides a standardized protocol for communications within and between computing file system layers worldwide.[25]

Good models also exist in the public sector at the international level. The Open Contracting Data Standard targets contracts in general and enables disclosure of data and documents at all stages of the contracting process by defining a common data model.

> > >

BOX 7 - Actionable Insight: Governments Should Standardize Data

The central agency may develop standards for data formats and interoperability through engagement with line ministries. The creation of data governance councils and nomination of ministry- and agency-wise data stewards help ensure standard quality in data sharing. Also, engagement with stakeholders to develop consensus using Requests for Comments is successful among international standards organizations.

24. See Request for Comments (8259), "The JavaScript Object Notation (JSON) Data Interchange Format," Internet Engineering Task Force, at https://tools.ietf.org/html/rfc8259.
25. The Open Group 2018.

Open, consistent standards and methodologies are the ground-level blueprints for a successful whole-of-government implementation of AI technologies. Prospects are that a global governance standard for data will likely emerge over time. This notion of a data standard can extend further to include suggested best practices for developing interoperable data fabrics within and across governments. Their practices may differ substantially from one another, though the technical processes for accountability and integrity share common standardized infrastructural patterns.

By enforcing standards, the international community of policymakers can achieve an intergovernmental vision of AI interoperability. By leveraging standards for document data storage, and APIs common to the databases supporting various existing silos, governments can deploy system integrations that evolve continuously with the trends and advancements in AI at a national and international scale.

Access to data is the key to managing governments at all levels of AI deployment. The software platforms and solutions that do the actual computation often provide APIs that access standardized databases. Developers and data scientists alike may be able to access a data fabric over network interfaces and conduct experimental research that helps determine the proper course in developing permanent AI solutions to common problems in government. This will also provide avenues for more effective data collection, aggregation, experimentation, policy management, and access control.

Data are more accessible than governments may realize. Leveraging data stored in existing silos should be the essential tenet of any digital transformation strategy. The majority of ERP, custom-developed, or open-source solutions these days provide some type of data access control through direct communication with the database layers that these systems utilize. Therefore, siloed solutions do not require forced obsolescence either. Governments may continue to utilize them while they transition to newer data fabric oriented architectures. Limitations do exist among mainframe systems developed before the turn of the millennium, which require custom programming to extract data from COBOL (or common business-oriented language) and other flat file systems.

An application within a data fabric can query existing databases for new records and feed the data to an ingestion layer, which routes information to the appropriate hardware resource controllers within the data fabric. While it is true that granular data access can lead to a "spaghetti" dependency structure, entirely independent distributed services can be tuned to execute any arbitrary set of applications, especially whole-of-government AI models over the long run. More granular information about advanced connectivity is available in Appendix A.

In conclusion, a data fabric offers an intrinsically resilient, adaptive, and decentralized architecture that has no single point of failure. Trends are moving toward AI as an operating system among developed governments. The introduction of the FedRAMP marketplace established by the United States, which provides vendors with stipulated standards, requirements, and guidelines for being authorized to provide cloud services to federal agencies, provides an indicative direction of the emerging trends.

Within whole-of-government systems, standardized access to data enables many types of practitioners to experiment and design all kinds of use cases for AI. In reality, the AI application layer can contain tens of thousands of AI models for all types of purposes. Each application can easily leverage all types of information simultaneously stored within the architecture—data such as text, audio, video, and biometrics. This allows for better solutions over the long run by enabling a fail-fast approach using data access as a baseline. Ultimately, governments can develop long-term strategies in AI innovation that count on standards. There is little doubt that a tidal shift is unfolding for governments that are serious about improving their long-term strategic advantages in AI.

To proceed and formulate a more in-depth view of AI systems, see Appendix A. It dives into the core concepts of AI in practical applications. The concepts are meant to inform the reader of the basic, advanced, and real-world AI applications. Again, understanding these foundational concepts demystifies much of the jargon and hype orbiting the topic of AI. The key topics that Appendix A depicts in greater detail include:

- Project development patterns.
- Cloud, hybrid, and on-premise architectures.
- AI connectivity.
- Microservices.
- AI models.
- AI workflows.
- Distributed ledger technology (DLT) in AI architectures.

7

>>>

Conclusions

AI is still a new area even for many of the advanced digital economies, but its rapid diffusion in every facet of private and public life is increasingly more visible. The enormity of development challenges requires exploring modern approaches, tools, and techniques. AI offers immense opportunities to address some of these challenges. However, it has inherent risks that can have profound consequences for society. Governments have to lead the efforts to manage these risks while promoting the use of AI in the private and the public sectors. This paper distills existing knowledge on these aspects for client governments. Conclusions as well as priorities going forward are highlighted here.

Human-centric AI design is a key principle to guide the development and deployment of AI. AI will not eliminate human oversight in decision-making. Also, entirely externalizing decision-making using AI is unrealistic due to bias, which is impossible to eliminate but reasonably controllable. Public sector AI technology must remain under the guidance of humans because it has the potential to affect trust, human health, safety, and overall well-being. Fortunately, the state of the art in AI demands it and all mission-critical AI deployments keep humans "in the loop" to varying degrees.

Governance and government practices benefit from transparency and evidence-based decision-making. AI systems must operate with transparency, human oversight, and neutrality while attempting to manage and disclose bias, which humans will never fully eradicate from AI solutions. However, well-managed AI solutions yield a repeatable model that may provide fundamental services through an open-source consortium of international collaborators. Therefore, while this paper encourages collaboration, any general government AI solution must take security, privacy, and data protection into full account to protect the sanctity and privacy of people and their governments. Currently, close to 135 governments are implementing privacy and data protection in their legislation, which applies to AI for the benefit of stakeholders.

The process of AI implementation is a journey. It starts with the most critical basic foundation: the acquisition, aggregation, management, and storage of reliable data. With quality data in hand, policymakers, data scientists, and AI engineers can perform introspective and comprehensive iterative deployments to expose the possibilities for full-scale AI systems. The journey requires coordination and collaboration between teams of stakeholders at all levels of government. It also demands that the outcomes earn the citizens' trust through disclosure, explainability, and transparency wherever bias is a concern. Where necessary, administrators may provision AI algorithm audits, especially in cases requiring forensic investigations.

Governments need to adopt a large-scale data fabric architecture to serve as the common denominator for standardized data interchange among a fully digital whole-government infrastructure. This approach enables robust AI solutions to grow and evolve with changing needs. The fundamental shift in the mindset of developing countries involves an emphasis on interoperability and IPC through standardization and API enablement.

The promise of AI is riddled with commercial marketing hype, but the fundamental value of the introspection cannot be overstated. AI systems offer a mechanism for qualitative predictions using quantitative measures of information. The various patterns of AI analysis provide tools for attacking a multitude of problems that are emerging in the face of increasingly intricate governance systems. Regardless of the flavor of governance employed, one thing remains clear: AI has the potential to revolutionize human intelligence in unprecedented ways. Despite the hype associated with being at the forefront of innovation by being the first to deploy one or more cleverly marketed solutions, the real focus should be on solving problems for internal governance and citizens. Also, government agencies must be willing to adopt standards and practices that enable fast and agile delivery, with an acceptable degree of failure risk.

A myopic view of AI is counterproductive. Immediate problems are like individual fires in a forest ablaze. Governments must avoid this tendency and commit to building a whole-of-government infrastructure that allows line agencies to operate interdependently. Systems at this scale require the collective efforts of nearly everyone in the scope of government influence to learn, trust, and invest. By creating fabrics of information, governments can promote their missions of better governance, transparency, accountability, and efficiency.

Priorities Going Forward

Based on the issues highlighted in the paper, several priorities could be considered by policymakers.

- **Governments must adopt policies and governance frameworks that promote human-centric AI while maximizing opportunities.** A few aspects of the policy framework are mentioned below:
  - » **Ethical AI requires the adoption of an AI policy and strategy.** It could be tailored to specific settings but should be approved at the policy level to provide the authorizing environment. Governments in many settings have issued AI strategies approved by the parliament, president, prime minister, or cabinet. These policies should be based on ethical principles. Governance and operational framework are essential to specify broad guidelines and institutional arrangements. An innovation hub could be established to pool talent, establish partnerships with academia and the private sector, promote research, and facilitate experimentation by line ministries. The innovation hub should source the best talent through adequate incentives. Innovative procurement approaches should be adopted to leverage private sector skills with agility to allow iterative, problem-driven approaches to the RFP. The implementation teams should also manage the risks associated with AI, including bias, security, and unintended consequences, among others.
  - » **Promote transparency and accountability through inclusion and multi-stakeholder engagement at every step of the AI policy design and implementation.** Affected communities and populations should be informed and provided with avenues for contesting AI logic without delays and hurdles.
  - » **Adverse ethical implications of AI could be managed through broader economic policies.** These could include industrial policy, tax policy, competition policy, human capital policy, among others.
  - » **These policies should also promote digital skills, education, and redeployment efforts to support people as they adjust to the shifting nature of work in the coming decades.** Unskilled people and disadvantaged groups should be given special attention.

- » **A policy framework to fight online propaganda, misinformation, libel, and cybercrimes should be given priority.** Also, governments could establish agency mandates to monitor policy compliance and track, prevent, and investigate disinformation to protect their citizens. Engagement with social media Big Tech—Facebook, Instagram, and Twitter—should aim at encouraging the deployment of AI tools and professional fact-check partnerships to take down content that is malicious, hateful, propagandist, and false.

- » **Strengthen privacy, data protection, and civil liberties and monitor compliance, which is typically weak in most settings.** Promoting full disclosure of information being tracked by AI and robots through transparency frameworks should also be strengthened.

- **Investments should be made in human capital and digital infrastructure.** AI research, digital skills, AI entrepreneurship, and foundational digital technologies could be prioritized.

  - » **Investments should be directed to fund research, education, and digital skills development programs in general and in AI in particular.** They could include scholarships, apprenticeships, and research funding in AI, computer science, STEM education, and AI-related disciplines such as data science. Special emphasis could be given to disadvantaged groups such as women, minorities, and those at risk of being left behind.

  - » **Innovative entrepreneurship could be promoted.** This could be done through an innovation fund, loan programs through state development banks, income-contingent loans for students or others, and small business loan programs. Variations of these funding modalities are already used in China, Brazil, Denmark, the European Union, Finland, Germany, Israel, and the United States (Mazzucato 2015). AI could be one of the areas to be incentivized through these programs.

  - » **The innovation hub should be staffed with the appropriate talent on market-based salaries.** These skills are in high demand and could easily drain overseas.

  - » **Data fabric architecture, including interoperability, should be considered for investments.** This will overcome silos, and leverage data assets for decision-making, compliance monitoring, and analytics. The initial focus should be on interoperability, open data, and data standardization. A hybrid cloud option should be explored to leverage the computing power at much lesser costs to pilot AI solutions.

  - » **Proof-of-concept and pilot AI projects could be the starting point for exploring opportunities.** Many governments have deployed AI to solve specific problems. Key use cases include citizen engagement, service delivery, regulatory compliance, decision analytics, fraud, and anti-corruption. Hackathons promote emerging talents and start-ups as seen in Austria, Estonia, India, Pakistan, Poland, and the United States.

- **Risks should be identified and managed, rather than avoided.** They could be mitigated through self-assessments, peer reviews, and inclusion.

>>>

Appendix A. AI Technical Primer

Appendix A explains a large subset of technical and operational details about AI project management, architecture, types, methods, and models, as a self-contained primer. This is based on the best industry advice from practitioners. The consolidated information herein intends to benefit the reader. For even more information, practical guidance is available in many books, blogs, articles, and other technical resources.

Project Development Patterns

Agile Development

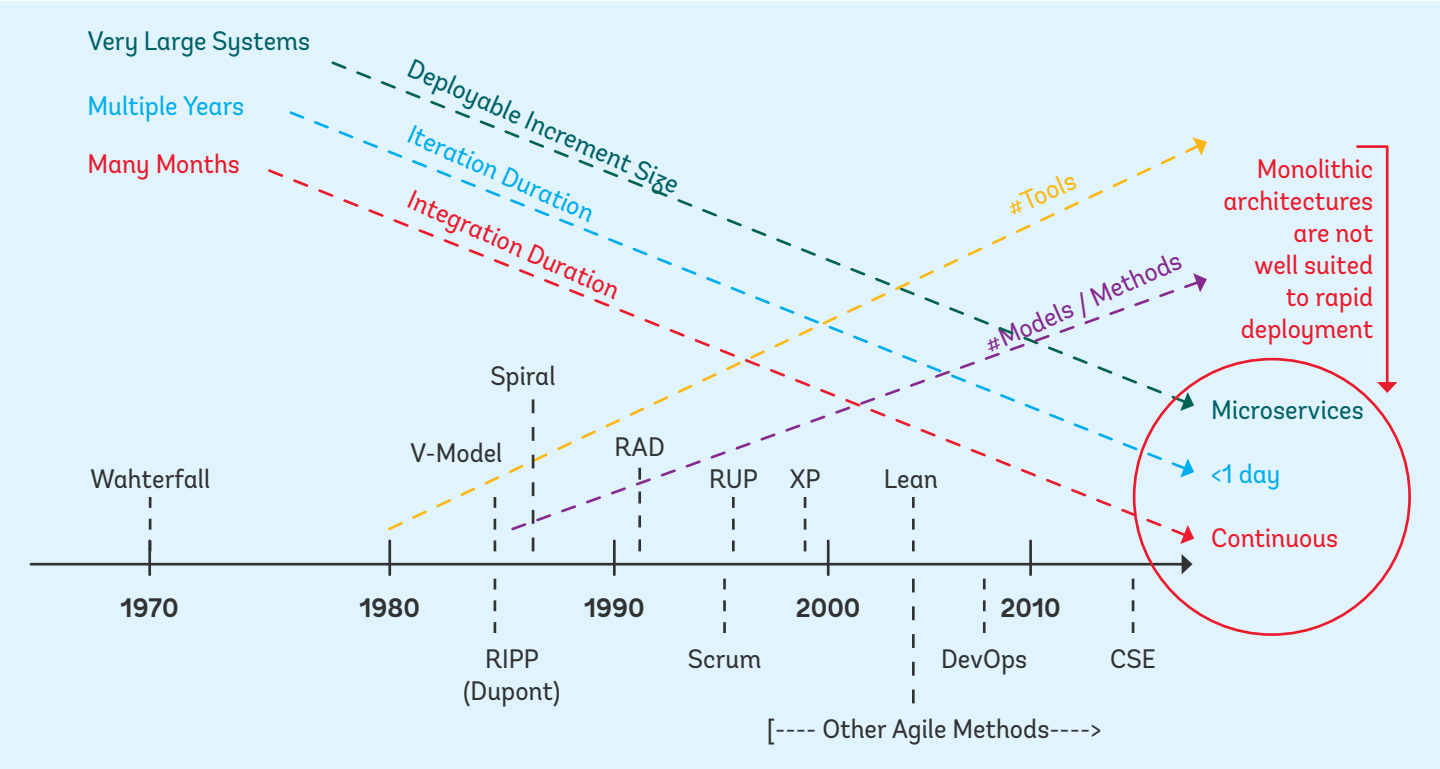
Iterative, agile development is the key to the steady adoption of any technology. The agile methodology offers an adaptive model for destructuring complex projects into manageable stages with discreet goals, short term development intervals, and continuous delivery. Agile teams convene regularly, often daily, in scrum meetings to disclose incremental progress and dependencies. Agile methodology is a longstanding backbone among organizations of all sizes.

Agile offers a method for execution that complements a goal-setting methodology consisting of objectives and key results (OKRs). They keep all levels of organization, especially individuals, holistically accountable to the project. Because OKRs are typically disclosed publicly to all stakeholders, the whole organization may audit the development process for measurable progress. Of course, process management is not a panacea for projects attempting to reinvent the wheel, or parts therein. To that end, there are ample turn-key solutions that ship with a unique set of caveats to consider.

Iteration times have been steadily decreasing in the decades since the 1980s. Early waterfall-based methodologies "iterated" through projects over months up to a year. The 1990s brought the adoption of the Rational Unified Process, an early precursor to Agile Development and eXtreme Programming (XP). These advances in management timelines reduced development iterations down to two to three weeks. The unit of code development has also decreased significantly since the 1980s with the advent of Service-oriented architectures and microservices. Today, Continuous Deployment techniques allow high performance organizations to release microservice applications to production several times a day. Figure A.1 illustrates the changes in iteration time and code volumes. Figure A.2 illustrates the change in unit of code over the previous 15 years prior to 2020.

> > >

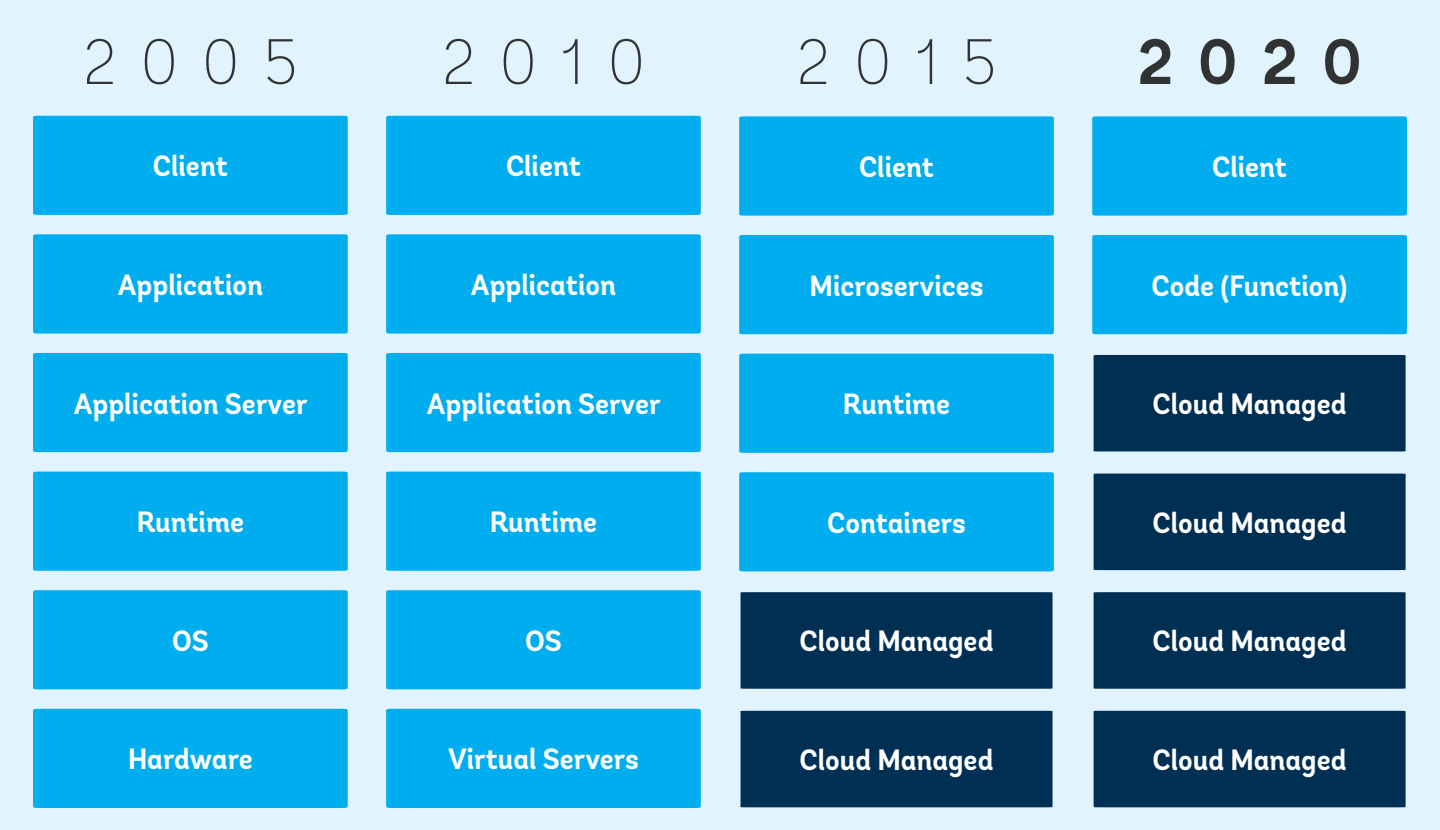
FIGURE A.1. - Iteration Times and Code Volumes versus Time



Source: Reproduced with permission from ©Paul Clarke; further permission required for reuse.
Note: Paul Clark, "Computer Science Lecture Notes," Dublin City University and Lero, the Irish Software Research Center.

> > >

FIGURE A.2. - Unit of Code Scale Change



Source: Reproduced with permission from ©Paul Clarke; further permission required for reuse.

Wherever possible, problems will require a consistent process for reducing complexity and establishing a manageable scope of execution. Project management plays a critical role in establishing a track record of success for software developments at the national level. The temptation to boil the ocean and attach a blanketed solution to a broad scope of operating requirements is often irresistible within novel undertakings. Project managers help mitigate risk and counteract scope creep by coordinating and elucidating the requirements and steps necessary for projects during the planning phase. This practice does not necessarily assume end-to-end project development of every possible use of a fundamentally valuable data infrastructure. Rather, project managers (PMs) solidify first-level operational requirements for a data infrastructure, the application layer, and critical external dependencies. These requirements capture the general requirements necessary for the future development of siloed executable applications, each assigned a dedicated PM that coordinates with the central development team to ensure consistent standards capable of supporting the various permutations of core and secondary systems. With proper coordination, standardization, and management among government stakeholders, government efforts will secure a development process that ensures that projects reach completion for timelines, which span changes in elected officials, and survive the varied political landscape.

