United Nations | Department of Economic and Social Affairs

Lessons Learned from Sandboxing, Piloting and Policy Experimentation with AI and Other Digital Initiatives: Part 1, Summary Report

Prepared by Steven M. Miller

30 June 2025

United Nations | Department of Economic and Social Affairs

Lessons Learned From Sandboxing, Piloting and Policy Experimentation with AI and Other Digital Initiatives: Part 1, Summary Report

Prepared by Steven M. Miller

30 June 2025

About the Department of Economic and Social Affairs

The Department of Economic and Social Affairs of the United Nations Secretariat is a vital interface between global policies in the economic, social, and environmental spheres and national action. The Department works in three main interlinked areas: (i) it compiles, generates, and analyses a wide range of economic, social, and environmental data and information on which Member States of the United Nations draw to review common problems and to take stock of policy options; (ii) it facilitates the negotiations of Member States in many intergovernmental bodies on a joint course of action to address ongoing or emerging global challenges; and (iii) it advises interested Governments on the ways and means of translating policy frameworks developed in United Nations conferences and summits into programs at the country level and, through technical assistance, helps build national capacities.

About the Economic and Social Commission for Asia and the Pacific (ESCAP)

The Economic and Social Commission for Asia and the Pacific (ESCAP) is the most inclusive intergovernmental platform in the Asia-Pacific region. The Commission promotes cooperation among its 53 Member States and 9 associate members in pursuit of solutions to sustainable development challenges. ESCAP is one of the five regional commissions of the United Nations. The ESCAP secretariat supports inclusive, resilient and sustainable development in the region by generating action-oriented knowledge, and by providing technical assistance and capacity-building services in support of national development objectives, regional agreements and the implementation of the 2030 Agenda for Sustainable Development. ESCAP also provides support to partners at the national level. ESCAP's national offer is rooted in and linked with the implementation of global and regional intergovernmental frameworks, agreements, and other instruments.

About the UN Development Account

The Development Account is a capacity development programme of the United Nations Secretariat aiming at enhancing capacities of developing countries in the priority areas of the United Nations Development Agenda. The Development Account is funded from the Secretariat's regular budget and implemented by 10 entities of the UN Secretariat including DESA and ESCAP. Under the United Nations Development Account, the DA2124B project has been conceived with the objective to enhance the institutional capacity of selected countries in special situations to understand and develop policy experimentation and regulatory sandboxes to accelerate the progress of the 2030 Agenda for Sustainable Development. The DA 2124B project aims to support recipient countries to have the demonstrated ability to identify and establish means for policy experimentation and regulatory sandboxes for targeted digital initiatives of strategic national importance. Through DA 2124B, this project has been jointly implemented by the Division for Public Institutions and Digital Government (DPIDG) of United Nations Department of Economic and Social Affairs (DESA), and the Information and Communications Technology and Disaster Risk Reduction Division (IDD) of United Nations Economic and Social Commission for Asia and the Pacific (ESCAP).

Author

Steven M Miller, Consultant, DESA
Professor Emeritus of Information Systems, Singapore Management University

Disclaimer

Contents

Acknowledgements

The success of this project effort and the resulting two-part report (Part 1: Summary Report, and Part 2: Ten In-Depth Interviews) is attributed to the invaluable support received from numerous individuals and organizations. We would like to express our profound appreciation to all of those who agreed to be interviewed for this project. This includes those interviewed who were previously involved with the sandboxing and policy experimentation projects in Bangladesh, Kazakhstan and Maldives that were sponsored by the UN Development Account Project, Frontier Technology Policy Experimentation and Regulatory Sandboxes in Asia and the Pacific (DA 2124B). It also includes those interviewed from the EU, New Zealand, Rwanda, Singapore, United States and Uzbekistan who have been actively involved in public sector AI application projects and related policy experimentation efforts in their own country and in some cases, in other countries as well.