Prior stepping into the domain of applied solutions in other sections, this appendix to the paper will step through some processes for reducing and scoping problems into solutions. These are by no means a comprehensive list of project management best practices, but they facilitate understanding of the intricacies of software development planning and execution. These processes are also not intended to replace the acumen of an experienced project management professional. Each project has a specific set of requirements that accompany project-specific nuances. Furthermore, the number of variables involved during execution necessitate careful coordination, investigation, and execution by professional planners and managers. There is no one-size-fits-all solution in AI project management, just processes based on the type, scope, and timeline required for execution.

Project Management

Avoid solutions looking for problems. All too often, technologists invent a groundbreaking solution in a theoretical environment and apply the solution to problems that simply do not exist. As unimaginable as this may seem, the promise of technology may outweigh the actual benefit when practical solutions fail to emerge from concrete and battle-tested best practices, even if those practices are manual or fragmented in nature. Governments tend to silo operations within the scope of an agency or department due to budgetary firewalls. Solutions developed by localized operations successfully focus on the problem at hand, incorporating turnkey solutions and consulting opportunities for highly specialized services deemed vital to deliverables needed within a budgetary window of opportunity. This practice leads to fragmentation and a lack of interoperability, often solving problems many steps ahead of the current set of requirements.

Conversely, as experimental technologies emerge from academic organizations and professional firms alike, their applied implementations may be directed toward advanced problems that ignore the scope of current operation. This leads to the presentation of a technology that is inconsistent with the immediate requirements of any organization, much less those of the entire group of organizations comprising government. In brief, no one has engineered anything resembling a "government-in-a-box," but many consultancies come close to selling solutions as a panacea for the most mission-critical problems. They do so in a manner that opens a dialogue that often demands full-scale adoption of a technological product that requires immense customization or otherwise a total replacement of the existing infrastructure that is incompatible with the fragmented, siloed solutions described previously.

AI technologies and automations offer myriad possibilities for enhancing the decision-making process used within manual systems. Replacing a manual system from day zero is not necessarily the best solution because of the lack of introspection. By establishing proper ground-level data infrastructure, solution architects can coordinate with data scientists to study the quality of information produced by the government and carefully scrutinize prospective product solutions with a knowledge of the internal workings of data. Thus, teams can subsequently derive quick wins before the need for advanced analytics or sophisticated software emerges in actuality. Therefore, it is a wise strategy to organize data early on through the implementation of policies that standardize data within a very large distributed data fabric that supports further development and the integration of proprietary software tools by providing APIs for filesystem data access.

> Entities should not be multiplied without necessity.
>
> *William of Ockham*

Resource Management

Among human resources, create a stakeholder hierarchy that is largely decentralized across areas impacted by the infrastructure. This means that organizational decision-makers need representation from inception. Eligible representatives include IT directors, executives, and project leaders: people who are central to planning, policy-making, standards, and execution. Choose only the key members required to communicate and address project requirements. These members will take responsibility for communication with supporting members of their respective teams. Only including the necessary individuals occupying central roles prevents paralysis by analysis. Supporting team members will have the power to comment on policies and standards that emerge by issuing documents in RFC format. Teams review comments as they funnel into the project and conduct discourse to evaluate and settle on a final specification for project requirements. Ultimately the top-level representatives ensure that the needs of their organizations are met. Standards may evolve over time to reflect changing architectural requirements.

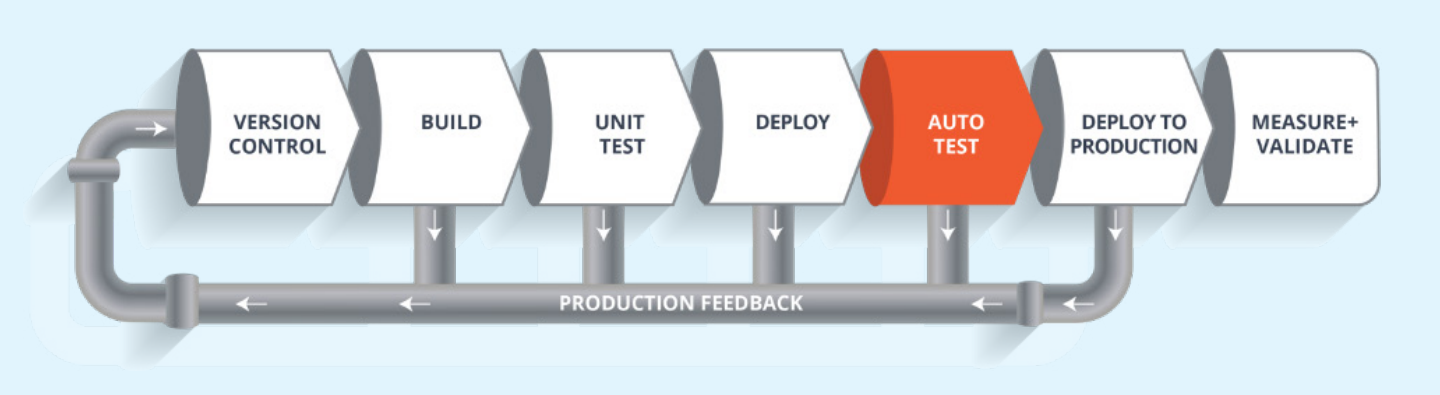
Concerning AI systems infrastructure, prepare to manage tens, hundreds, even thousands of experimental projects of varying scale. There is no one-size-fits-all infrastructure that picks all the locks and opens all the doors. There is no panacea, no completely turn-key solution. AI requires work in layers of application infrastructure built on large volumes of data. There are no fewer than hundreds of tools available for AI engineering across a variety of free and open source and paid licensed solutions from private software firms. Whatever the path taken to develop a formal production model, supporting members of the AI engineering team will analyze data many ways during the process. Once in production, the compounding effect of various deep learning applications will continue to extend the infrastructure. By using commodity hardware systems on-premises and considering a hybrid cloud infrastructure for experimentation, administrators can attain considerable flexibility and adapt to the changing demands of uncertain futures. Capacity with headroom will ensure that new models and new data are able to proliferate. A safe general guideline is to maintain a minimum 20 percent of additional capacity, whether on-premise or in-cloud, in order to have burst capabilities for new initiatives in machine learning and artificial intelligence.

Continuous Deployment and Automation

The final key concept in understanding the rapid development of any technology, especially those in the space of AI microservices development within a data fabric is continuous development and service deployment automation. The actual programming portion of large-scale systems development and deployment is a fraction of the entire delivery pipeline. This is important to note in light of the possible solutions that exist. Even COTS systems require continual development and releases to address bug fixes and the deployment of new features. It is therefore quite useful to understand the continuous deployment pipeline.

> > >

FIGURE A.3. - Unit of Code Scale Change



Source: The World Bank

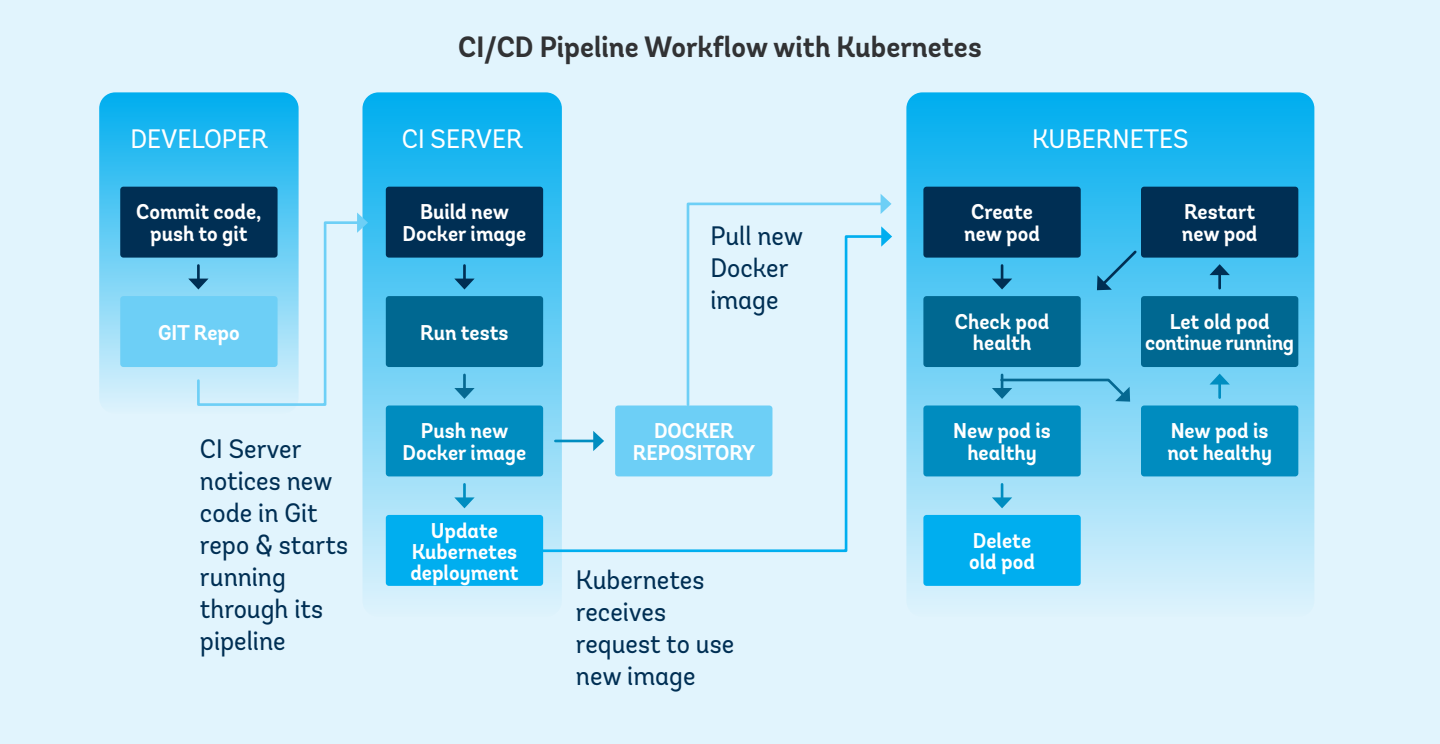
Essentially, any application within a larger architecture contains source code. Engineers specify criteria for the proper function of that source code in the form of various tests. Engineers store the code in source control systems. They then edit the code and commit changes to the central repository for their preferred source control system. The tests must run to validate the functionality of the application before new changes propagate to production. Every change requires that all tests pass in order for the change to be validated for release into the wild. Breaking changes prevent release. Passing changes merge into the master branch of the source code tree, and the product evolves. This is the continuous development pipeline.

Continuous integration (CI) systems automate this process so that engineers can work on the core of the product and continuously deploy code to production. The early days of manual testing are long gone for all modern software enterprises. Thus, the expectation today is that developers make new features and fixes in real time and deploy these to production without hesitation, sometimes several times a day among various teams managing various projects tied to the many applications supporting a microservices architecture.

To further the example, application containers offer engineers dependency management at previously unprecedented scale. Applications with dependencies for particular software modules with specific versions get packaged into small images called containers that are encapsulated to run in an isolated instance alongside many other application containers without overlapping dependencies. Thus, a small number of computing nodes can run a large number of application instances independently. This further reduces the complexity of dependency management among commodity infrastructures and lowers TCO over time. The initial investment of setting up this application deployment environment pays handsomely to teams with limited resources such as those of government agencies, their respective contracting consultants, and in-house technical management teams. There is little to deny the virtue of pursuing this course for any government wishing to develop a long-term plan for successful AI infrastructure. The following diagram illustrates the continuous integration and continuous deployment of an application using containers within Kubernetes, which is the gold standard in application container management among world-class software engineers and architects.

> > >

FIGURE A.4. - Continuous Integration and Continuous Deployment Pipeline Workflow with Kubernetes



Source: ReactiveOps

Brief History of Computing Architecture

Computing history is a history of the levels of logical abstraction. Early concerns with physical disk sectors gave rise to operating systems. Low-level languages gave rise to dynamic modern interpreted languages, thanks to layers of abstraction. The same applies to architectures leading to the boon in artificial intelligence. Before diving into types of architecture, it is worth rewinding through brief history to better understand how technology evolved.

As ancient computers became logical computing languages that gave rise to operating systems, a key principle of computer system architecture emerged; it was called the Single Responsibility Principle (SRP). SRP dictates that programs do one thing and do it well: work together, and handle text streams as a universal interface. In short, single-focus programs, when strung together, may perform a varied and complex assortment of tasks. A more detailed explanation may be found in The Unix Programming Environment, the book by Brian Kernighan and Rob Pike.

The engineering community largely forgot the SRP pattern in favor of the object oriented paradigm during the late 1980s and early 1990s, as the languages C++ and Java gained popularity. The promised vision of object orientation and code reuse was never fully realized due to ironic problems with polymorphism. This period in history also gave rise to monolithic systems that often had millions of lines of code buried in one executable.

The 2000s brought the introduction of network-aware applications and the application server model, which was characterized by large monolithic code bases, massive relational databases —with stored procedures for query optimization—and Common Object Request Broker Architecture (CORBA) and Common Object Model (COM) for distributed communication and application interoperability.[26]

The revolutionary 2000s also gave rise to XML[27] as a means to configure and communicate. The open standards community developed the Simple Object Access Protocol (SOAP) as a "superior" alternative to CORBA and COM. Because SOAP is text-based, it had better interoperability although it was still cumbersome compared to modern, gRPC,[28] JSON[29] and RESTful APIs.[30] Hence, the Software-as-a-Service (SaaS) gained traction among Web companies, and the industry began its shift toward the Web as the primary application service delivery mechanism.

Increasing pressure to deliver rapid iterations of both customer and internal-facing software systems led to the shift toward open-source software systems, led by organizations such as Apache and GNU/Linux.[31] This shift was an irreversible bifurcation that took power away from enterprise leaders and democratized innovation at the hands of hobbyists, start-ups, and academics at a previously unprecedented rate. Where once everyone waited on centralized policy-making, now the community offered unparalleled power and agility that led to the advent of cloud computing.

Amazon Web Services launched its Elastic Computing Cloud (EC2) in 2006, Google Compute Engine followed suit in 2008, and Microsoft Azure in 2010. In 2019, Amazon Web Services (AWS) reported revenues of $35 billion, indicating the extent of the seismic shift in the software industry, which sought out the most innovative software tooling developers for leadership and not the enterprise software vendors. As cost models shifted away from large up-front capital expenditures to lower ongoing operating costs, scaling and resources could be used and paid for on-demand, and the entire deployment stack transformed into a DevOps[32] infrastructure as code with the advent of CI and continuous deployment services.

Open source software and operational expenditure fueled a resurgence of the Unix Philosophy and gave rise to the microservices architecture: many small, fine-grained services that perform a single function all trying to achieve the goal of distributed networked components. Microservices gave rise to an engineering culture that embraces automated testing and deployment and embraces failure with unprecedented levels of fault tolerance. Microservices teams have the power to work on independent, deployable units of application code that are

26. The Common Object Request Broker Architecture (CORBA) is a legacy binary communication protocol that was popularized in the early 2000s. The Common Object Model was a Microsoft specification and alternative to CORBA. RESTful APIs and gRPC replaced both technologies.
27. XML – Extensible Markup Language
28. gRPC is a modern, open source remote procedure call (RPC) framework that can run anywhere. It enables client and server applications to communicate transparently and makes it easier to build connected systems: https://grpc.io/.
29. JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate: https://www.json.org/.
30. RESTful API design (Representational State Transfer) is designed to take advantage of existing protocols: https://restfulapi.net/.
31. The GNU/Linux operating system is free software that is an alternative to Microsoft Windows and macOS: https://www.gnu.org/.
32. DevOps is a set of practices that combines software development (Dev) and IT operations (Ops). It aims to shorten the systems development life cycle and provide continuous delivery with high software quality. DevOps is complementary with Agile software development; several DevOps aspects came from Agile methodology.

elastic, resilient, minimal, and complete. These applications scale individually and horizontally.

The underlying idea of microservices existed since the 1970s. Distributed systems are a permanent and enduring realization of the power of decentralized, democratized computing as a process and superset of products. Modern cloud infrastructures are built on microservices. Rapid, continuous integration and deployment pipelines are the reason for the overwhelming success of cloud computing platforms, where new features can move directly to production without human intervention after passing a stringent series of automated tests.

To conclude this brief excursion through computer architecture and summarize the experience, it is evident that the main driver for the evolution of computing architecture is speed of deployment. The demand to get code into production is unrelenting. Early results lead to rapid innovations regardless of the product. Cloud-native computing services open the clearest path to achieving the goal of rapid engineering and deployment at blistering rates.

Basic Components of AI Architecture

A nation's production of AI infrastructure requires that data centers be built on commoditized goods and services. The key to infrastructure cost savings and operating strategy is a commodity of goods and services. By definition, commoditized hardware is limited to components that are readily available at economies of scale for general purposes in computing—drives, racks, switches, routers, coolers, power supplies, etc. Highly successful data centers use commoditized hardware to maximize the procurement of parts for repairing equipment failures and performing system maintenance. Much effort goes into minimizing failure rates across large-scale computing infrastructure. Failure is unavoidable. Generally, data centers frown upon specialized computing infrastructure. Research shows that the type and utilization of commodity hardware can reduce failure by orders of magnitude. Commoditized hardware minimizes TCO, a key metric for financial viability of any data center. Analysts often use TCO for cost-benefit analysis when deciding whether to pursue cloud software systems such as AWS, Microsoft Azure, Google Compute Platform, or others.

Consider using cloud services wherever possible. On the upside, cloud services minimize TCO by multiples for operations of all sizes, especially during early phases of development. Cloud systems offer a variety of commoditized hardware and services. Cloud offerings range in complexity and computational power; customers may purchase bare, dedicated systems and turn-key AI solutions alike through an all-in-one interface. The downside, however, is the lack of ownership and geographical disadvantage that the location of data centers presents. A lack of data centers is not uncommon in nations with nascent computing industries, and likewise limited domestic control of cloud-based data infrastructure. Nations with data infrastructure located in foreign nations face the real potential for disruptions to corporate agreements due to unanticipated geopolitical tensions spurred by sanctions during periods of conflict. Although this is uncommon, strategic vulnerability remains a key reason for the slow adoption of cloud computing in government infrastructure among developing nations.

Consider a hybrid-native cloud services approach in order to maximize redundancy and protect data. Cloud services offer undeniable benefits and minimize TCO by offering comprehensive lists of commoditized services on demand. One particularly valuable benefit is the ability to extend on-premise infrastructure with dedicated cloud infrastructure. The reasons for developing the hybrid infrastructure model are primarily centered around redundancy and specialization. Data redundancy is essential to successful operations. Systems fail in all environments, without exception, and data are always at risk for total loss. The operating cost associated with archival storage may not be equitable in the long run as data volume increases. Similarly, databases and data processing systems often require redundant nodes to guarantee serviceability and failover and eliminate any SPOF. Scalability relies on the mitigation of these factors. A hybrid model offers effective strategic reserves for growing infrastructural demands by providing burst capabilities for adjusting to unanticipated demand during phases of growth. Furthermore, growth requires investment in innovation, and innovation requires specialization of novel services. Rather than develop new service infrastructure in an on-premise environment, cloud service providers offer a wide gamut of specialized artificial intelligence infrastructure suitable for experimentation using on-demand billing agreements. The incurred expense is limited to only what is used by the organization. In either redundant or specialized use cases, on-premise infrastructure gets extended over the network and organizations have the power to control the security and topology entirely.

General AI Architectures

Cloud computing originated with the need to run virtual machines on standardized hardware inside remote data centers—what is commonly known as Infrastructure-as-a-Service (IaaS). In the present day, cloud computing services span a vast array of on-demand services that address all sizes of computing tasks. Three major competitors dominate the global market for cloud-native services: Amazon, Google, and Microsoft. These provide customers with similar service offerings, which are listed in Table A.1.

> > >

TABLE A.1 - Cloud Service Counts Services

Service type	AWS	Google	Azure
Compute	14	8	17
Data and storage	13	12	12
Network	6	8	13
Developer	9	13	9
AI and Machine Learning	11	15	35
Other (e.g. IoT)	56	33	24
Total	109	89	110

Source: World Bank.

The service landscape continues to transform as new technologies enter the market, and ground-breaking work from publicly traded technology companies continues to evolve in the open source communities, especially the work in AI and ML. Therefore, the goal for executive leadership is to understand architectural principles and compose services into systems designed to achieve specific business goals. Systems may target general solutions, such as storage for AI experimentation, or specific siloed solutions, that detect a specific form of fraudulent activity within data streaming from a discrete source. By maintaining a general inventory of service types, practitioners can zero in on desirable results.