We also express our appreciation to our UN project sponsors from the Division for Public Institutions and Digital Government (DPIDG) of the United Nations Department of Economic and Social Affairs (DESA), and the Information and Communications Technology and Disaster Risk Reduction Division (IDD) of United Nations Economic and Social Commission for Asia and the Pacific (ESCAP). Staff members and interns from these organizations who played important roles in either creating this project, providing access to in-country experts, reviewing drafts of the report or helping with preparation for publication included Vincenzo Acquaro, Wai Min Kwok, Deniz Susar, Arpine Korekyan, Junho Lee, Aiai Li and Youngae Namkung.

Executive Summary

This report, "Lessons Learned from Sandboxing, Piloting and Policy Experimentation with AI and Other Digital Initiatives," captures insights and experiences from project experts involved in recent digital innovation initiatives with the governments of Bangladesh, Maldives, and Kazakhstan, and from project experts actively involved with the use of AI for delivering government digital services in the EU, New Zealand, Rwanda, Singapore, United States, and Uzbekistan. The ten in-depth interview write-ups produced from these nine different country settings provide a small but highly informative sample of rich descriptions of some of the important realities, approaches, nuances, issues and challenges related to testing and piloting public sector AI-enabled digital services and other digital innovation efforts. These interview write-ups and the related summary conclusions will help government officials better understand some of the important micro aspects and broader policy aspects of planning, piloting, and deploying AI and digital technology projects.

This report contains two parts. Part 1 is the summary report. It provides the motivation and background for the project effort, a listing of who was interviewed and the topics summarised in each interview write-up, a condensed summary of each full-length interview write-up, a summary of the project methodology, and conclusions and recommendations based on my assessment and interpretation of the interview content.

Key points from the concluding comments in the Part 1 Summary Report include:

- While R&D and new technology aspects of AI and other digital technologies move at a fast pace, the process of carefully testing, piloting, validating and evaluating the performance and broader impacts of these systems in the context of real-world public sector use cases and conditions necessarily moves at a much slower pace, creating an inherent and ever-present tension that will not disappear.
- Public sector officials at all levels need to grasp that the effort required to do careful and reliable validation and policy experimentation through a combination of sandboxing and field piloting requires persistent and patient effort over extended time periods, with longer (even multi-year) timescales required when the AI applications are being used to support more consequential and higher impact decisions.
- The initial steps of technically testing AI models using only historical data sets and other available information sources can sometimes proceed much more quickly because this type of testing does not involve any type of user testing or trials in the real domain context. Yet, the follow-on phases of doing higher fidelity, more realistic sandboxing and field piloting cannot move as fast or be completed as quickly due to all the complexities involved with real public sector use cases and live users, complex real-world domain requirements and constraints, and the time required to do validation, experimentation and evaluation.
- Public sector decision makers overseeing these AI and digital innovation projects need to pursue the strategy of disciplined selection and filtering to limit the number and scope of initiatives that move beyond the technical model testing phase and into the subsequent phases of sandboxing and field piloting. Simultaneously, over time, public sector decision makers also need to increase the cumulative quantity of efforts moving into the sandboxing

and field piloting phases by maintaining the steadiness of project flow through the validation pipeline while also implementing the supporting efforts to increase the capacity and speed of this pipeline.

- The terms related to “test”, “sandbox”, “pilot”, “field pilot”, and “deploy” are used throughout the ten interview write-ups and the meaning of these terms sometimes differs across the contexts of the various interviews. A framework is given for understanding these terms in the context of a systematic progression of four phases used to test, validate, evaluate and eventually operationally deploy a new AI-based application or any type of complex digital solution.
- Public sector officials involved in reviewing and overseeing AI efforts and other digital technology innovation efforts need to understand the nature and importance of all four of these phases, the different meaning of testing and validation within each of these phases, and the special importance of the phases of sandboxing and field piloting and the way they enable policy experimentation.
- A public sector organisation must be able to realistically assess its internal ability at any given point in time to accomplish these four phases of testing and validation through a combination of using internal staff and through procuring external vendor and consultant services. The public sector organisation may even need external help to do this type of assessment of their internal ability to the necessary testing, validation and experimentation and how to get the work done given internal capability gaps.
- A piloting effort is so much more than just getting the technical aspects of an AI model to work as it is necessary go beyond that and observe, test, validate and evaluate the larger socio-technical system involved.
- Partnerships between public sector organisation and universities have been an effective mechanism to support public sector efforts to develop AI-based decision support models, to test and evaluate AI solutions and their impacts, and to develop relevant manpower.