Overall, achieving the most effective AI IaaS model relies on understanding four pillars: Architecture, Development, Operations, and AI. To understand is to pursue the following respective inquiries with the goal of depth and breadth, underscoring a clear strategy of experimentation and execution.

- **ARCHITECTURE**: What are the architectural patterns for adopting AI computing infrastructure?
- **DEVELOPMENT**: What are the best development tools, frameworks, and best practices?
- **OPERATIONS**: What are the best practices to deploy and manage services in production?
- **AI**: What are the available ML/Data Services? How can problems be best solved with these tools?

> > >

FIGURE A.5. - Pillars of Effective AI Architecture

Architecture	Operations	AI	Development
Microservices RPA Protocols Messaging Queueing Events Data Models Cloud AI Services Computation Networking ERPs COTS	CI/CD Logging Monitoring Performance Analytics Databases Security	Data Science Deep Learning Chat Bots GANs NLP Text-to-Speech Speech-to-Text Machine Learning Preprocessing Prediction	Frameworks Tooling Debudding IDEs Technologies Cloud Service APIs Engineering

Source: The World Bank.

Function-as-a-Service Architecture

Serverless software is the state of the art in software service development today. There is no "official" definition of the term "serverless," therefore the following is adopted as a working definition:

"Serverless computing is a form of cloud utility computing where the cloud provider dynamically manages the underlying resources for the user of the service. It provides a level of abstraction over the underlying infrastructure, removing the burden of management from the end user."

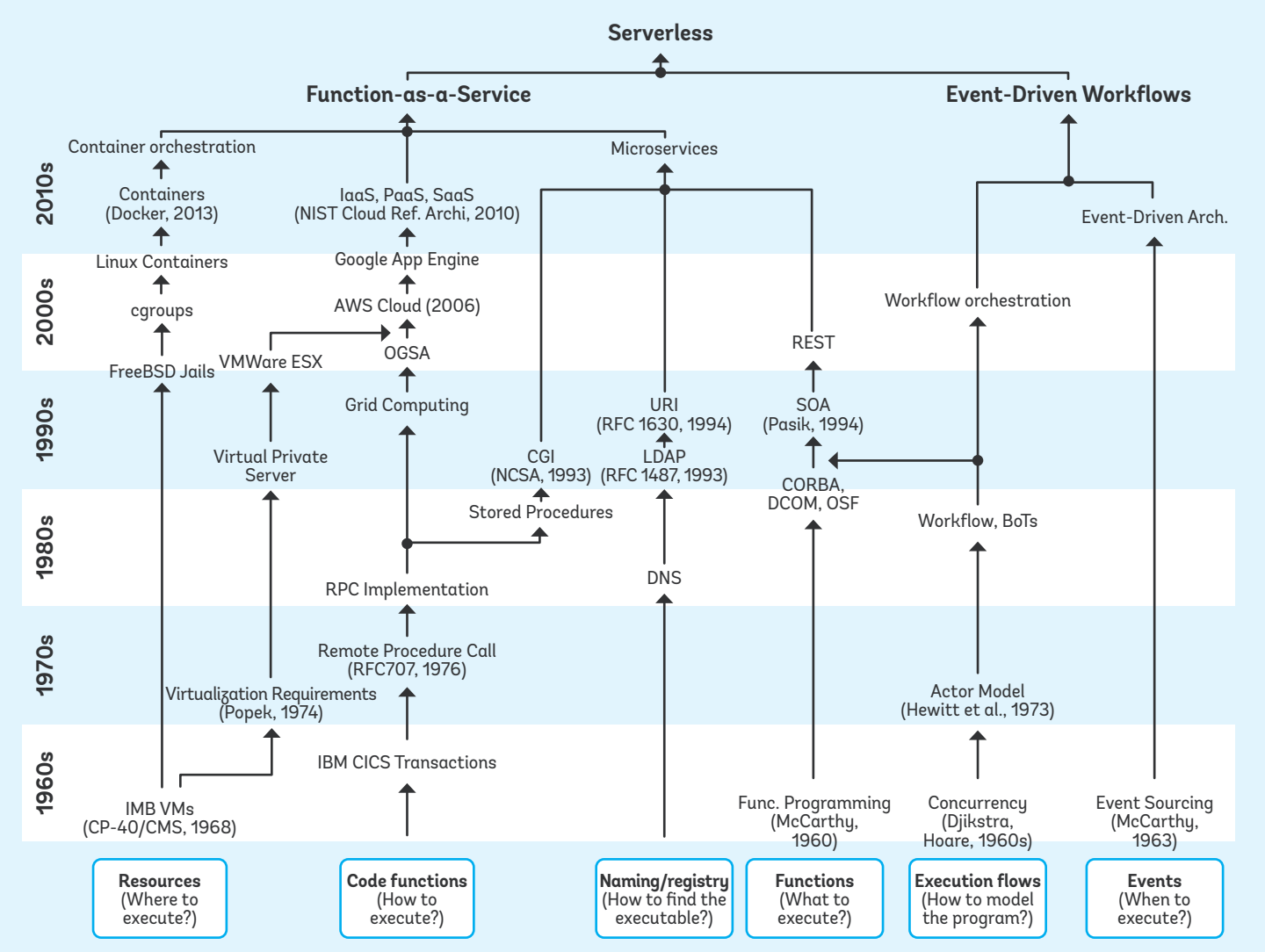
Serverless software avoids the explicit creation and management of infrastructure, such as servers and containers. Instead, functions that are managed and run by the cloud service provider replace these traditional computer resources. This category of cloud computing services is called Function-as-a-Service (FaaS), the overarching pattern serverless architecture. It is important not to be misled by the term serverless. Data stores and persistent layers are available to FaaS applications. The point is that practitioners no longer concentrate on the infrastructure because the cloud vendor provides a layer of abstraction over all underlying infrastructure.

The principles of FaaS (serverless) computing architecture can be summarized as follows:

- On-Demand cloud functions replace all servers and containers.
- Preference goes to managed services and third-party APIs over custom-built resources.
- Architectures are event driven and distributed.
- Engineers focus on developing the core product, not the low-level infrastructure.

> > >

FIGURE A.6. - Evolution of Architectures: A History of Computing Concepts Leading to Serverless Computing



Source: van Eyk et al. (2018).

Infrastructure-as-a-Service Architecture

Also known as cloud-native solutions, infrastructure as a Service (IaaS) is common to AWS, Azure, and Google Compute Engine. AWS offers a dedicated government infrastructure for qualifying customers. IaaS is based on the premise that infrastructure costs are significantly reduced when shared multiple tenants maximize rack resources in a data center. Tenancy does not affect security. Individual systems operate on virtualized machines referred to as Virtual Private Clouds (VPC). VPCs are isolated from one another; the end user "sees" them as physical machines when really the resources are constrained according to a virtualization policy specific to the customer's individual requirements and cost selections.

COTS AI Architecture

Most commercially available AI toolkits abstract the learning process with models developed for specific uses in a siloed environment. These pre-trained models require a specific dataset with custom features. Data inputs must have the prescribed features in order to realize accurate predictions. The end user simply needs to make known data available to a specific COTS product, and activities take off. This absolves the end user from possessing an in-depth knowledge of the underlying AI models, but still requires the end user to collect, clean, and provide as much data with relevant features as possible to the COTS solution.

Training is an important part of the AI process. Intelligence is the result of an emergence of outcomes that are trained and tested repeatedly over countless cycles of iteration depending on the model and methods employed. Most of the computational cost—and the biggest barrier to entry overall—lies in the fact that training requires a large volume of data processing on compute-intensive resources. Therefore, it is beneficial to approach new problems with an understanding of pre-trained AI models available from COTS and cloud service providers.

Cloud service providers discussed here have several applications and services available to attack common AI problems. These span a wide gamut of topics including document analysis, speech recognition, sentiment analysis, object detection, recommendation, and forecasting. Table A.2 lists common cloud AI services. This section of the Appendix discusses how to employ several of the services listed in Table A.2 to address notable problems in the final chapters, which contain practical examples of AI systems.

> > >

TABLE A.2 - AI Applications and Services

Application	Use	Service
Natural Language Processing	Machine Translation	AWS Translate
	Document Analysis	AWS Textract
	Key Phrases	AWS Comprehend
	Sentiment Analysis	
	Topic Modelling	
	Document Classification	
	Entity Extraction	
Conversational Interfaces	Chatbots	AWS Lex
Speech	Speech-To-Text	AWS Transcribe
	Text-To-Speech	AWS Polly
Machine Vision	Object, scene, and activity detection	AWS Rekognition
	Facial recognition	
	Facial analysis	
	Text in images	
Others	Time Series Forescasting	AWS Forecast
	Real-time personalization and recommendation	AWS Personalize

Source: Amazon Web Services

Cloud-Native Hybrid Architecture

Redundancy, persistence, and service access underscore the value proposition when considering whether to extend a native architecture into the cloud. All the major cloud services providers offer an assortment of general ML and AI services using an on-demand commodity, software-as-a-service (SaaS) model. The services include both software and hardware solutions for common problems in ML and AI such as image object detection, natural language processing, computer vision, and speech recognition. Although the on-demand price of services may appear to be a significant expense for any "24/7" solution, customers may pay only for the time and computation they utilize for a specific task, experiment, or stage within the project lifecycle.

A cloud-native hybrid solution allows discrete access control and infrastructure integration. Overall, the physical data center can extend its topology with a virtual topology in the cloud called a virtual private cloud (VPC) that is nearly identical to a Virtual Private Network, governed by access and security policies that the government's Development Operations (DevOps) Manager has control over. The extension behaves as though it is on-premises. Whitelisted infrastructure communicates internally, using private DNS network addresses. DevOps administrators may enforce one or more firewall proxies to grant access to vetted external components and services.

Implementation occurs in several phases depending on the desired objectives and key results. First, the team determines the purpose of external cloud services. If there is a need for bespoke AI compute services, then managers may surmise estimates by using existing local development processes that may address computational shortcomings. Per the information in the previous section on cloud architecture costs, from these shortcomings, DevOps Administrators may estimate the hourly TCO of cloud services based on the anticipated service requirements. Similarly, if there is a need for storage redundancy, then managers will estimate storage durability (level of redundancy) and availability (time-to-access), based on the current footprint and operating requirements. With anticipated estimates in hand, budgets may be appropriated, and resources deployed according to the prescribed needs of the project. DevOps will manage the deployed resources and ensure that operating requirements are adequately resourced, which admins may choose to automate using cloud management tools for the long run.

By using cloud management tools, the size of cloud-native services hybrid architectures can expand and contract automatically, on-demand. Although it requires an investment of time to develop a formalized topology of services and storage nodes, the benefits of investment pay off in spades over the long run. Once services are operating as planned, depending upon their intended use among citizens (public) or agencies (private), a configuration and template system will generally manage the scale of the infrastructure operation. Typically, a YAML[32] (YAML Ain't Markup Language) document will contain the topology requirements for the entire system and individually within containers that make up the constituent services.

Several containerization and instance management solutions are available through the major cloud service providers, but many of the best solutions are Free and Open-Source Software (FOSS). In particular, Kubernetes (K8s) paired with docker containers is among the most well-regarded solutions for full-scale application infrastructure management. Docker Containers are lightweight disk images containing an application—and all software dependencies—configured in a fully operational state. Containers will deploy on any computer (node) running docker software. Docker reduces deployment time, eliminates system dependency management, and allows nodes with different operating systems to run "dockerized" applications stored in docker containers. Capital allocated to DevOps and DevOps stretches much farther with a managed cloud application cluster. K8s deployments are clusters of nodes running dockerized service applications that employ easy-to-use configurations to scale with minimal human intervention.

Cloud-Native COTS Hybrid Architecture

Starting small is adequate for long-term proliferation of successful solutions, yet there are hybrid alternatives with commercial off-the-shelf (COTS) solutions that may extend AI capabilities. There are several commercial large-scale systems for transactional accounting and financial audit available. Many rely on proprietary cloud infrastructures in foreign data centers. This places significant barriers to entry for government teams facing long-term goals of developing a convergent data infrastructure on-premises. Governments that consider making an investment in small-scale development once a broad data infrastructure strategy is in place may have a higher likelihood of long-term success. By edifying a formal iterative agile process, small-scale projects can spiral upward through versions. The key to iterative development is failing fast and often: projects that invest long time-spans to realize products at any scale become burdensome and fail to garner enough momentum to endure or provide value in the face of changing economic and political landscape. Thus, it is important to start small and scale with experimentation through iteration in order to prove the effectiveness of novel solutions, especially those in artificial intelligence.

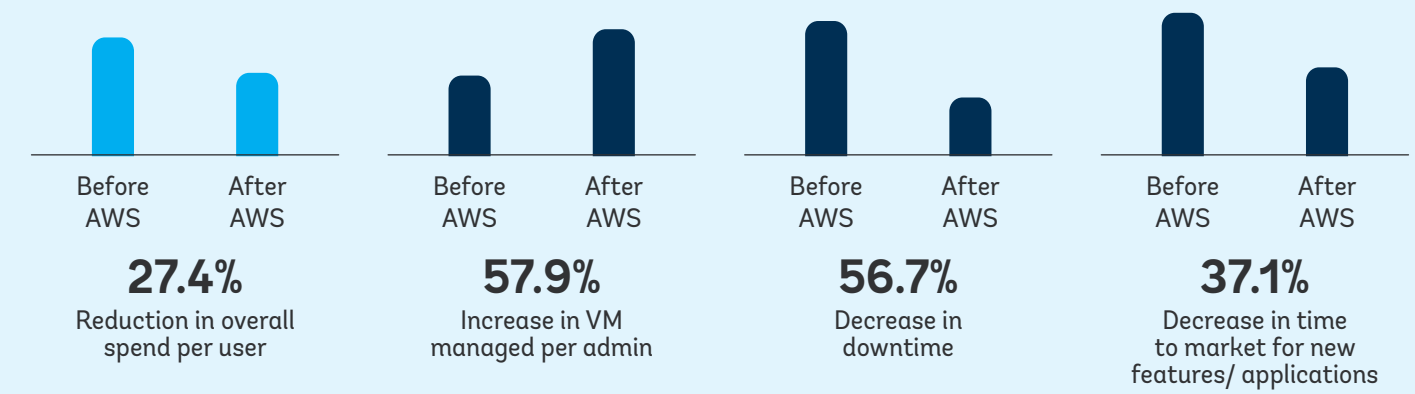
32. YAML ("YAML Ain't Markup Language") is a human-readable data-serialization language. It is commonly used for configuration files and in applications where data is being stored or transmitted.

Turn-key solutions may present quick and easy wins, but need careful vetting to prevent unnecessary technical debt from sneaking up. More nuanced than mere costs associated with cutting corners, technical debt is a hidden intrinsic cost of technological development, which emerges as bugs, unfinished tasks, improvements, features, and upgrades that accompany the process of engineering any software or hardware product. Technical debt is inherently unavoidable and impossible to eliminate. Yet, there are process management best practices for mitigating the risk of extensive technical debt. During initial stages, careful planning and scoping are the best measures for maximizing productivity without incurring debt. But some things cannot be anticipated, so it is commonly acceptable to commence in a small scope that addresses the key concepts that underpin a full-scale long-term solution.

Measure the costs associated with a cloud computing architecture in terms of TCO, which takes into account more than the cost of engineering and implementation. TCO factors in the cost of personnel, heating, ventilation, and air conditioning (HVAC), maintenance, monitoring, hardware, software, land, facilities, electricity, and innovation. By leveraging on-demand Infrastructure-as-a-Service (IaaS), projects forgo the cost of brick-and-mortar data center underutilization. Planning calls for expected server capacities that are guaranteed to fluctuate due to regular cycles of use on a daily basis. Moreover, although early projects demand fewer resources, planning for lateral growth to accommodate new deployments places a burden on resources that may naturally underwhelm the overall server infrastructure, leading to idle systems that demand step-wise investments for anticipated future demands. The overall efficiency of cloud-native architecture exceeds on-premise systems significantly, as illustrated by the graphs and statistics below, provided by AWS.

> > >

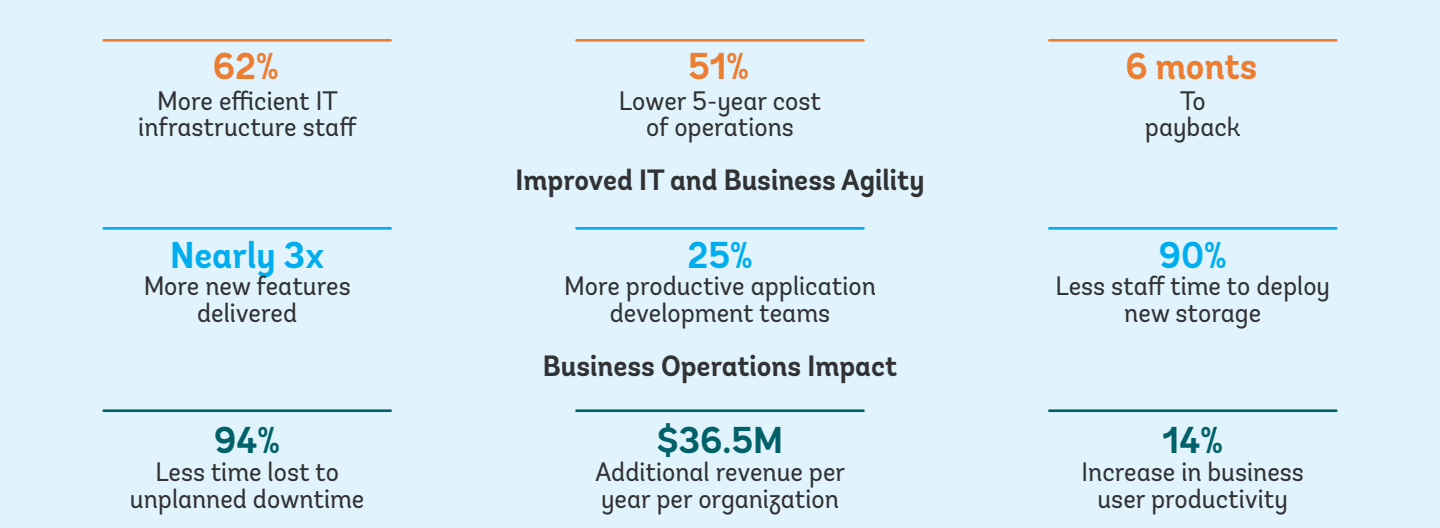
FIGURE A.7. - Overall Efficiency of Cloud-Native Architecture



Source: Amazon Web Services

> > >

FIGURE A.8. - Optimizing Cost of Providing IT Services and AWS Value



Source: Amazon Web Services

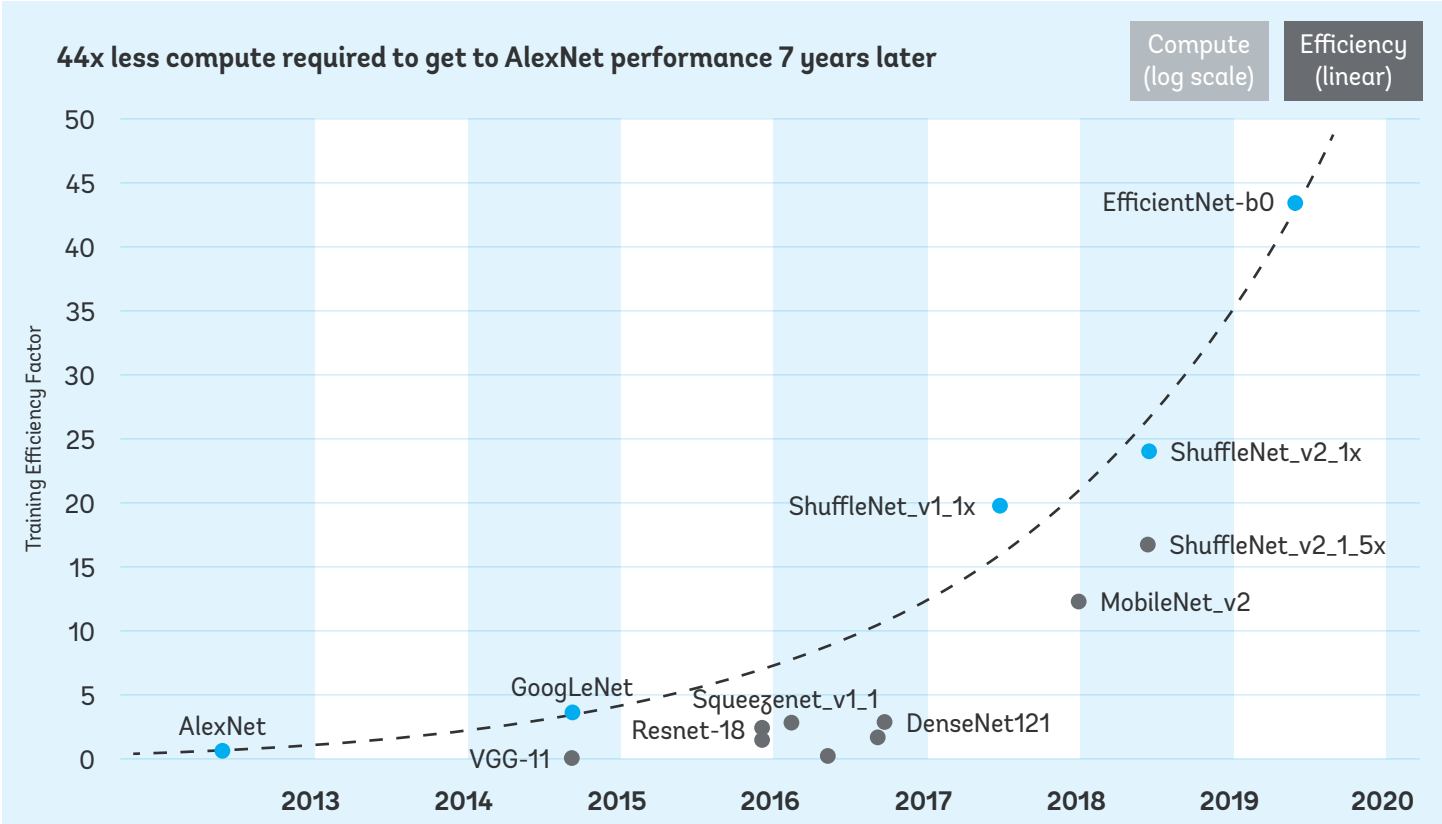
Cloud-native costs fall into one of four contract categories: reserved, partially-reserved, spot, and on-demand. Reserved instances require up-front payment for a period of one to three years with no additional monthly costs for central processing unit (CPU) and memory utilization. Partially-reserved instances require a partial payment for one to three years but require a reduced monthly service cost for CPU and memory. Both reserved options are contract-based solutions. Contracts can be sold at a rate prorated to the remaining duration of the contract period if a customer should deem the contract unusable. Contract options are also specific to machine types, which often ship with immutable memory and CPU configurations. Spot instances offer significant savings similar to reserved instances, but their availability is not guaranteed. Spot rates allow the customer to specify the maximum allowable cost per hour of VPC use for a given VPC configuration. Customers pay a reduced variable cost for the instance, but should the market cost exceed the customer's maximum, the instance can terminate, causing potential loss of data. The key is to set a high cost threshold and the VPC remains protected. Additional configuration can allow for persistent disk mounts that protect volumes of information from loss in the event of an unexpected termination. Spot instances are best for experimental development and skunkworks usage in which there are no risks to the general public. Lastly, on-demand instances offer no savings compared to reserved instances, but still remain very competitive with TCO of on-premise deployments. The market determines the on-demand rate.