Part 1 concludes with three types of recommendations:

- Recommendations for UN Development Account projects involving sandboxing and piloting with AI and other digital technology applications.
- Recommendations for future editions of the UN DESA E-Government Survey related to sandboxing and piloting with AI and other digital technology applications.
- One recommendation for public sector institutions implementing government digital services that use AI, which is to read some or all the full-length interview summaries in Part 2, as these are the “gems” of this project. No simply distilled, condensed summary of the full-length write ups, or briefly stated recommendation derived from them, can substitute for the richer experience of reading some of these full-length interview write-ups.

Part 2 of this report contains the full write-ups for each of the ten in-depth interviews.

1. Introduction

1.1 Project origins and motivations

There were two motivations behind embarking on the project which led to this report on "Lessons Learned From Sandboxing, Piloting and Policy Experimentation With AI and Other Digital Initiatives." The first motivation links back to UN DESA Policy Brief #123, "Sandboxing and experimenting digital technologies for sustainable development" published in December 2021.[1] That prior policy brief explained the concept and purpose of a regulatory sandbox, the four stages of sandboxing, why digital capacity development experiments done through sandboxing are relevant to making progress towards the sustainable development goals, and limitations of this approach. It also mentioned three national digital experimentation and sandboxing efforts ongoing at that time that UN DESA and UN ESCAP were jointly sponsoring with the governments of Bangladesh, Kazakhstan and Maldives through the UN Development Account project, "Frontier Technology Policy Experimentation and Regulatory Sandboxes in Asia and the Pacific" (DA Project 2124B).

The first motivation for this project was to provide a means of circling back to those three country-specific efforts after the initial phase of work co-sponsored by UN DESA and ESCAP had been completed to capture lessons learned about sandboxing and experimentation with digital initiatives through a small sample of in-depth interviews.

The second motivation is related to the UN E-Government Survey published bi-annually by UN DESA. Though the 2022 version of the survey report [2] noted that more governments were deploying artificial intelligence (AI) technology along with other forms of advanced digital technology, the term "artificial intelligence" was only mentioned in the main body of the text (excluding references) six times, reflecting growing awareness and usage, but a limited extent of diffusion and deployment.

In the 2024 version of the survey[3], the situation with respect to the presence of AI in digital government efforts across the world had changed dramatically. The term "artificial intelligence" was mentioned 36 times in the text of the main report (including the mentions in a new addendum on Artificial Intelligence and Digital Government that was added to the survey report for the first time), and was further mentioned in the survey's technical appendix (a separate document from the main report) which included its own addendum with a list of UN artificial intelligence initiatives.

While the substantially increased presence of artificial intelligence in the 2024 version of the survey report is partially due to the addition of several new questions specifically asking about artificial intelligence related plans, usage and governance, the 2024 survey report clearly indicates that the diffusion and use of AI had substantially expanded across the landscape of worldwide digital government efforts since the time of the prior 2022 survey.

Responding to this trend, the second motivation for this project was to attempt to capture lessons learned related to piloting and/or deploying AI systems across a range of different types of country or regional public sector settings, also using the approach of a small sample of in-depth interviews.

The overarching motivation of this project, spanning both sets of in-depth interviews, was to generate rich, qualitative content about lessons learned and related reflections that would provide insights into some of the important realities, challenges and nuances of implementing AI and other digital innovation initiatives. The content of these ten interviews will be disseminated as case

studies and used to help government officials better appreciate the issues of translating their high-level AI and digital technology related polices into actual implementation related practice, and to better understand important micro aspects of planning, piloting and deploying these types of projects.

This type of knowledge is meant to support the development of capacity to plan, test and implement AI and other digital initiatives. The combined content in Part 1 and Part 2 of this report will be informative to public sector officials responsible for overseeing or supervising AI and digital efforts across the full spectrum of socio-economic development: least developed countries, developing economics, economies in transition, and the highly developed economies.

In addition, targeted efforts will be made by UN DESA staff to distribute this report to the group of countries and communities involved with UN Development Account sponsored projects related to any type of digital innovation, as they should find this content to be especially useful.

A related motivation for compiling these case studies was the belief that this type of micro-level, in-depth exploration of these public sector AI and digital change efforts would provide helpful input to the UN DESA team that designs and administers the bi-annual E-Government Survey as they assess if they should make any revisions to the 2026 survey questions or report analysis related to managing the testing, piloting and evaluation AI initiatives. The UN DESA staff associated with the E-Government Survey will study this report with these considerations in mind as they do their planning for the 2026 survey.

1.2 A small first step in the direction of understanding bigger questions

The use of AI in e-government has important implications for the Frontier Technology Policy Experimentation and Regulatory Sandboxes in Asia and the Pacific (DA Project 2124B) effort, and for any subsequent follow-on UN Development Account efforts involving the need to pilot digital technology applications as a means of supporting sustainable development goals. Using AI within digital government for either internal or external service delivery creates governance challenges.

The current and future speed of AI capability development and the breadth of public sector organizational deployment efforts far outpaces the speed of traditional policymaking. This is compounded by the scale and reach of AI's potential impacts both on the positive side for productivity improvement and enablement of new and better services, and simultaneously, on the negative side for amplifying existing types of risks as well as for the potential to create new and sometimes unanticipated risks and problems.[4]

These challenges play out in every sector, including public sector. Governments must grapple with the challenge of how to integrate AI-related matters, including governance structures and the nature and extent of regulatory regimes, across existing structures, or whether to create new structures to deal with them.

Over the next 5 to 10 years, the UN DESA Division of Public Institutions and Digital Government, in conjunction with UN ESCAP and other UN partners involved in Digital Innovation related Development Account Efforts and the E-Government Survey, want to become better informed on the following trends and questions:

- To what extent are governments, especially those in developing countries or in smaller countries, creating or strengthening their institutional capabilities for agile and iterative-based early-stage experimentation with digital initiatives and AI applications? Also, to what

extent have they created structured ways of implementing, managing and evaluating the various pilot stages of projects? And similarly for projects where the government has chosen to proceed to larger-scale operational deployment?

- What are the issues encountered as they attempt to do this? Are their special issues associated with managing the public's sense of trustworthiness, or of responsible approaches to usage?
- What are some examples of approaches countries are taking to navigating the inherent trade-offs between managing risk, catalysing innovation, achieving economic benefit and viability, and building and retaining social trust?
- What has been the experiences of selected countries with respect to their efforts to put in place approaches to digital and AI related experimentation, piloting and operational deployment they felt were contextually relevant, systematic and appropriately governed?
- To the extent that some of these countries did this, was it helpful? Did a more structured approach to experimentation and piloting (including sandboxing as appropriate) lead to useful improvements in decision making and evaluation? Or to useful improvements in follow on efforts for operational deployment? How did the benefits of these efforts compare to the costs?
- As more digital government initiatives make use of various types of AI in one way or another, what are lessons learned for creating and using sandboxes, and for piloting? How can these earlier (per-operational deployment) phases be better used as a means to chart and steer the path forward in guiding the effective, inclusive, accountable and contextually relevant use of AI in a country's public sector? And additional, to achieving aims aligned with one or several of the UN sustainable development goals (SDGs)?

The pursuit of these types of questions requires the ability to understand and unravel many complex issues across a wide range of country specific settings, and that requires both a broad and deep investigation effort over an extended time period.

In contrast, this initial exploratory effort to harvest and synthesise lessons learned was small in scope and constrained for timeline and budget reasons to three key country specific case study interviews related to the Development Account projects, and seven additional supplementary interviews from several other countries related to public sector AI usage. This is only a very modest first step in the direction of addressing these bigger questions. Even with these limitations, we hope that this effort serves as a starting point for other digital government policy investigation and case study efforts to be pursued in the future by the UN as well as by other organizations and countries.