Also, important to note, disk drives (volumes), network input-output (I/O), monitoring, and dedicated VPCs are additional costs on top of the VPC cost in a cloud-native infrastructure. These are metered in fractions of a unit of computational payload (in bytes) and sold as add-ons, which still offer significant savings and added efficiency over the on-premise model.

The overall reduction in costs is compounded by the increase in efficiency of AI algorithms, which are outpacing predictions made by Moore's Law, which states that the number of transistors on a microchip doubles about every two years, though the cost of computers is halved. This leads to exponential increases in computational power. Coupled with the fact that research in AI algorithms is increasing their efficiency, AI is outpacing Moore's Law faster than expected. Figures A.9 and A.10 below illustrate the fact; the first is efficiency, and the second is compute according to a study conducted by researchers at OpenAI.

> > >

FIGURE A.9. - Less Compute Required to Get to AlexNet Performance 7 Years Later – Efficiency Level

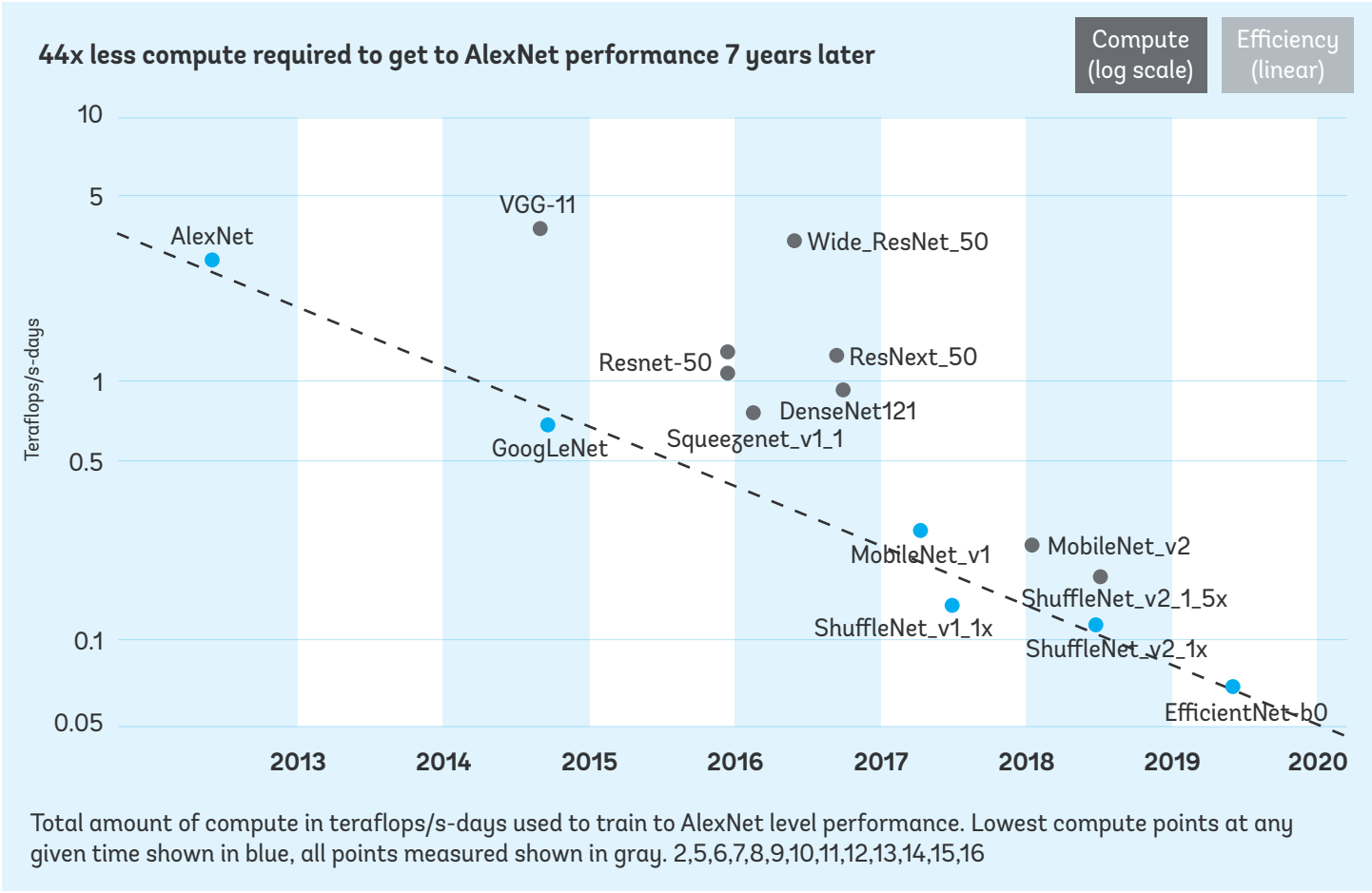


Total amount of compute in teraflops/s-days used to train to AlexNet level performance. Lowest compute points at any given time shown in blue, all points measured shown in gray. 2,5,6,7,8,9,10,11,12,13,14,15,16

Source: OpenAI

> > >

FIGURE A.10. - 44x Less Compute Required to Get to AlexNet Performance 7 Years Later – Compute (log scale)



Total amount of compute in teraflops/s-days used to train to AlexNet level performance. Lowest compute points at any given time shown in blue, all points measured shown in gray. 2,5,6,7,8,9,10,11,12,13,14,15,16

Source: OpenAI

Overall, the TCO for cloud-native and hybrid infrastructure makes a strong case for consideration in government systems, if only during planning and research phases of new initiatives, even after a government deploys the core on-premise infrastructure.

The necessity of a common data interchange standard as well as compact and accessible interfaces in the design of an AI capable infrastructure cannot be understated when pursuing the task of access to clean data processing pipelines.

Global Data Interchange

Plan globally for data interchange the long run and act locally for data processing in the short term. The coordination and management of a digital infrastructure requires conscious effort to identify existing requirements and plan the digitization of the most mission-critical systems. These systems may not have a national scope. They may be localized operations devoted to managing reporting under an integrated financial management information system, procurement, or asset management and exchange. Whatever the entry point, stakeholders must carefully consider the design of the underlying architecture early during the planning phase with a focus on accommodating long-term use of data to satisfy secondary operational requirements. Health, education, transportation, and public safety systems are examples of secondary operational requirements in a digital government environment.

Efficient Communication with Compact Data

In the early days following the shift from mainframe to distributed systems of computing, engineers began to address IPC between applications, computers, and data centers with often creative solutions. At the smallest scale, data communication between applications using XML and SOAP protocols allowed for independently specialized applications to share information somewhat effectively. At a very large scale, immense data transfers required the physical transport of magnetic reels and hard drives over land to mitigate the total cost of network transfer. Today's standards may require the occasional physical transport, but among emerging data technologies at massive scale, IPC is managed efficiently in real time using compact data standards and communication protocols. Two standards in structured data stand out above others: JSON and protocol buffers - also called protobuffers or protobufs.

JavaScript Object Notation

JSON is an object definition "language" standard that gives the practitioner the ability to define key-value relationships between any number of values, which may be primitive types such as strings, numbers, and Boolean values or complex types such as arrays and nested JSON objects. Engineers and data scientists refer to one of these comprehensive and completely self-contained units of information as a document. The attributes and values within a document are iterable—programs can "walk" the document to retrieve values—and mutable—programs can alter the values of the attributes.

JSON documents are the primary structure of document storage in several of the most successful databases and big-data storage solutions on the market. JSON is also the preferred format for data exchange among web services architectures throughout the world of software development. There are international standards for the structure and definition of JSON documents. More information about JSON is available at www.json.org, and many other resources exist that cover the subject matter exhaustively. The discussion of JSON in AI Architecture continues in the section on Leveraging Microservices.

It will suffice to write that JSON provides a very efficient IPC standard for virtually any application that will ever be engineered. It is fast, compact, semantically endowed for human consumption, and provides a low barrier to entry for practitioners in need of rapid interchange of data between specialized applications. In some instances, internal applications require even more performance, less readability of payload while maintaining semantic interoperability. This leads engineers to consider protobuffers.

Protocol Buffers

When speed of interchange and consistent structure is crucial for mission-critical applications—such as those in finance or infrastructure management—protobuffers provide a valuable alternative to JSON. Protobufs are platform-independent, language-independent extensible mechanisms for serializing structured data. Once in a document schema, developers structure the data, and any applications wishing to communicate with that data can simply implement an API that is automatically generated in any programming language on any platform.

A specialized remote procedure call framework further extends the power of protobuf's compact data interchange format with structured programmatic function definitions called gRPC. With gRPC in place, IPC occurs over any network topology by leveraging exposed functions capable of ingesting and outputting protobufs. This means that highly specialized and compact application services can be built to communicate in real time and process large volumes of information for large scale implementations of AI services. This is the technology at the heart of Google's global infrastructure. The standards and software supporting this technological breakthrough are FOSS.

Other Data Formats

Other standards in data do exist. Many are proprietary to the systems that leverage that data. One notable data format is called Parquet—a FOSS construct of the Apache Software Foundation. The purpose of Parquet was to provide a columnar data access format that interoperates well with Hadoop software systems. Hadoop came to prominence in the 2010s after Google Published a MapReduce white paper describing in detail the design and development of the original architecture that powered PageRank algorithms to notoriety. The report was reverse-engineered, and an open source MapReduce solution hit the market leading to a trend in Big Data technology, which never fully panned out for Hadoop and its consortium of supporters.

As a modern relic, Hadoop (and Parquet) technology proves primarily that industry hype can mislead practitioners in search of problems looking for a solution. In contrast, the more simple, streamlined, and effective long-tail solutions and patterns of application and data architectures continue to satisfy the requirements for modern Big Data best practices.

Leveraging Microservices

As mentioned, monolithic systems have critical faults that lead to eventual collapse, for reasons of obsolescence stemming from stifling complexity. Microservices, conversely, are a methodology of designing, architecting, and developing a wildly scalable infrastructure of highly specialized applications. Engineering teams focus on each application independently, while inter-process communication, especially when leveraging the power of gRPC, remains versioned through API standards. Thus, project management dependencies are limited to the scope of each independent component within the microservices application infrastructure. Teams can and often do work independently to achieve incredibly rapid results for very large scale systems. Thus, the future of AI systems engineering and data fabric infrastructure rests on this fundamentally advanced pattern of application architecture development regardless of the course of physical deployment, be it in the cloud, on-premise, or a hybrid of both.

A large volume of information exists on the subject of microservices development. This appendix to the paper does not delve into the subject further, but rather uses this mechanism as a talking point to illustrate the necessary jargon that is essential in understanding the factors allowing for the development of solutions.

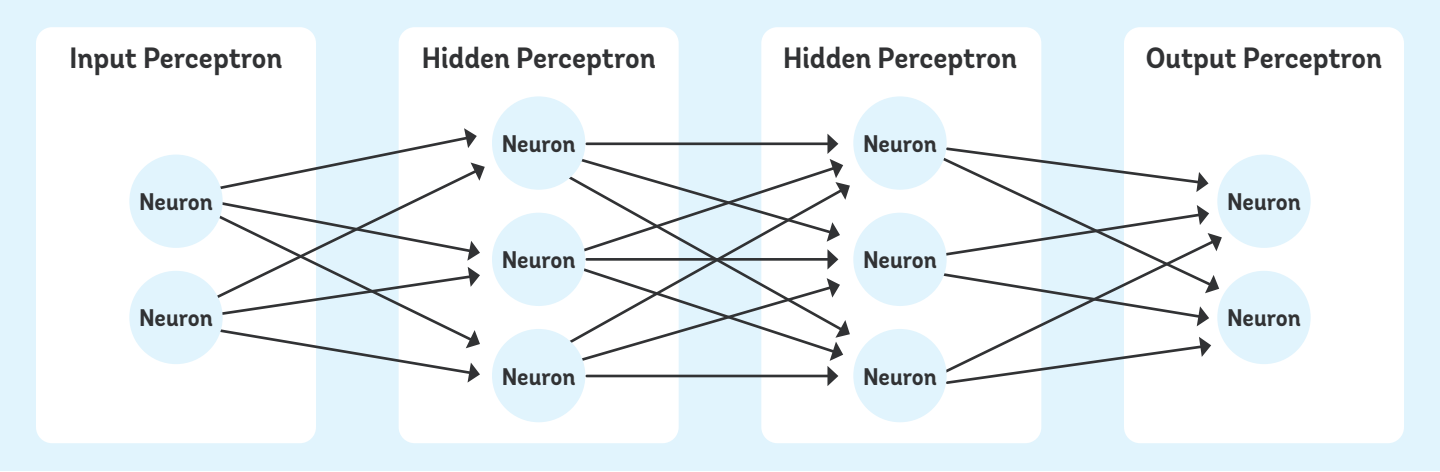
Advanced AI Models

Artificial Neural Networks

Neural networks are at the heart of advanced concepts in AI. Neural networks perform computations that derive potentially vast sets of self-selected features. Deep Learning relies on artificial neural networks (ANNs). First studied in the 1950s, ANNs have emerged today through several cycles of dormancy due in large part to the copious amount of raw computing power available in the cloud. At their core, ANNs are organized layers of decision nodes called perceptrons. Numbers enter an input layer and exit through an output layer. Hidden layers exist between the two. The goal of ANNs is to iteratively learn weights for each perceptron layer and produce an approximation of the desired result in the output layer. “Deep” refers to the number of hidden layers in an ANN, which may be as few as seven to eight but most often hundreds. Figure A.11 represents the basic ANN structure.

> > >

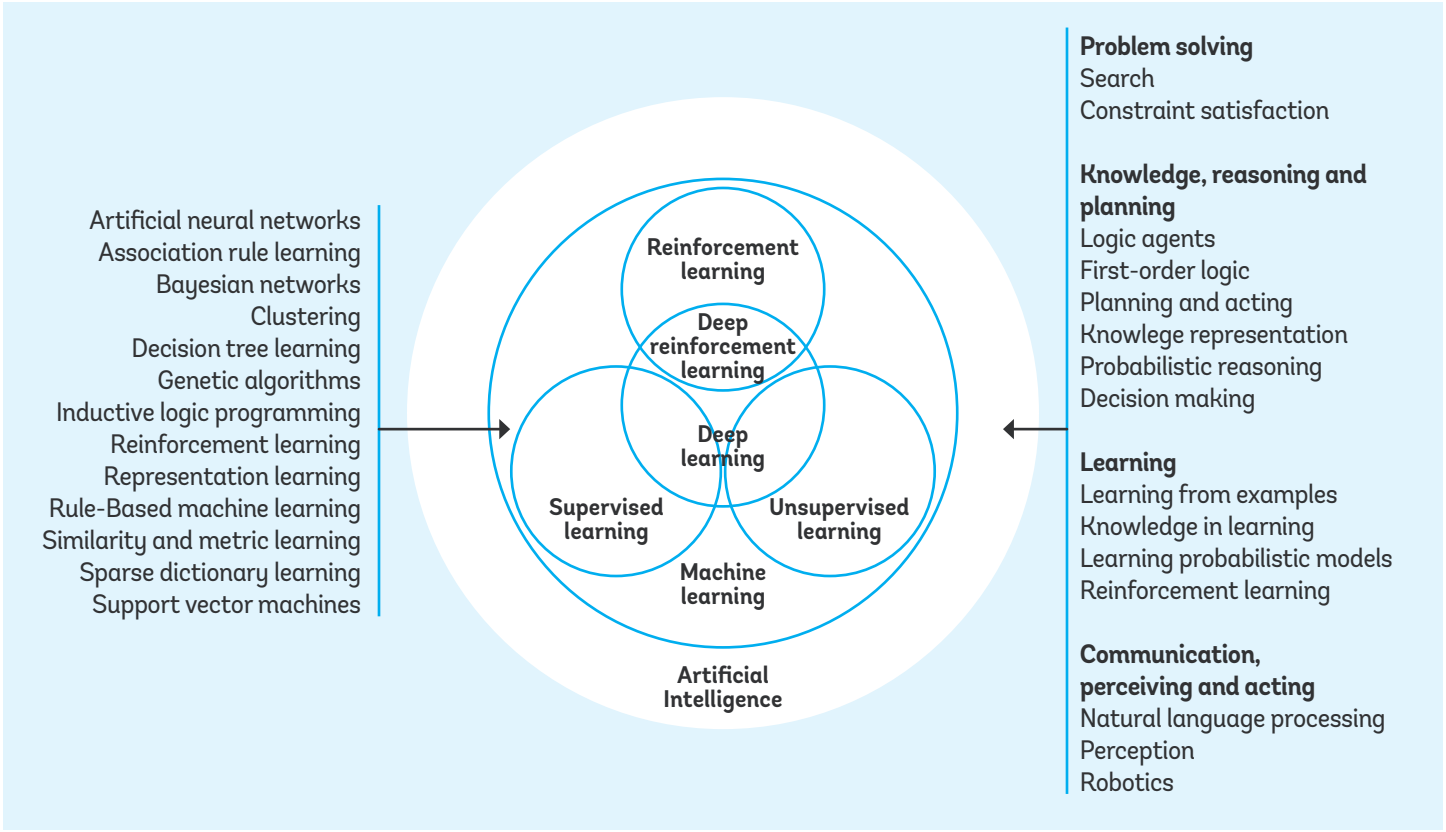
FIGURE A.11. - Basic Deep Neural Network Structure



Source: The World Bank.

> > >

FIGURE A.12. - AI and Machine Learning Algorithms and Applications



Source: Peter Elger and Eoin Shanaghy, AI as a Service, Manning 2020.

Natural Language Processing

Language is the most powerful and potent human mechanism. With great access to language data comes great responsibility. Today's commercial email, word processing, and voice communication tools are constantly scanning and interpreting human language with a goal of suggesting grammatical corrections, advertisements, and translating our conversations into written language. Smartphones and smarthomes alike respond to words, sometimes when a passive conversation is "overheard." The annals of news reporting are at the mercy of suggestions catered to individual indulgences. At the heart of all this is NLP.

Since around 2013, NLP and chatbots have gained presence nearly everywhere in society at large. Google search became smarter and more capable of interpreting more human-like inquiries. Smartphone auto-correct and auto-complete followed suit, and the emergence of personalized phone assistants began to gain traction. In government, NLP began to emerge as a tool for combating corruption and giving a voice to citizens. One project called Hack Oregon used natural language campaign finance data to find connections between political donors because it seemed that politicians were hiding their donors' identities behind obfuscating language in their campaign finance filings.