1.3 Administrative mechanism for sponsoring this project

This project was funded by the UN Development Account for Frontier Technology Policy Experimentation and Regulatory Sandboxes in Asia and the Pacific (DA Project 2124B) as a small, follow-on supplement to the three main project efforts sponsored by that account with the governments of Bangladesh, Kazakhstan and the Maldives. The project was set up as a consultancy on "Digital Policy Experimentation and Artificial Intelligence (AI) Sandboxes for SDGs," and

supervised by the UN DESA Department of Public Institutions and Digital Government, Digital Government Unit. The Information and Communications Technology and Disaster Risk Reduction Division (IDD) of United Nations Economic and Social Commission for Asia and the Pacific (ESCAP) was also involved in the approval and final review of this project. Staff members from these UN organizations were involved in providing access to in-country experts and in reviewing drafts of the report.

2. The ten in-depth interviews for capturing lessons learned

The ten people interviewed, and the topics discussed within each interview, are introduced in this section in Table 1 and Table 2 respectively. The full written summary of each interview is contained in the companion document, Lessons Learned From Sandboxing, Piloting and Policy Experimentation with AI and Other Digital Initiatives: Part 2, Ten In-Depth Interviews. The overview information given in Table 1 and Table 2 will aid in the determination of which of these ten interviews, as well as which parts of each interview, you may be interested in reading in more depth by viewing the content in the Part 2 companion document.

2.1 Framework for presenting the ten people interviewed

In Table 1, the ten interviews are presented in two main clusters (A and B). Interview Cluster A contains four interviews that all deal with digital innovation projects for enhancing national capacity building. The first three of these interviews are based on the policy experimenting and sandboxing efforts co-sponsored by UN DESA and UN ESCAP in conjunction with the governments of Bangladesh, Maldives and Kazakhstan. The fourth interview focuses on Singapore's own efforts (without UN support) to build up the capability to securely use the commercial cloud for delivering a growing number of government digital services.

Artificial intelligence (AI) is not even mentioned in the first three interviews. That does not mean that the use of AI in any form was totally absent from these three projects. It means that the use of AI, if it was used at all, was not a key part of these three efforts, at least not during the initial sandboxing and experimentation phase that was the focus of the interview. As such, in reviewing lessons learned for these first three projects, the three people interviewed focused on various aspects of managing the sandboxing, experimentation and policy making aspects related to their respective digital transformation efforts without seeing the need to make any mention of AI related issues.

The fourth interview in Cluster A provides a bridge between the three prior interviews in that cluster that deal with complex digital transformation innovation efforts that did not involve the use of AI in any significant way and those following in Interview Cluster B which do. That interview, with Singapore's former Chief Digital Technology Officer, focuses on issues related to managing the complex digital transformation of transitioning to government usage of commercial cloud services. While this interview is not focused on the use of AI, it notes that transitioning to commercial cloud usage is an important enabler for a government to make use of large AI models and for enhancing overall digital application development and deployment capabilities which increasingly include AI-enabled applications.

Interview Cluster B contains six interviews that all explicitly focus on using AI-based applications for delivering digital government services, and some of the interviews also address related policy making and governance aspects of managing AI usage across the economy. Two of

these interviews provide big picture overviews of public sector AI usage and related policy experimentation and sandboxing. The interview with Carnegie Mellon University Professor Ramayya Krishnan is mostly based on his extensive knowledge of these topics across federal, state and local levels of government in the United States, and also includes some mentions of other global examples. The interview with Professor Gianluca Misuraca is based on his extensive knowledge of these topics across the member states of the European Union and at the overall EU level as well.

The next three interviews in Cluster B focus on managing country specific efforts to use AI for the delivery of government digital services. Government officials directly involved in supporting and overseeing such efforts from Uzbekistan, Rwanda and Singapore share their experiences with piloting and deploying AI-enabled government digital services in their respective country settings. The last interview in Cluster B focuses on the use of AI for supporting decision making in one particular public sector domain setting: local level (city, municipal, or county/district level) social services related decision making for highly sensitive and high-risk types of decisions that case workers need to make. Professor Rhema Viathianathan started this line of work in her home country of New Zealand and has supported local government agencies in multiple locations across the United States and in several other countries to development, test, pilot and evaluate these types of applications.

Table 1: Thematic grouping of the 10 in-depth interviews

Thematic Grouping of Interviews	Item #	Country focus of interview content	Name and title of people interviewed
Interview Cluster A: Digital Innovation Projects for National Capacity Building			
Policy Experimentation and Sandboxing Efforts Co-sponsored By UN DESA and UN ESCAP	1.	Bangladesh	**Bijon Islam** CEO, Lightcastle Partners consulting firm; Expert on national development projects in Bangladesh; Consultant for Bangladesh-UN project to sandbox and pilot a solution to increase access to capital for micro and small enterprises.
	2.	Maldives and multiple other countries doing CBDC pilots	**Gordon Clarke** Managing Director, Monetics Pte Ltd consulting firm; Expert on e-payment, Central Bank Digital Currency (CBCD) and fintech; Consultant for Maldives-UN project to plan a CBDC sandbox effort, and consultant for numerous other CBDC national projects.
	3.	Kazakhstan	**Sayran Suleimenov** Formerly with the Project Management Department, KOREM (owner of the centralized electricity trading market in Kazakhstan); Participant in the Kazakhstan-UN project to improve the Kazakhstan electricity industry infrastructure.
Singapore's Transition To Using the Commercial Cloud for Selected Government Digital Services	4.	Singapore	**Cheow Hoe Chan** Senior Advisor, Singapore Economic Development Board; Former Government Chief Digital Technology Officer of Singapore; Former Deputy Chief Executive of the Government Technology Agency of Singapore.
Interview Cluster B: Using AI in the Public Sector			
Big Picture Overviews of Policy Experimentation and Sandboxing	5.	United States, with some other global examples	**Ramayya Krishnan** Professor and Dean of Heinz College of Information Systems and Public Policy, Carnegie Mellon University; Member, US National AI Advisory Committee; Faculty Director, CMU Block Center for Technology and Society; Lead Research Coordinator for the CMU/NIST AI Measurement Science & Engineering Cooperative Research Center (AIMSEC).
	6.	European Union	**Gianluca Misuraca** Professor, Polytechnic University of Madrid; Executive Director of AI4Gov International Masters on AI for Public Service co-sponsored by the EU; Founder and Vice President of Technology Diplomacy at the Inspiring Futures

			consulting firm; Consultant to the UN E-Government Survey; former Senior Scientist at the EU's Joint Research Centre.
Country Specific Overviews of Policy Experimentation, Sandboxing and Piloting	7.	Uzbekistan	**Dilshat Saitov** Head of Division for Cooperation with International Rating Organizations, Digital Government Projects Management Centre, Ministry of Digital Technologies. **Nigmatullo Sharafutdinov** Head of Division of Introduction of Electronic Public Services and Interdepartmental Electronic Cooperation, Ministry of Digital Technologies. **Jahongir Topildiyev** Chief Specialist of Division of Introduction of Electronic Public Services and Interdepartmental Electronic Cooperation, Ministry of Digital Technologies.
	8.	Rwanda	**Esther Kunda** Director General, Innovation & Emerging Technologies, Ministry of ICT & Innovation.
	9.	Singapore	**Dominic Chan** Chief Information Officer and Assistant Chief Executive for Product Management, Government Technology Agency of Singapore.
Domain Specific Overview of Policy Experimentation, Sandboxing and Piloting In High-Risk Social Services Applications	10.	US and New Zealand	**Rhema Vaithianathan** Professor, Auckland University of Technology; Director, AUT Centre for Social Data Analytics; Consultant to numerous social service agency efforts to use AI-based predictive risk models to provide decision support to social service case worker in areas related to child protection and housing assistance for homeless people.

2.2 Listing of main topics covered in each interview

Table 2 provides a listing of the main topics covered in each of the ten interview summaries. The table preserves the structure of Table 1 in terms of grouping and ordering the interviews by Cluster A and Cluster B.

In the 4th column of Table 2 titled "Main Topics Covered in Each Interview Summary," the bullet points are extracted from the headers and sub-headers that appear in each of the full-length interview summary write-ups contained in Part 2 of this report. As such, these bullet points provide an accurate way of identifying the topics covered in each interview and provide a sense of the abundance of rich and relevant content captured across these ten interviews.