> Language is the foundation upon which we build our shared sense of humanity.
>
> *Dr. Arwen Griffioen, Senior Data Scientist - Research, Zendesk*

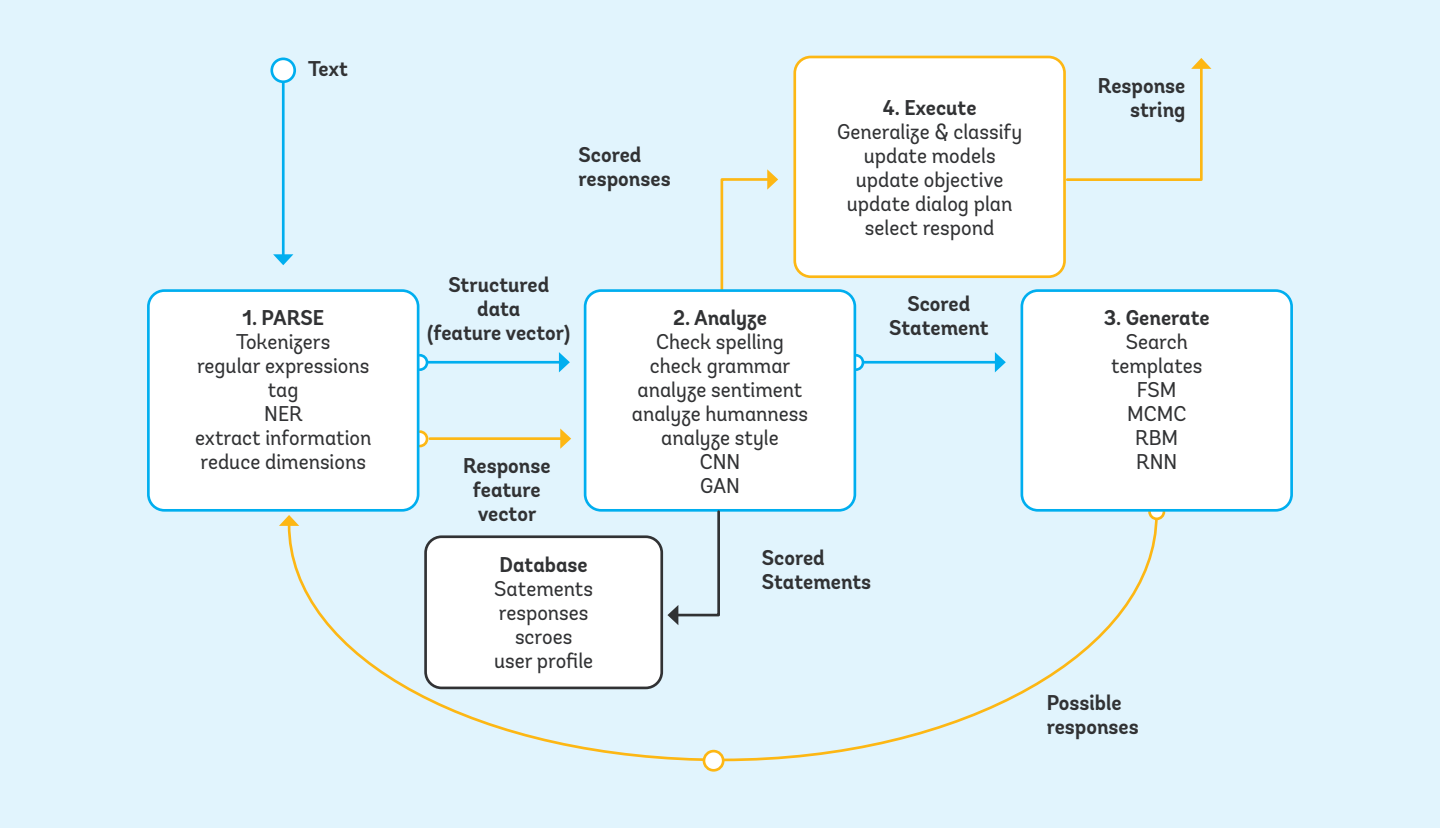
Basic NLP systems track term frequency relative to inverse document frequency (TF-IDF). These evolved to "chain" clusters of word frequencies in order so that predictions could be made about the best "next" word, also called Markov chains. These conditional, probabilistic distributions have evolved since into very sophisticated systems of interpreting, "understanding," and formulating language into topics with semantic meaning using math alone. Fascinating barely begins to describe the power of NLP.

In government and beyond, the necessity of beneficial machines with prosocial be-

havior that leads to greater cooperation among actors remains a key focus of ongoing NLP research. Governments are able to leverage NLP for interfacing with citizens for many purposes, the least of which is gathering information about the quality of service within the government. As news sources become increasingly aligned with the indulgences and personal preferences detected among patrons of various internet service providers, the quality of information revealed to citizens in relation to the government also falls into question. A simplified process of NLP operations in the AI algorithm is depicted below in Figure A.13.

> > >

FIGURE A.13. - Chat Box Recirculating (Recurrent) Pipeline



Source: Lane, Howard, and Hapke (2019).

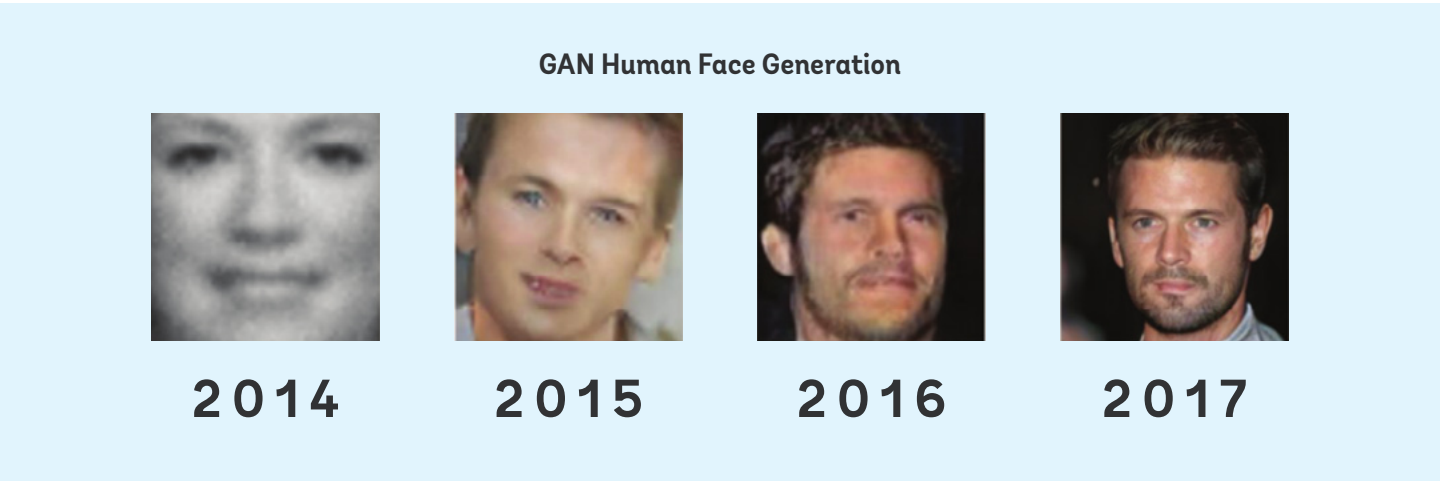
To highlight the many use cases, NLP makes it possible to review contract submissions, resumes, proposals, campaign advertisements, published documents, and financial transactions for authenticity with minimal bias. The mathematical models that enable these technologies to perform such important tasks fall outside the scope of this paper. Instead a quick exploration of how an NLP architecture operates precedes later sections that enumerate the examples of NLP in action within government. NLP is among the most interesting topics in AI that will make a lasting impact on the way in which human beings interact with computers, organizations, the environment, and each other for decades to come.

Generative Adversarial Networks

Generative adversarial networks (GANs), introduced to the AI ecosystem in 2014 by Ian Goodfellow, enable computers to generate realistic data by using two separate neural networks. Although these were not the first computer programs used to generate data, their results and versatility set them apart from all the rest. GANs achieve remarkable and often alarmingly convincing results that were previously considered virtually impossible for artificial systems, such as the ability to generate fake images (and videos) with real-world quality. GANs can turn a scribble into a photographic image or turn video footage of a horse into a zebra—all without the need for incredibly large painstakingly labeled data. A staggering example of how far machine data generation is able to advance because of GANs is the synthesis of human faces—see Figure A.14.

> > >

FIGURE A.14. - Progress in Synthetic Human Face Generation, 2014–2017



Source: Brundage et al. (2018).

By 2017, GANs enabled computers to synthesize fake faces rivaling high-resolution photographs. Most notably, GANs produced fake videos of notable celebrities and political figures whose speech and countenance are virtually indistinguishable from real life recordings simply by "mutating" the face of any recorded individual to appear as the synthesized individual, as shown in Figure A.15. This is of particular interest to government policymakers due to the fact that fake-news videos can be produced and proliferated by anyone with access to GAN modeling toolkits in order to misinform and manipulate the public with practically any video content imaginable.

> > >

FIGURE A.15. - GAN Transformation of One Politician into Another[33]

Source: Bansal (2018).

33. Watch the video: https://www.youtube.com/watch?v=F51RCdDIuUw.

GANs are a class of machine learning techniques that consist of two simultaneously trained models competing with one another as adversaries: one (the Generator) trained to generate fake data, and the other (the Discriminator) trained to discern the fake data from real-world examples.

The term "generative" refers to the overall purpose of the model: to create "new" data. The data that GANs learn to generate depends on the choice of the training set. In the example mentioned above, if a practitioner wants a GAN to generate images that look like the president of any country, they will use a training dataset of the president's face.

The term "adversarial" refers to the game-like competitive dynamic between the two models that constitute the GAN framework. The Generator creates examples that are indistinguishable from the real-world data in the training set: fake images of the public figure. The Discriminator verifies the authenticity of the images believed to be the president. The two networks are continually trying to outwit each other: the better the output of the Generator, the better the Discriminator needs to be at distinguishing real examples from the fake ones. The term "network" indicates the class of machine learning models most commonly used to represent the Generator and Discriminator: deep neural networks. The complexity of the artificial neural networks employed varies from simple to extreme and the results are unimaginably concerning for policymakers interested in preserving public trust and ensuring the safety of representatives of government charged with protecting national security.

GANs is explored further in the section about AI in policy. Before progressing to the topic of general artificial intelligence, it is worth noting that technological advancements in AI also enable concerns with voice synthesis in addition to image synthesis. Present day AI technologies allow the mimicry of human speech with relative ease. Thus, with moderate effort, AI models can produce human speech samples that are practically indistinguishable from actual human voice, furthering the concern over influence due to intentional disinformation spread through social media and the news. All hope is not lost, however, with the introduction of authentication mechanisms that practitioners can implement to prevent the loss of integrity for state-sponsored messaging using asset encryption and cryptographic signing, which produces a digital watermark using state-sponsored media. Despite this possible solution, as AI methods continue to improve, the need for more robust authentication and prevention mechanisms will accelerate in order to keep pace with more advanced methods of image and audio forgery.

General Artificial Intelligence

The ultimate goal of artificial intelligence is to emulate the intelligence of humans and animals by modeling the behavior of neural pathways and the brain. General artificial intelligence takes that goal one step further by pursuing the ability to learn how to learn. The mention of General AI conjures visions of a singularity and the domination of mankind by sentient machines. This is fodder for science fiction and cinema. Learning to learn—sentience—is beyond AI's current capabilities. Presently, all AI practitioners operate within the confines of artificial methods of guided learning and modeling based mostly on advanced statistical models built to process vast amounts of information in order to assist with specific decision-making goals. They cultivate interpretive data models that train computers to provide sound decision-making similarly to humans, by emulating the physiological design of the brain. General AI is a proverbial mecca on the AI horizon that aims to eliminate the need for human influence over the learning process. There are no known instances of this phenomenon in current employment among AI practitioners, although researchers have made significant contributions to reach this ultimate goal. There is no doubt that current AI resources will be instrumental in the emergence of General AI, however, the timeline for the realization of the singularity is uncertain. Therefore, it is outside the scope of the paper to explore this topic any further, however many resources exist for those interested in learning more about General AI.

Real World AI Workflows

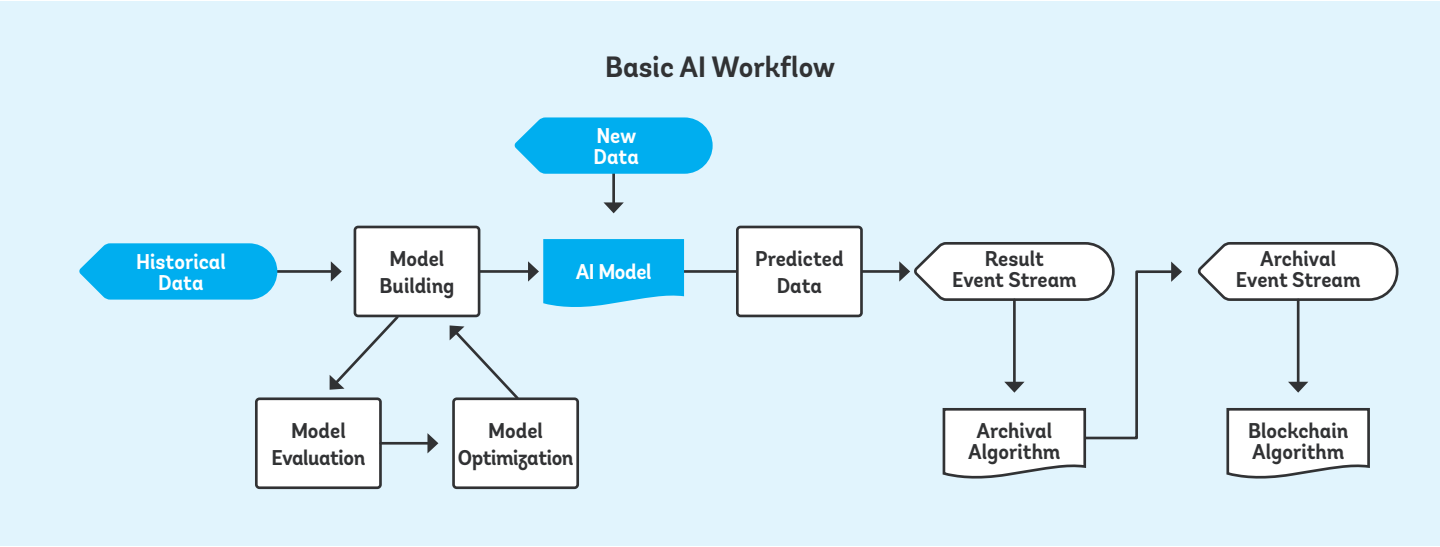
The real world AI workflow has five main components: data preparation, model building, evaluation, optimization, and predictions on new data. Although applying these steps has an inherent order, most real-world applications revisit each step multiple times using an iterative process. Practitioners first build a model using historical input data from a particular ML algorithm. Next, they iteratively evaluate model performance and optimize for accuracy and scalability to fit the project requirements. Last, they use the final model to make predictions on new data inputs to the system. Historic data helps build the model, and new data flows into the resulting AI model to create predicted data. Predicted data flows into data streams that may be useful in applications for additional computational workloads and eventual archival storage in distributed ledger technologies (DLT). This appendix touches upon DLT—a useful tool in combating long-term data tampering—in later sections.

Human-Out-of-the-Loop Workflow

The basic AI workflow, absent of human oversight, is also referred to as an Out-of-the-Loop workflow. This simply refers to the fact that humans do not evaluate the predicted outcomes before applications take additional action. It is worth noting that this is a simplified representation of an AI system, and the following human-out-of-the-loop table does not account for the steps one must take to optimize the AI model building process, such as feature engineering and model tuning. Overall, an Out-of-the-Loop approach is a useful way to approach non-critical decision-making systems such as recommendation engines and general classification engines—see Figure A.16. For mission-critical applications that result in consequential collateral actions—the detection of fraud and other criminal acts—practitioners must employ advanced AI workflows with human intervention built into the AI loop.

> > >

FIGURE A.16. - Basic Out-of-the-Loop AI Workflow



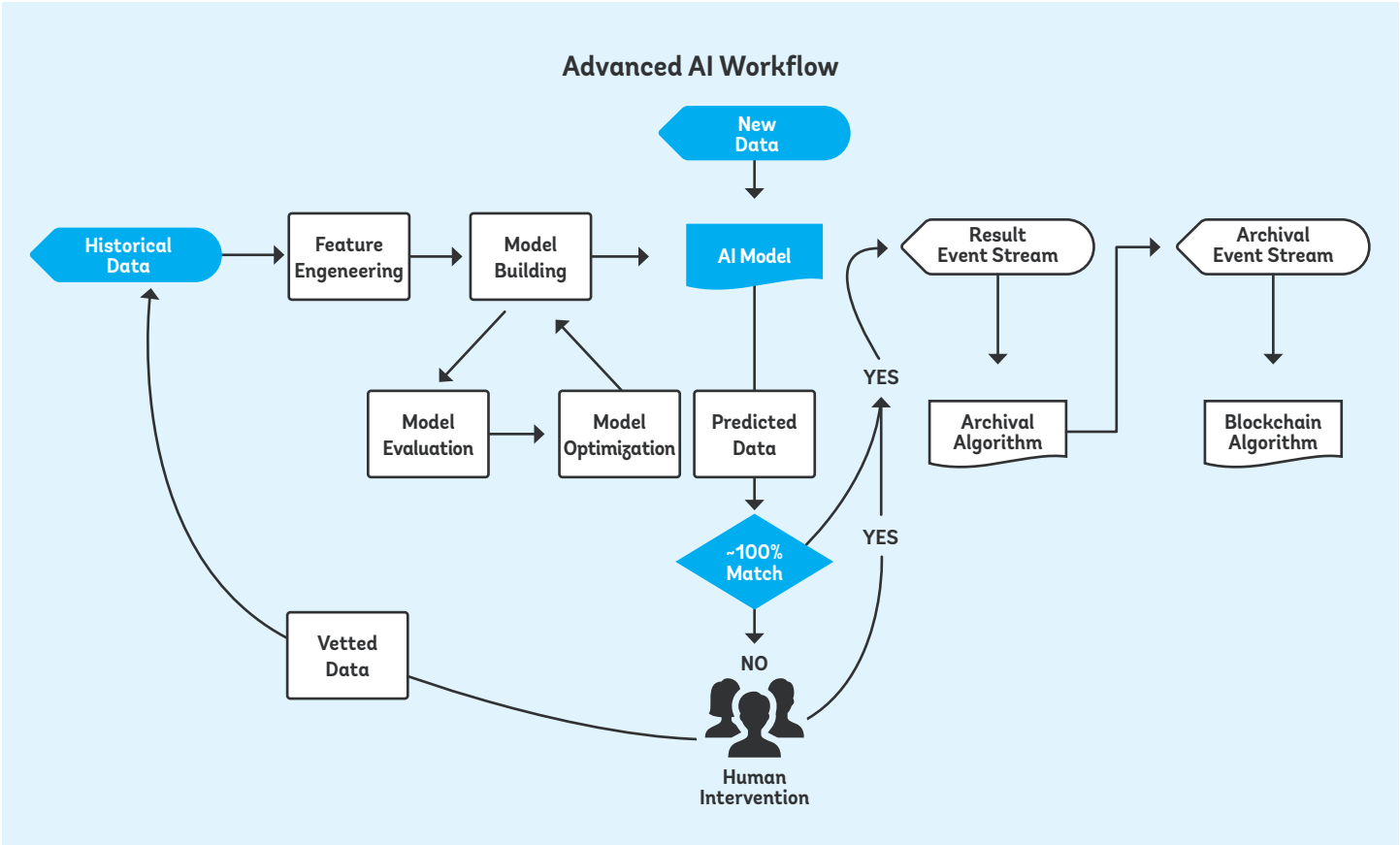
Source: The World Bank.

Human-in-the-Loop Workflows

Of the many forms of advanced AI workflows, the simplified—yet advanced—flow depicted in Figure A.17 has the same components of the basic workflow, but with an additional human element that improves model performance and prevents unintended consequences in mission-critical systems. Workflows with human intervention loops are referred to as human-in-the-loop and human-over-the-loop workflows.

> > >

FIGURE A.17. - Basic Out-of-the-Loop AI Workflow



Source: The World Bank.

The Human-in-the-Loop workflow incorporates an additional logical step in the overall workflow that presents all predicted data to a human for further intervention. This logical gate may enrich the result data, complete with additional information related to the outcome. There, a human being can review the information and select the proper classification. This is useful when there must be no doubt in the accuracy of predicted results. The Human-over-the-Loop workflow selectively gates results that demonstrate a high probability of accuracy for additional human intervention. This allows humans to concentrate on other mission-critical tasks when the certainty of predicted results is uncompromising, and it allows for intervention when the quality of predicted results falls below a threshold of desired probability.

Both variants of the advanced AI workflow may feed human-vetted predicted results back to the historical data stream for model retraining before continuing into the data stream that captures and distributes results to subsequent applications. Vetted data are particularly useful for improving the quality of decision-making over time, because human intervention lessens the gap of uncertainty and provides the model with increasing accuracy.

Feature engineering is also present in the advanced AI workflow. All problem domains require specific knowledge when deciding what data to collect. This valuable domain knowledge can also be used to extract value from collected data. Creating new data from existing data are called feature engineering. This phase occurs prior to model building. Once the AI loop is functioning adequately, practitioners often find the majority of their time going into this part of the optimization process. This is the more creative part of developing AI solutions since it requires imagination and knowledge to invent ways to improve the model by extracting hidden value from standard data. Common examples of feature extraction include converting dates and times to times of day/week/year, location-wrangling, in conjunction with census data, and object detection in land use imaging data that is useful in classification.

The mention of data streams deserves some attention when discussing real-world AI. Data streams are an important component in hybrid AI architectures. During development, data scientists may load data from comma-separated or tab-delimited data files for cleaning and processing. In practice, data flows from input to output in a constant stream. Thus, the term "data stream" is applicable. It is fair to wonder "what" exactly streams the data. Data streams are usually event-driven applications that "listen" for specific events within the system architecture. These are powered by open source technologies, particularly Kafka, that consume data from various sources through a standardized API, such as those offered by a relational database or ERP system. Data streams are especially useful because they offer data to one or more authorized applications that are capable of "listening" to the data stream producers using APIs over the network.

Distributed Ledger Technology

Any mention of DLT among the general public stirs the topic of cryptocurrency and alternative currency markets, particularly Bitcoin. However, while Bitcoin is a particularly popular example of DLT, when put to novel use, the subject of DLT spans a much broader variety of topics. At the core, DLT is a distributed network of computing nodes that manage identical ledgers containing blocks of data. Each ledger contains blocks linked together in a manner that prevents tampering by enforcing consensus requirements across the distributed network. At a minimum, each block contains an index that references the block order, a timestamp that references creation time, a cryptographic hash that is a signature of the data contents "salted" with the previous hash, a hash referencing the preceding block, and the rows of data, which may be individually encrypted for added security.

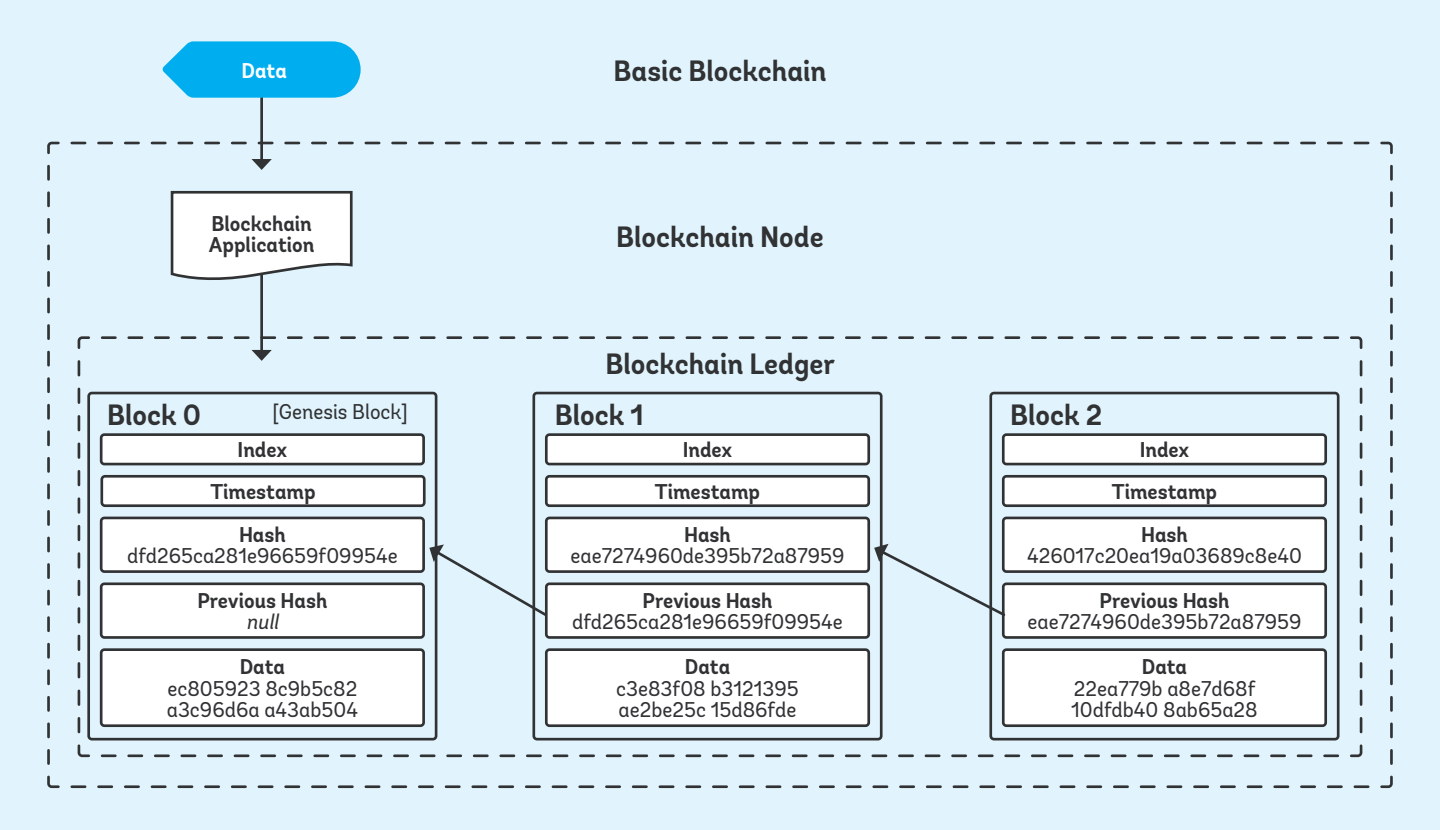
DLT Architecture

A block is defined simply as a collection of batched data. Block size is determined by the fault tolerance of the blockchain network due to distributed denial-of-service attacks and other factors related to network capacity. The typical block size is 1MB but that can be tuned to the needs of a particular application. The data stored in blocks is typically metadata on the order of kilobytes in scale. Storing large files in blocks is contraindicated to the functionality of the standard blockchain. Typically, large files are stored in a filesystem while the information describing their contents such as location, author, and perhaps a hashed checksum that serves as a signature for the integrity of the file, is written to the batched data buffer that eventually becomes part of the block. Data stored in a block can be encrypted and later deciphered upon retrieval. When batched data reaches the block size limit, the block is hashed using a cryptographic algorithm, and the blockchain algorithm places a request to add the block to the distributed ledgers throughout the network. Multiple requests may be placed from different nodes in the network; these are handled in sequential order, and the consensus mechanism aids in orderly propagation of data.

Because a block hash is unique to data contained within the block and the blockchain links blocks using hashes generated from the previous hash and current, any mutation to the data within a prior block will change the reference in subsequent blocks, thereby breaking the blockchain. Figure A.18 illustrates the design of a single computing node within the DLT network.

> > >

FIGURE A.18. - Basic Blockchain Node



Source: The World Bank.

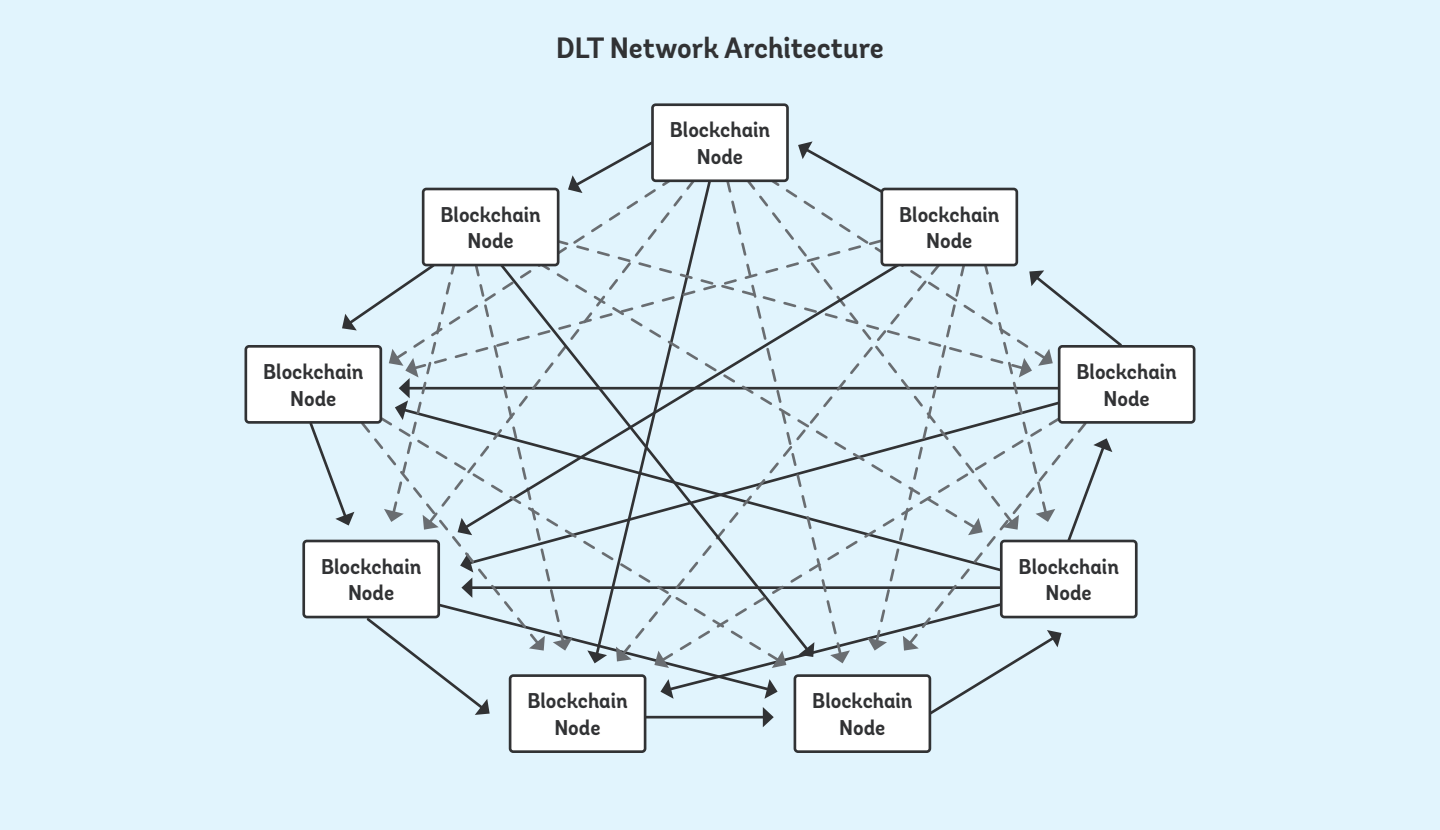
Blockchain security is compelling for archival data storage in government systems. Yet, a single node is insufficient for establishing a proper DLT network. The owner of one centralized node can simply alter any block arbitrarily and rewrite the hashes for all the subsequent blocks! Thus, the power lies in decentralization. DLT architectures distribute the entire ledger to nodes qualified to participate in the DLT network and require consensus before new blocks may append to the blockchain. DLT requires consensus to prevent the Byzantine Generals Problem, which arises when actors attempt conflicting actions such as overwriting or altering the blockchain with nefarious intent.

Consensus

There are several mechanisms to achieve consensus: proof-of-work, proof-of-stake, proof-of-bid, and the list goes on. The goal of each consensus algorithm is to validate the blockchain integrity before the block append operation is distributed to all the remaining nodes in the network. Energy consumption is the reason so many forms of consensus exist. A network consensus generally consumes a tremendous amount of computational power—thus, electricity—and utilizes a large amount of network resources. Therefore, it is imperative for the DLT network architecture to implement an efficient consensus mechanism. Figure A.19 illustrates the architecture of a DLT network. When one of the nodes in the network captures enough data to write a block to its local blockchain, it issues a consensus request to other nodes in the network according to the rules of the consensus mechanism. When the network reaches a consensus, the source node writes the block, and the block propagates throughout the remaining nodes in the network.

> > >

FIGURE A.18. - Basic Blockchain Node



Source: The World Bank.

Government information systems can benefit significantly from the use of DLT for maintaining the integrity of mission-critical data. Data for procurement, FMIS, and other systems generating transactions are the primary benefactors. When a transaction is generated, AI systems can send data to archival DLT nodes for archiving. Archived data stored in a DLT architecture helps maintain integrity throughout the network of participating nodes for reasons that should be obvious, given the context of preceding sections. Overall, a network of government agencies, or even departments within an agency, may become a stand-alone DLT network that is capable of maintaining, authenticating, and honoring long-term commitments to data integrity within the government. Should any participating node attempt to sabotage the integrity of the transactional blockchain archive, a mechanism can be established to alert overseers of the transgression and preventative action can be taken to investigate the problem and take appropriate action to prevent any fraud or corruption.

The most prolific government projects leveraging DLT for the purposes of distributed trust are those of central banks. Banks leverage DLT for Treasury Single Accounts that are host to foreign exchange transactions between central banks to speed the settlement of international exchange. Several experimental models are under consideration by the Monetary Authority of Singapore in ongoing research conducted through Project Ubin (https://www.mas.gov.sg/schemes-and-initiatives/Project-Ubin).

Additionally, governments can benefit significantly from implementing DLT along with AI processes in procurement and logistics. By tracing the procurement process with a distributed ledger, equipment, raw materials, and various critical resources can be transferred between parties with granular control.

>>>

Appendix B. AI and the Sectors

AI is gaining traction as an invaluable tool in urban planning, resource utilization, energy management, and climate change. Several practical applications are in development due in part to successful academic research funded by private enterprise. This trend will continue as humans occupy more densely populated urban areas that make use of natural resources in all manners. The scope of development and land use is enormous considering that most of human resource management touches on nearly every aspect of society in some form. Appendix B attempts to highlight many solutions that rely on AI for improvements in efficiency, scientific analysis, and prediction within the disciplines mentioned above. Figure B.1 illustrates the timeline of AI innovation in the environment and potential impact over the next 20 years.

> > >

FIGURE B.1. - Timeline of AI innovation in the Environment and Impact Over the Next 20 Years

Source: http://www3.weforum.org/docs/Harnessing_Artificial_Intelligence_for_the_Earth_report_2018.pdf.

The list of sectors covers energy, agriculture, materials science, transportation, climate management, and urban planning. The overall effort is toward a more effective feedback loop that mitigates risks brought on by overpopulation and resource scarcity in all of these sectors.

Agriculture

Agricultural innovators are currently using AI to model several interdependent factors in an effort to maximize food production yields. By consuming vast amounts of weather conditions, satellite and drone imaging, temperature, water use, soil conditions, crop rotation, and annual yields, AI systems are able to suggest optimal planting patterns that guide heavy equipment using geospatial precision. AI monitoring assists with managing water distribution during the growing season. As harvest approaches, AI leverages hundreds of thousands of data points on the ground from the Internet of Things (IoT) devices combined with satellite or drone imagery to determine optimal harvest quality and accuracy, which minimizes food spoilage in the post-harvest supply chain. In parts of Africa and Asia, AI helps maximize food production given the increasing dearth of annual rainfall, which forces farmers to become more precise in their forecasting and planning. The use of computer vision in combination with deep learning methods can detect potential fluctuations in pests, disease, water shortage, and harvestability. This is all a part of an emerging discipline called precision agriculture.

More specifically, a project called Ag-Analytics is collecting farmland data in the cloud and making it available to farmers for precision agriculture. Ag-Analytics uses sensors to collect soil, tillage, and yield-data for specific plots of farmland (https://analytics.ag/Home/HowItWorks). Microsoft Azure stores the data and shares the information with farmers through user-friendly APIs to lower costs, improve yields, and minimize the environmental cost of agriculture.

AI is also assisting with labor shortages in agriculture. As society becomes more urbanized, the supply of labor continues to move toward urban centers. Seasonal agricultural demand is faced with consistent shortages. Companies like

Root.AI are developing robotic harvest systems to bridge the gap between supply and demand of labor during harvest. Advanced methods in autonomous robotics, computer vision, botany, and biotechnology form the basis for the production of large scale operations capable of detecting ripeness and continuous harvesting at the peak of efficiency.

Chatbots also enable farmers to share information and resolve problems in the supply chain. The proliferation of chatbots in AI is made possible through the use of advanced NLP frameworks. Farmers can turn to chatbots for difficulties in production planning and resource management that are common to agriculture.

Agricultural monitoring by whole-of-government systems, using a data fabric, can leverage resource production and prevent state capture events from occurring in underrepresented regions. Many of the methods in agriculture are also relevant to mineral resources and energy production, so investment in these technologies is worth considering for the advancement of digital government systems.

Ecology, Climate, and Conservation

Deforestation and land degradation are major problems for ecosystems. Governments and NGOs are using AI to monitor the steady decline of forests worldwide. By using multi-agent AI systems (MAS), resource utilization scenarios can better understand the impact that agricultural expansion has on forest decline. MAS has the ability to manage complex systems with several stakeholders to allow the exploration of alternative forest and land management systems. Moreover, MAS serves as a tool for learning and understanding, rather than predictive analysis. Reinforcement learning (RL) methods using computer vision and transfer learning are most suitable for forest management and conservation.

Climate change stemming from deforestation also requires a comprehensive understanding of additional factors in the overall health of both local and global ecosystems. Several AI subdomains are necessary for the comprehensive analysis of such a monumental topic. Table B.1 illustrates the various subdomains relating to AI that are currently employed for climate impact mitigation.

> > >

TABLE B.1 - AI for climate impact mitigation

	Computer Vision	NLP	Time-series analysis	Unsurpevised learning	RL & Control	Casual inference	Uncertainty qualification	Transfer learning	Interpretable ML	Other
Electricity Systems	1	1.1	1.1 1.2	1	1.1		1.1 1.2	1.3	1.1	1.1
Transportation	2.1 2.2 2.4		2	2.1 2.4	2	2.1 2.4	2	2.1 2.4	2	
Building & Cities	3.2	3.3	3	3	3.1	3.1	3.3	3		
Industry	4.1 4.3		4.3	4.3	4	4.2 4.3		4.2 4.3	4.3	
Farms & Forests	5.1 5.3 5.4				5.2			5.4		
CO_2 Removal			6.3				6.3	6.3		6.2
Climate Prediction	7.1		7				7.3		7	
Societal Impacts	8.1 8.4	8.4	8.2 8.3		8.2	8.3	8.2	8.1	8.3	
Solar Geoengineering			9.3		9.4		9.3 9.4			9.2
Tools for Individuals	10.1	10.1	10.2	10.3	10.2	10.1			10.2	10.2
Tools for Society		11.1	11.2 11.1	11.3	11.2 11.1	11.1 11.3	11.1	11	11.1	11.1 11.3
Education		12.2			12.1					
Finance		13.2	13				13.2			

*Source: https://miro.medium.com/max/1400/0*7_IIv_JRbf85ClQj.*

The world's oceans are under increasing threat due to human overpopulation. A project called OceanMind is using satellites and AI to preserve biodiversity, protect the livelihoods of fishermen, and prevent slavery in the fishing industry. It collaborates with governments to prevent illegal, unreported, and unregulated fishing by analyzing vessel movements in real time. AI algorithms detect anomalous behavior that OceanMind shares with regulatory agencies to direct ocean patrols more efficiently.

In forest management and conservation, SilviaTerra is transforming how conservationists and landowners measure and monitor forests (https://www.silviaterra.com). The system tracks an inventory of forest resources for the protection and management of ecological, social, and economic health. SilviaTerra uses AI frameworks on Microsoft Azure to study the effects of climate change and improve habitats using high-resolution satellite imagery, U.S. Forest Service inventory, and field data to train AI models to measure forest values.

In species conservation to fight extinction, Wild Me is leveraging computer vision, citizen science, and deep learning algorithms to power Wildbook (http://www.wildbook.org/doku.php). Wildbook scans and identifies individual animals and species. Wildbook is notably an open source platform. It provides scalable and collaborative wildlife data storage and management, extensible easy-to-use software tools, API support, data exposure to external biodiversity resources, and animal biometrics that support easy data access. This robust design for data interchange using APIs makes it a stellar example of a system that will integrate well with a whole-government data fabric architecture. (Wildbook, Software to Combat Extinction) Another project in the same domain is Protection Assistant for Wildlife Security (PAWS). PAWS uses AI to aid conservationists in the fight against poaching by utilizing AI for learning, planning, and behavior modeling. PAWS collects information from previous poaching activities and then generates predictions about poaching locations and optimal patrol routes, resulting in more effective patrols and better use of resources in the fight against poaching endangered animal species (Fang 2013).

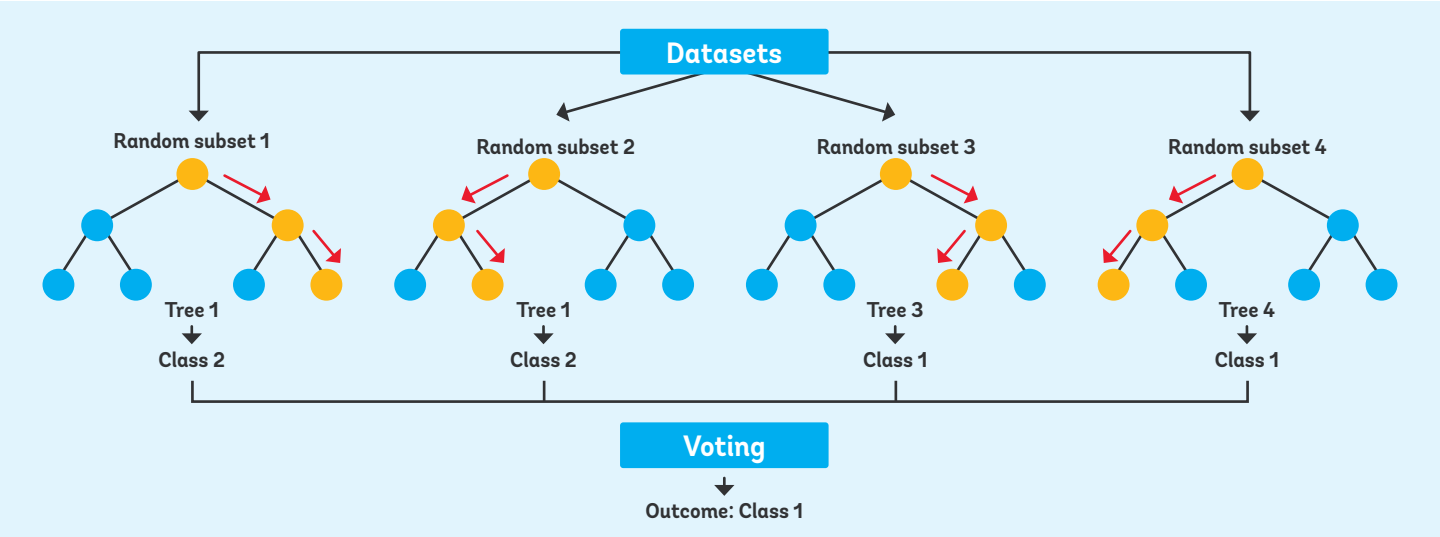
More technical information about the goal of tackling climate change with AI is available from a technical report published by a consortium of researchers from many prominent universities worldwide (Rolnick et. al. 2019).

Urban Planning

In one prominent example, researchers leverage advanced methods in predictive analysis using AI for urban planning. By using cellular automata in conjunction with evolutionary algorithms and AI, a mathematical model for predicting evolving spatial patterns examines the impact of policy and geography on the outcomes of various urban planning scenarios (Yang et. al. 2019). In plain English, this means they are using math to model the evolution of any urban environment over time. This framework optimizes Urban Development Demand by leveraging a model to synthesize changes in urban growth boundaries (UGB). The model uses historical observations of different time intervals and per-capita land requirements. Next, a patch-based cellular automata (CA) model simulates urban growth by estimating urban development probability using a random forest machine learning algorithm (Figure B.2).

> > >

FIGURE A.18. - Basic Blockchain Node

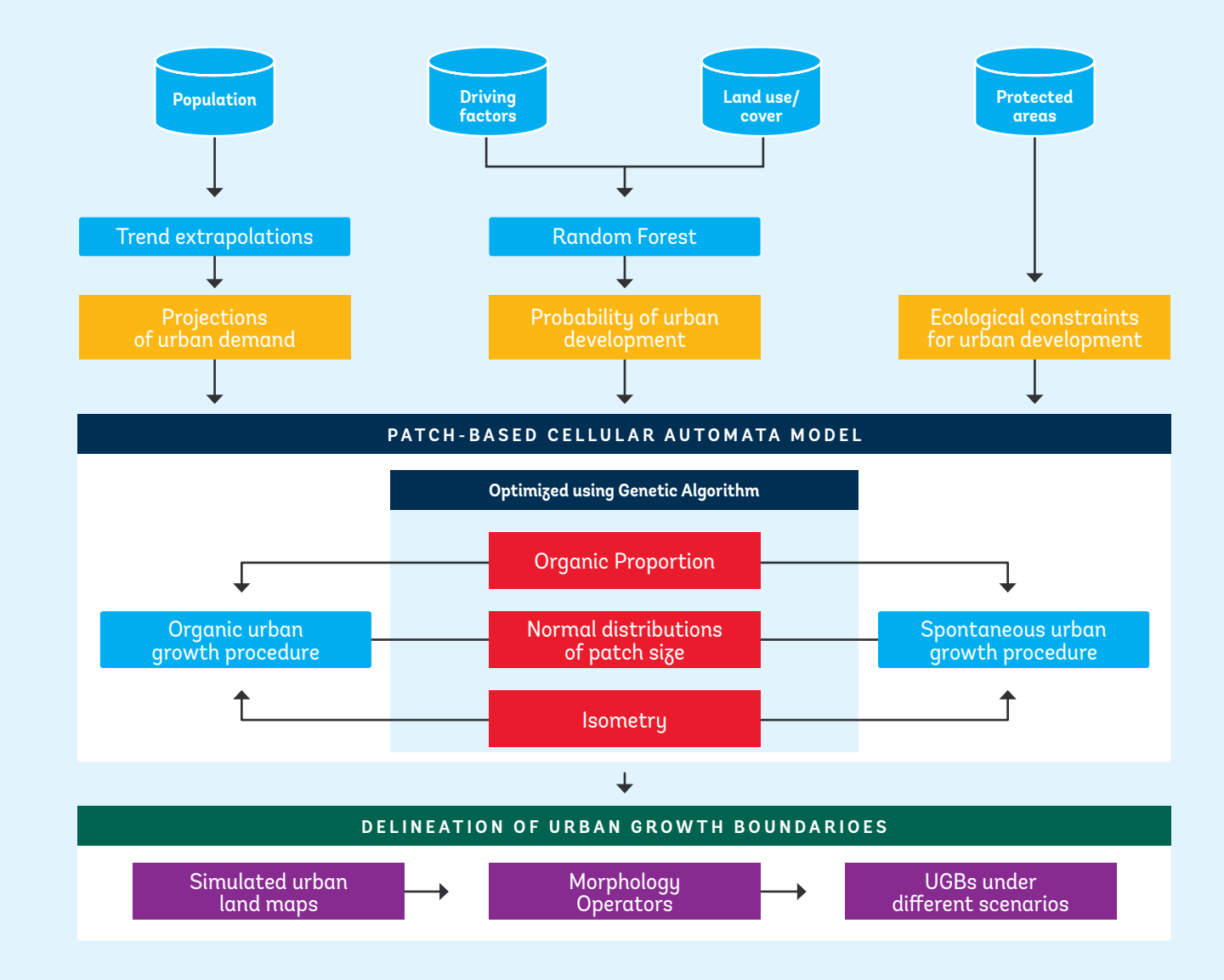


Source: Yang et. al. 2019.

The “patches” represent plots of land. Then, genetic algorithms optimize key model parameters, and finally the system aggregates land maps from multiple model runs to generate UGB alternatives. The random forest (RF) algorithm models a classification hierarchy using a strategy that creates a “forest” of individual decision trees. RF is hardly the only model in AI, but it is the most useful in this case. Each “tree” in the RF model makes independent decisions based on the feature variables and a random selection of observations derived from training data. Final outputs are the resulting averages of the decisions of the individual trees, which is considered a “voting strategy” that generates the resulting outcome. The RF method is insensitive to outliers, noise, and overfitting. Figure B.3 illustrates the workflow of modules within this predictive UGB framework.

> > >

FIGURE B.3. - Workflow of Modules within Predictive UGB Framework

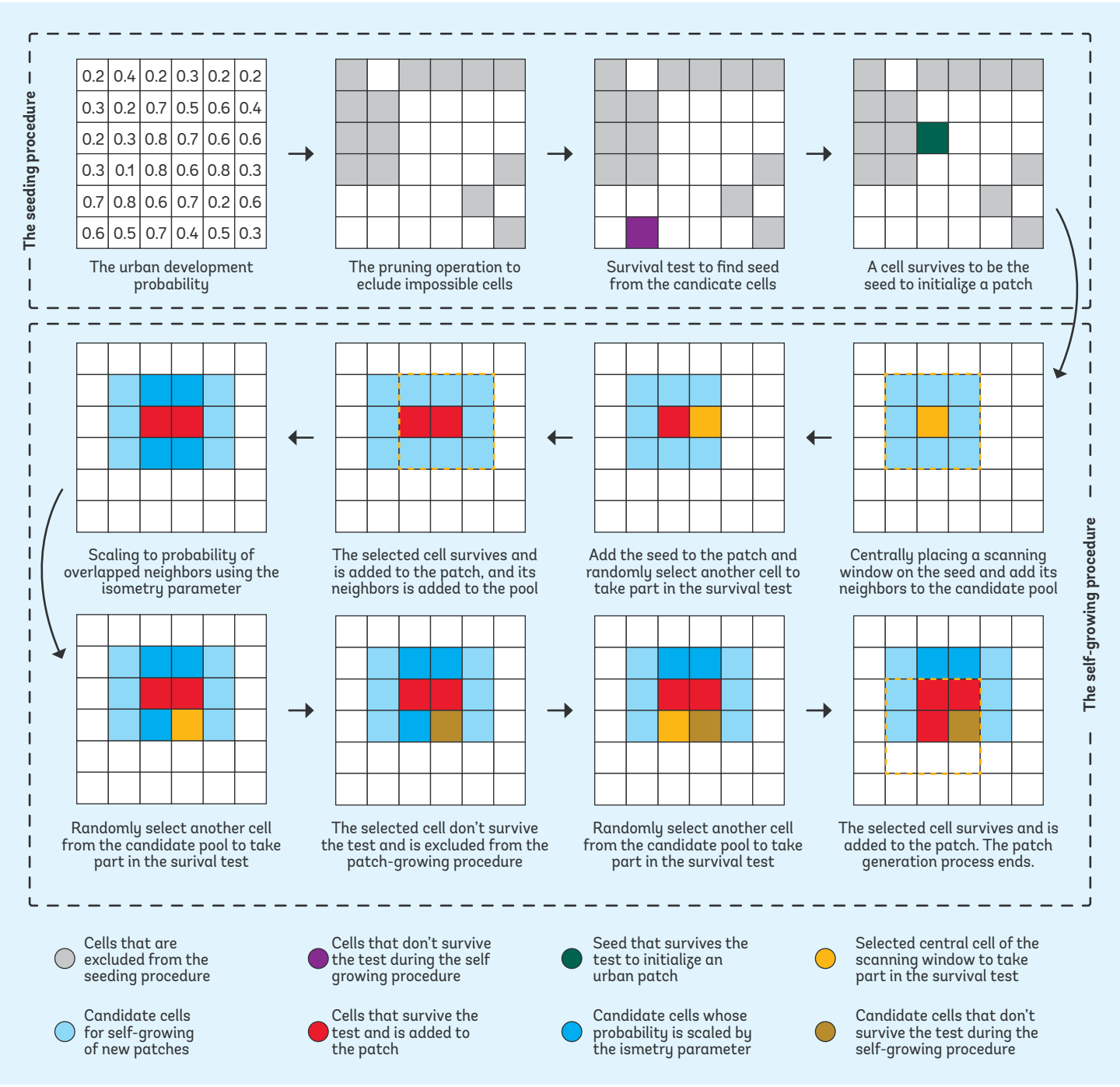


Source: Yang et. al. 2019.

Figure B.4 shows the CA patch generation function with a size of three cells. Think of the cells as patches of land with land use probabilities. Note again that the cells represent plots of land area.

> > >

FIGURE B.4. - Cellular Automata Patch Generation Function with a Size of Three Cells

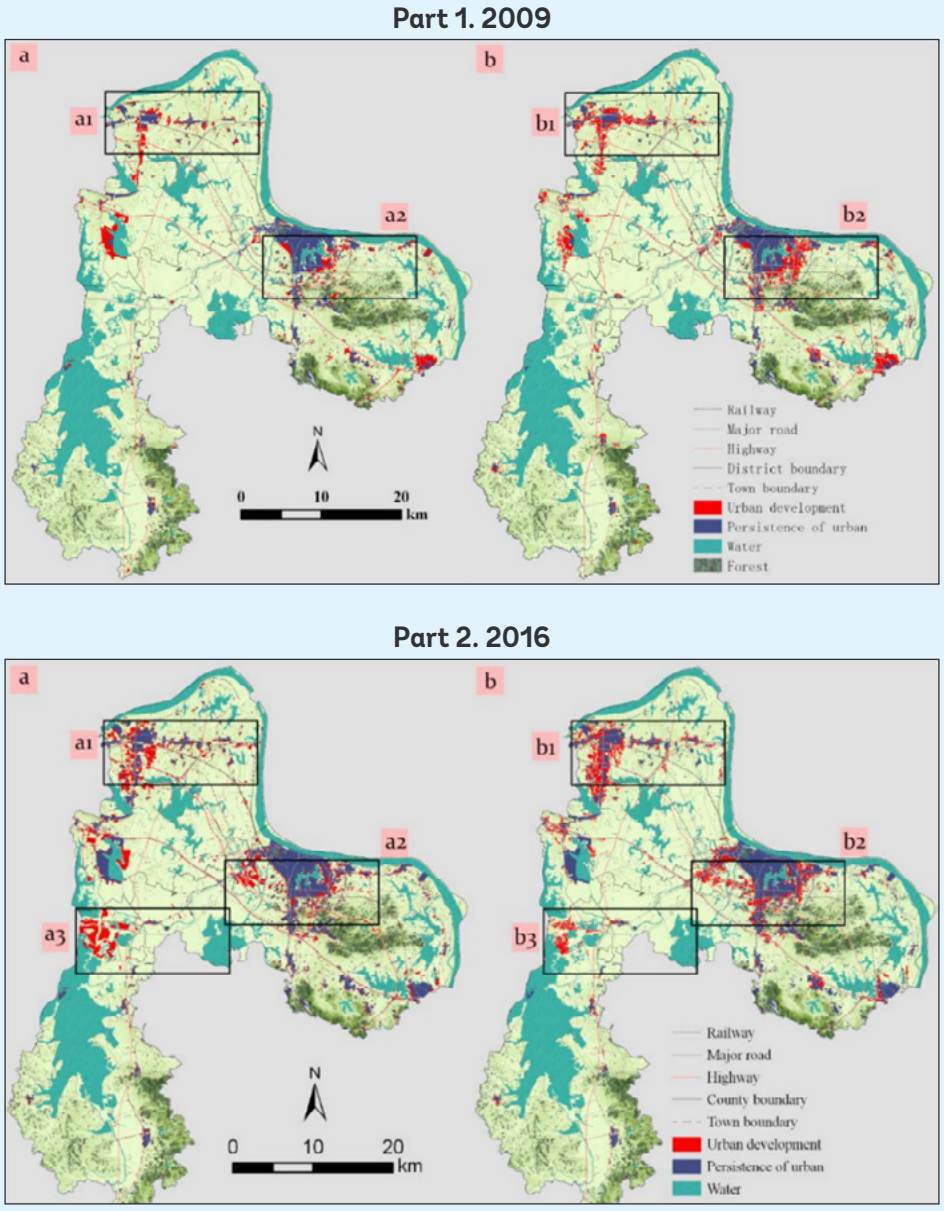


Source: Yang et. al. 2019.

Lastly, Figure B.5 illustrates the simulated and observed land map in 2009 and 2016. Map "a" is the observed land map, while map "b" is the simulated solution with the highest fitness score, meaning the best fitting results of the genetic algorithm. Note the accuracy of the predictions. In the real world, this model was put to use in an undisclosed rapidly growing city in China and revealed high reliability in the simulation of urban growth and the delineation of UGBs.

> > >

FIGURE B.5. - Simulated and Observed Land Maps in 2009 and 2016:



Note: Map "a" is the observed land map; map "b" is the simulated solution with the highest fitness score.
Source: Yang et. al. 2019.

The patch-based CA model, which represents urban growth as an organic and spontaneous process can simulate more realistic urban landscapes by coupling the spatial process with the pattern of urban development. The RF model can successfully show the relationship between driving policy factors and the urban development probability. Key model parameter calibration is achieved through genetic algorithms that capture the landscape characteristics of historical urban changes quite well and can therefore be used for future projections. The results also suggest that empirical (observed) knowledge from historical observations can assist the genetic algorithm with avoiding overfitting, to some extent. Although this model leverages simple population projection methods, the factors that drive future urban development can be further enhanced, and government planners can derive a deeper understanding and analysis of the resulting planning scenarios with more comprehensive data.

Energy and Smart Cities

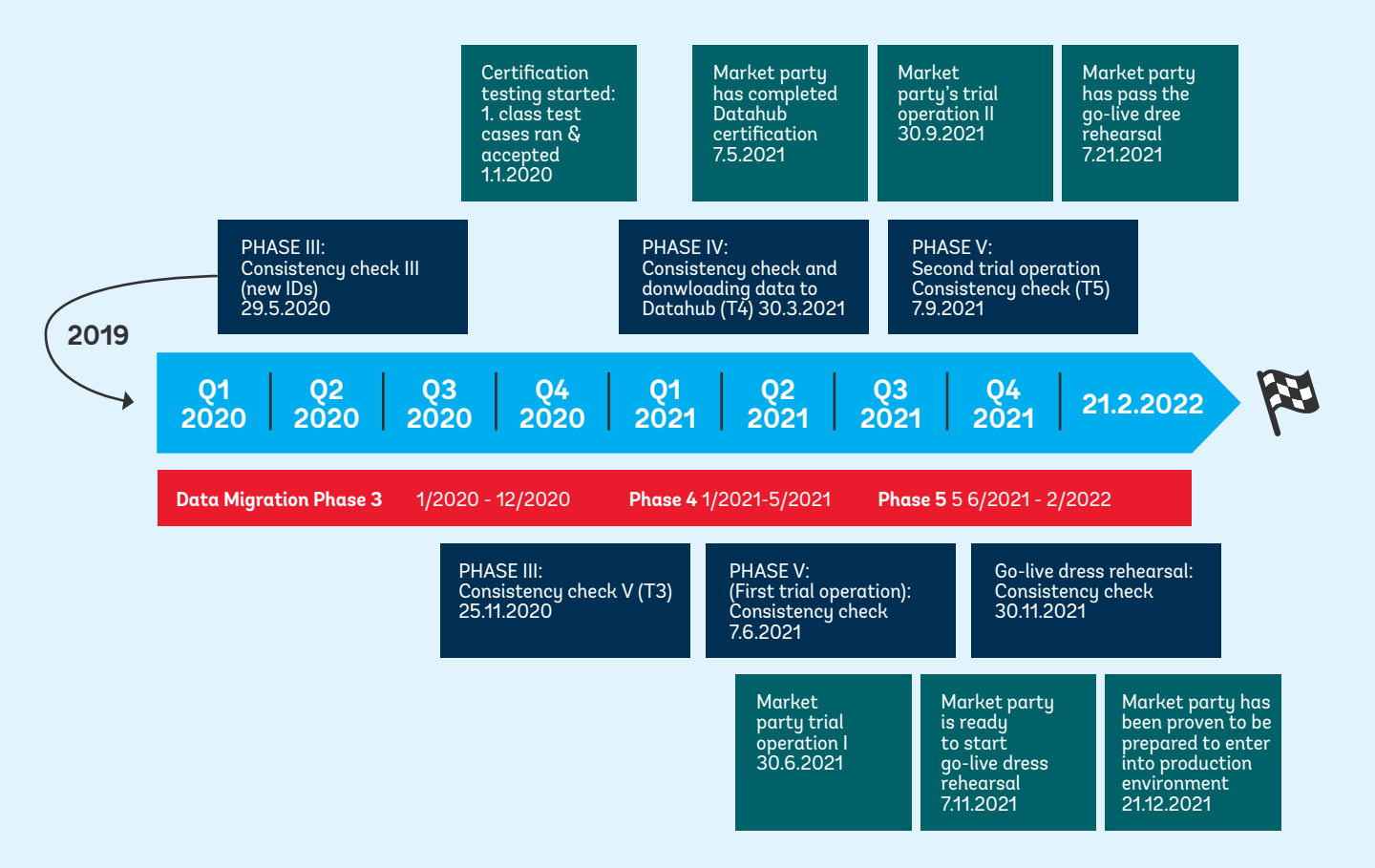
The introduction of IoT in cities around the world enables the use of AI in the management and planning of electricity. IoT smart meters transmit information over Wi-Fi using passive communication between a grid of meters, distributed in high and moderate density urban environments. The data allows energy companies to adjust electricity production to nearly real-time accuracy. Prior to the advent of smart grid technology, power companies needed to predict demand based on a combination of environmental predictions using weather and temperature forecasting and almanac predictions. This led to massive inefficiencies and wasteful production of electricity.

The invention of large scale data processing systems and introduction of data fabric infrastructure allowed power companies to transition to consuming massive pipelines of information about electricity use, thereby reducing the impact of electricity production on the environment and improving the overall efficiency of the electricity marketplace. The utilization of electricity in modern cities is now burgeoning as a result of widespread transition to connected electric vehicles, which serve as distributed reserves of electricity.

One notable example of AI in a data fabric stems from the inception of SmartGrid AI systems using a large scale data layer that transformed power grid utilization in Ontario, Canada. The project serves over 70 regional distribution companies handling reads from over four million meters and processing over 100 million transactions per day (KX Systems 2014). A similar project is called FinGrid, which is run by the primary transmission provider for Finland (KX Systems 2018). FinGrid will process data from 3.7 million locations to deliver 15-minute imbalance settlements between electricity suppliers and consumers, an EU regulatory requirement, by December 2020. This architecture is called DataHub. The data migration and go-live planning are important examples of how existing systems can transition to entirely new data architectures with minimal disruptions to existing mission-critical services. Figure B.6 illustrates the proposed transition plan for FinGrid (Fingrid Datahub 2019).

> > >

FIGURE B.6. - Proposed Transition Plan for FinGrid

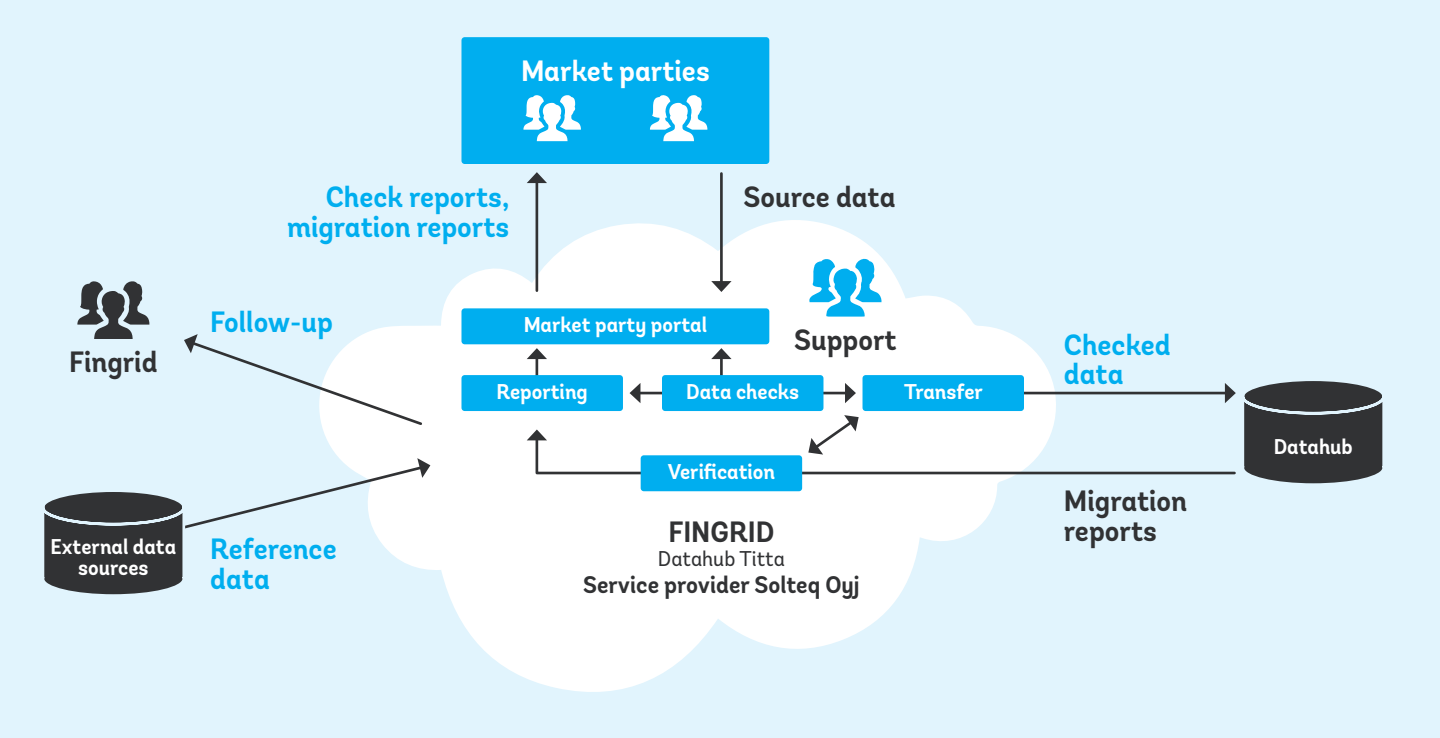


Source: Fingrid Datahub 2019.

Figure B.7 illustrates the workflow of the data migration project for FinGrid using kdb+, which is defined here as a column-based relational time series database (TSDB) with in-memory database (IMDB) abilities. It is commonly used in high-frequency data sets needing storage and retrieval of large data sets at high speed.

> > >

FIGURE B.7. - The Workflow of the Data Migration Project for FinGrid Using kdb+



Source: Fingrid Datahub (2019).

For more information about Datahub and Fingrid, visit the Fingrid website at https://www.ediel.fi/en/datahub/business-processes/business-process-other-datahub-instructions.

>>>

Glossary

Big Data	One or more databases containing extremely large data from various sources.
Data Dominion	The scope of a government's ownership and use of data, applications, and infrastructure defined by geographic, political, and national boundaries.
Data Fabric, Data Lake	An interconnected data storage infrastructure that provides a common set of interfaces and access control layers for "Big Data" operations spread across thousands of servers that may be geographically distributed.
Data Silo	An architecture that is isolated due to the absence of a common application programming interface (API) for inter-process communication.
Document	A self-contained JSON object specifying the attributes and values in a comprehensive unit of information that is iterable, transactable, and mutable.
Dummy Variable	A binary feature that indicates that an observation is (or is not) a member of a category.
Features	The input attributes that are used to predict the target, which may be numerical or categorical.
Feature Engineering	A form of machine learning optimization that leverages collected data to extract features.
Generative Adversarial Network	GANs are a class of machine learning techniques that consist of two simultaneously trained models competing as adversaries with one another: one (the Generator) trained to generate fake data, and the other (the Discriminator) trained to discern the fake data from real examples.
Ground Truth	The value of a known target variable or label for a training or test set.
Instance (Or Example)	A single object, observation, transaction, or record.
Model	A mathematical object describing the relationship between features and the target.
Online Machine Learning	A form of machine learning in which predictions are made, and the model is updated, for each example.
Preprocessing	The process of cleaning and correcting errors and inconsistencies in collected data. Also referred to as data munging or data wrangling.
Protocol Buffers, Protobuffers, Protobuf	A language-neutral, platform-neutral, extensible mechanism for serializing structured data.
Recall	Using a model to predict a target or label.
Supervised Machine Learning	Machine learning in which, given examples for which the output value is known, the training process infers a function that relates input values to the output.
Target (Or Label)	The numerical or categorical (label) attribute of interest. This is the variable to be predicted for each new instance.
Training Data	The set of instances with a known target to be used to fit a ML model.
Unsupervised Machine Learning	Machine learning techniques that do not rely on labeled examples, but rather attempt to find hidden structure in unlabeled data.

>>>

References

ACT-IAC (American Council for Technology and Industry Advisory Council). 2020. *Artificial Intelligence (AI) Playbook for the U.S. Federal Government*. Fairfax, VA: ACT-IAC. https://www.actiac.org/system/files/AI%20Playbook_1.pdf.

ANI (Asian News International). 2019. "Bahrain and UK First in the World to Pilot New Artificial Intelligence Procurement Guidelines Across Government." *Business Standard,* July 4, 2019. https://www.business-standard.com/article/news-ani/bahrain-and-uk-first-in-the-world-to-pilot-new-artificial-intelligence-procurement-guidelines-across-government-119070401389_1.html.

Bansal, Aayush. 2018. "Donald Trump to Barack Obama." August 11, 2018. YouTube video, 0:06. https://www.youtube.com/watch?v=F51RCdDIuUw.

Berryhill, Jamie, Kévin Kok Heang, Rob Clogher, and Keegan McBride. 2019. *Hello, World: Artificial Intelligence and its Use in the Public Sector.* Paris: OECD Publishing. https://oecd-opsi.org/wp-content/uploads/2019/11/AI-report-Online.pdf.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei. 2018. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Oxford, UK: Future of Humanity Institute. https://maliciousaireport.com/.

Buchanan, Ben. 2020. 2020. "A National Security Research Agenda for Cybersecurity and Artificial Intelligence." CSET Issue Brief, Center for Security and Emerging Technology, Washington, DC. https://cset.georgetown.edu/wp-content/uploads/CSET-A-National-Security-Research-Agenda-for-Cybersecurity-and-Artificial-Intelligence.pdf.

Bughin, Jacques, Jeongmin Seong, James Manyika, Michael Chui, and Raoul Joshi. 2018. "Notes from the AI Frontier: Modeling the Impact of AI on the World Economy." report, McKinsey & Company, Washington, DC. https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy#.

Chen, Stephen. 2019. "Is Fraud-Busting AI Systems Being Turned Off for Being Too Efficient?" *South China Morning Post,* February 4, 2019. https://www.scmp.com/news/china/science/article/2184857/chinas-corruption-busting-ai-system-zero-trust-being-turned-being.

Chilamkurthy, Kowshik. 2020. “Reinforcement Learning for Covid-19: Simulation and Optimal Policy.” *Towards Data Science* (blog), March 31, 2020. https://towardsdatascience.com/reinforcement-learning-for-covid-19-simulation-and-optimal-policy-b90719820a7f.

Coursera. 2019.

Craddock, M. 2019. “UN Global Platform.” Retrieved June 27, 2020, from https://unstats.un.org/unsd/bigdata/conferences/2019/presentations/seminar/day1/5th%20Big%20Data%20External%20Workshop%20Slides%20-%20UN%20Global%20Platform.pdf.

Dandekar, Raj, and George Barbastathis. 2020. “Quantifying the Effect of Quarantine Control in Covid-19 Infectious Spread Using Machine Learning.” *medRxiv;* 2020. DOI: 10.1101/2020.04.03.20052084. https://www.medrxiv.org/content/10.1101/2020.04.03.20052084v1.

Data Center Map. 2020.

Dignan, Larry. 2017. “IBM’s Rometty Lays Out AI Considerations, Ethical Principles.” *Between the Lines* (blog), June 17, 2017. https://www.zdnet.com/article/ibms-rometty-lays-out-ai-considerations-ethical-principles/.

Dutton, Tim. 2018. *Building an AI World Report On National and Regional AI Strategies*. Toronto, Canada: CIFAR. https://www.cifar.ca/docs/default-source/ai-society/buildinganaiworld_eng.pdf.

EC (European Commission). 2019. *Ethics Guidelines for Trustworthy AI*. An independent Report by the High-Level Expert Group on Artificial Intelligence. Brussels: European Commission.

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. “Robust Physical-World Attacks on Deep Learning Models.” *arXiv:*1707.08945v5. https://arxiv.org/abs/1707.08945.

Fang, Fei. 2013. “Protection Assistant for Wildlife Security.” *Societal Computing* (Applied Systems and Infrastructure), November 13, 2013. https://sc.cs.cmu.edu/research-detail/102-protection-assistant-for-wildlife-security.

Federico, C., and T. Thompson. 2019. “Do IRS Computers Dream About Tax Cheats? Artificial Intelligence and Big Data in Tax Enforcement and Compliance.” *Journal of Tax Practice & Procedure* February–March 2019: 43–47. https://www.crowell.com/files/2019-Feb-March-Do-IRS-Computers-Dream-About-Tax-Cheats-Federico.pdf.

Feldstein, Steven. 2019. “The Global Expansion of AI Surveillance.” Working Paper, Carnegie Endowment for International Peace, Washington, DC. https://carnegieendowment.org/files/WP-Feldstein-AISurveillance_final1.pdf.

Fingrid Datahub. 2019. *Go-Live Plan for Centralized Information Exchange Services (Datahub) for Electricity Market*. Helsinki, Finland: Fingrid Datahub Oy. (May 28, 2019). https://www.ediel.fi/sites/default/files/Go-Live%20plan%20for%20centralised%20information%20exchange%20services%20%28Datahub%29%20for%20electricity%20market.pdf.

Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI.* Cambridge, MA: Berkman Klein Center for Internet and Society. https://dash.harvard.edu/handle/1/42160420.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph *Luetge*, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. 2018. *AI4People's Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*. Brussels: Atomium–European Institute for Science, Media and Democracy Atomium. https://www.eismd.eu/wp-content/uploads/2019/11/AI4People%E2%80%99s-Ethical-Framework-for-a-Good-AI-Society_compressed.pdf.

Gartner. 2019. "Gartner Identifies Top 10 Data and Analytics Technology Trends for 2019." Press Release, February 18, 2019. https://www.gartner.com/en/newsroom/press-releases/2019-02-18-gartner-identifies-top-10-data-and-analytics-technolo.

GDS (Government Digital Service) and OAI (Office for Artificial Intelligence). 2019. "A Guide to Using Artificial Intelligence in the Public Sector." GOV.UK, June 10. https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector#contents.

Government of Canada. 2019a. "Directive on Automated Decision-Making." Government of Canada, modified February 5, 2019. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

IEEE (Institute of Electrical and Electronics Engineers). 2019. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems,* First Edition. Piscataway, NJ: IEEE. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf.

IOTA (Intra-European Organisation of Tax Administrators). 2018. *Impact of Digitalisation on the Transformation of Tax Administrations*. Budapest, Hungary: IOTA. https://www.iota-tax.org/sites/default/files/publications/public_files/impact-of-digitalisation-online-final.pdf.

ITU (International Telecommunication Union). 2019. *Measuring Digital Development: Facts and Figures 2019*. Geneva: ITU. https://www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2019.pdf.

Kernighan, Brian and Rob Pike. 1984. *The Unix Programming Environment*. New Jersey: Prentice Hall.

KX Systems. 2014. "KX Systems' kdb+ Chosen by Ontario Electric Grid Operator." Press Release, June 22, 2014. https://kx.com/news/kdb-technology-chosen-for-retrieval-and-querying-of-smart-meter-data-processed-by-ontario-smart-metering-system/.

KX Systems. 2018. "European Energy Market Contract Win with FinGrid." Press Release,

July 16, 2018. https://kx.com/news/european-energy-market-contract-win/.

Lane, Hobson, Hannes Hapke, and Cole Howard. 2019. *Natural Language Processing in Action*. Shelter Island, NY: Manning Publications.

McKinsey Global Institute. 2017. *Harnessing Automation for a Future that Works*. Washington, DC: McKinsey & Company.

Mazzucato, Mariana. 2015. "Re-Igniting Public and Private Investments in Innovation." Report presented at the U.S. Senate Forum of the Middle Class Prosperity Project "Building the Economy of the Future: Why Federal Investments in Science and Innovation Matter." Washington, DC, July 27. https://marianamazzucato.com/wp-content/uploads/2015/07/Mazzucato-Statement-Middle-Class-Prosperity-Project-.pdf.

Mozur, Paul, and Lin Qiqing. 2019. "Hong Kong Takes Symbolic Stand Against China's High-Tech Controls." *New York Times,* October 3, 2019. https://www.nytimes.com/2019/10/03/technology/hong-kong-china-tech-surveillance.html.

Nakasone, Keith. "Game Changers: Artificial Intelligence Part II; Artificial Intelligence and the Federal *Government*." Statement of Keith Nakasone, Deputy Assistant Commissioner, Acquisition Operations, Office of Information Technology Category (ITC), U.S. General Services Administration, before the Subcommittee on Information Technology of the Committee on Oversight and Government Reform, Washington, DC, March 7, 2018. https://republicans-oversight.house.gov/wp-content/uploads/2018/03/Nakasone-GSA-Statement-AI-II-3-7.pdf.

2154 Rayburn House Office Building

Ntoutsi, Eirini., Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernandez, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. 2020. "Bias in Data-Driven Artificial Intelligence Systems—An introductory Survey." *WIREs Data Mining and Knowledge Discovery* 10 (3).

O'Brien, Tim, Steve Sweetman, Natasha Crampton, and Venky Veeraraghavan. 2020. "How Global Tech Companies Can Champion Ethical AI." World Economic Forum Annual Meeting, Davos-Klosters, Switzerland, January 21-24, 2020. https://www.weforum.org/agenda/2020/01/tech-companies-ethics-responsible-ai-microsoft/.

OECD (Organisation for Economic Co-operation and Development). 2016. *Preventing Corruption in Public Procurement*. Paris: OECD Publishing. http://www.oecd.org/gov/ethics/Corruption-Public-Procurement-Brochure.pdf.

OECD (Organisation for Economic Co-operation and Development). 2019.

"Recommendation of the Council on Artificial Intelligence." OECD Legal Instruments, May 21, 2019. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

PDPC (Personal Data Protection Commission). 2020. *Model Artificial Intelligence Governance Framework* (Second Edition). 2020. Mapletree Business City, Singapore: Infocomm Media Development Authority and PDPC. https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf.

Perrault, Raymond, Yoav Shoham, Erik Brynjolfsson, Jack Clark, John Etchemendy, Barbara Grosz, Terah Lyons, James Manyika, Saurabh Mishra, and Juan Carlos Niebles. 2019. *The AI Index 2019 Annual Report*. Stanford, CA: AI Index Steering Committee, Human-Centered AI Institute, Stanford University.

Public-Private Analytic Exchange Program. 2018. *AI: Using Standards to Mitigate Risks*. Washington, DC: U.S. Department of Homeland Security. https://www.dhs.gov/sites/default/files/publications/2018_AEP_Artificial_Intelligence.pdf.

Rolnick, David, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording,Carla Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. 2019. "Tackling Climate Change with Machine Learning." *arXiv:* arXiv:1906.05433v2. https://arxiv.org/pdf/1906.05433v2.pdf.

Rossi, Francesca. 2019. "Building Trust In Artificial Intelligence." *Journal of International Affairs* 72 (1). https://jia.sipa.columbia.edu/building-trust-artificial-intelligence.

Stiglitz. 2018. https://royalsociety.org/science-events-and-lectures/2018/09/you-and-ai/
The Open Group. 2018. *The Open Group Base Specifications Issue 7.* San Francisco: The Open Group. https://pubs.opengroup.org/onlinepubs/9699919799/.

UN (United Nations). 2004. *United Nations Convention Against Corruption*. New York: United Nations. https://www.unodc.org/documents/treaties/UNCAC/Publications/Convention/08-50026_E.pdf.

UNESCO (United Nations Educational, Scientific, and Cultural Organization). 2020. "UNESCO Appoints International Expert Group to Draft Global Recommendation on the Ethics of AI." Press Release, March 11, 2020. https://en.unesco.org/news/unesco-appoints-international-expert-group-draft-global-recommendation-ethics-ai.

U.S. Department of the Treasury. 2017. *2017 Annual Privacy, Data Mining, and Section 803 Reports*. Washington, DC: Department of the Treasury. https://home.treasury.gov/system/files/236/annual-privacy-data-mining-Report-and-section-803-Report-final-2.pdf.

van Eyk, E., L. Toader, S. Talluri, L. Versluis, A. Uta, and A. Iosup. 2018. "Serverless Is More: From PaaS to Present Cloud Computing." *IEEE Internet Computing* 22 (5): 8–17.
Venkateswaran, T.V. 2020. "AI Isn't Unbiased because Humans are Biased." *The Eighth Column* (blog), February 19, 2020. https://thefederal.com/the-eighth-column/artificial-intelligence-algorithms-unbiased-humans-biased/.

WEF (World Economic Forum). 2018. *Harnessing Artificial Intelligence for the Earth* Cologny, Switzerland: World Economic Forum. http://www3.weforum.org/docs/Harnessing_Artificial_Intelligence_for_the_Earth_Report_2018.pdf.

WEF (World Economic Forum). 2020. “AI Procurement in a Box: AI Government Procurement Guidelines.” Toolkit June 2020, World Economic Forum, Cologny, Switzerland. http://www3.weforum.org/docs/WEF_AI_Procurement_in_a_Box_AI_Government_Procurement_Guidelines_2020.pdf.

West, Darrell. 2018. “Will Robots and AI Take Your Job? The Economic and Political Consequences of Automation.” *TechTank* (blog), April 18, 2018. https://www.brookings.edu/blog/techtank/2018/04/18/will-robots-and-ai-take-your-job-the-economic-and-political-consequences-of-automation/.

World Bank. 2016. *World Development Report 2016: Digital Dividends.* Washington, DC: World Bank. https://www.worldbank.org/en/publication/wdr2016.

Yang, J., J. Gong, W. Tang, Y. Shen, C. Liu, and J. Gao. 2019. “Delineation of Urban Growth Boundaries Using a Patch-Based Cellular Automata Model under Multiple Spatial and Socio-Economic Scenarios.” *Sustainability* 11 (21): 6159. https://www.mdpi.com/2071-1050/11/21/6159.

Zheng, Stephan, Alex Trott, Sunil Srinivasa, Nikhil Naik, Melvin Gruesbeck, David Parkes, and Richard Socher. 2020. “The AI Economist: Improving Equality and Productivity with AI-Driven Tax Policies.” *Salesforce Research* (blog), April 28, 2020. https://blog.einstein.ai/the-ai-economist/.

Supported by the GovTech Global Partnership: www.worldbank.org/govtech

GovTech
Putting people first

Federal Ministry
Republic of Austria
Finance

Ministry of Economy
and Finance
Republic of Korea

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs,
Education and Research EAER
State Secretariat for Economic Affairs SECO

WORLD BANK GROUP