Another summary of each of these ten interviews in the form of a text narrative is given in Section 3.

Table 2: Main topics covered in each interview summary

Thematic Grouping of Interviews	#	Country Focus of Interview Content	Main Topics Covered in Each Interview Summary
Interview Cluster A: Digital Innovation Projects for National Capacity Building			
Policy Experimentation and Sandboxing Efforts Co-sponsored By UN DESA and ESCAP	1.	Bangladesh Bijon Islam	• Introduction to Bijon Islam • Background on Bijon's involvement with the Bangladesh project • The meaning of policy experimentation and sandboxing in the context of this project • Lessons learned from this sandboxing and pilot effort • Lesson #1: Don't make the scope of the pilot too wide. • Lesson #2: Quickly find partners that can move quickly to implement the pilot. • Lesson #3: Don't underestimate the technological complexity needed to implement the sandbox and pilot. • Lesson #4: Don't "over-workshop" with large workshops to the extent of causing long delays in starting actual piloting. • Lesson #5: Maximize the synergy with the few key government entities most relevant to the narrowed pilot scope as early as possible. • Lesson #6: Do in-depth knowledge exchange with external experts early on. • Lesson #7: Get strong private sector partners with the right motivations involved in the pilot as early as possible. • Other key points. • The role of sandboxing in piloting policies, rules and governance. • The importance of having government rules that allow for sandboxing and policy experimentation within the sandbox. • Using the sandbox to navigate across the spectrum of existing rules that can be applicable, that cannot be practically applied in the new digital setting, or that might not exist at all. • Risks associated with the digital services solution being tested in the sandbox. • Cyber crime and cyber security. • The Digital Divide. • The benefits of better CMSME access to cash & finance being piloted in the sandbox. • Potential implications of successfully piloting and scaling this access to cash platform on existing microfinance providers. • Summary reflections on the phases of sandboxing. • The conceptualisation phase of the sandbox. • The operations phase of the sandbox. • The evaluation phase of the sandbox.

	2.	Central Bank Digital Currency sandbox efforts in multiple countries, including Maldives Gordon Clarke	• Introduction to Gordon Clarke. • The Maldives UN ESCAP project: Planning for a Central Bank Digital Currency and supporting sandboxing for policy, regulatory and execution experimentation. • The follow-on effort to publish a Global Toolkit on regulatory sandboxing for Central Bank Digital Currency and FinTech. • The need for clarity on economic and financial policy goals when moving ahead with sandboxing for CBDC projects. • CBDC projects can be retail focused or wholesale focused. • Examples of Retail focused CBDC projects in developing countries. • Examples of wholesale focused CBDC projects in developed countries. • The origin and evolution of regulatory sandboxing in the context of digital innovation projects. • Clarifying the terms testing, sandboxing and piloting in the context of CBDC and other FinTech projects. • Testing. • The importance of including ease of use (usability) testing in addition to technical and functionality testing in the test phase. • Sandboxing. • Considerations for the types of participants in the sandbox phase. • Piloting and examples of recent CBDC pilots. • Large scale CBDC pilot efforts. • Continuing with monitoring, adaptation and regulatory discovery after the sandbox phase and into the pilot phase. • Elaborating on the purpose of sandboxing. • The risks of bypassing the sandbox phase and jumping directly to the pilot phase when operating in a regulated environment. • Applying regulatory sandboxing concepts and practices to other industries beyond Financial Services.
	3.	Kazakhstan Sayran Suleimenov	• Introduction to Sayran Suleimenov. • Origin of the concept of a digital platform for Kazakhstan's electric power industry. • The Innovation Award from UN DESA and the Kazakhstan Ministry of Digital Development for the Digital Energy Platform. • Getting started on initiating a regulatory sandbox and on building the Digital Energy Platform. • Pilot testing of the Digital Energy Platform. • Lessons learned from our sandbox piloting and testing the Digital Energy Platform. • Intention to expand the usage and scope of the Digital Energy Platform to meet the needs of Kazakhstan's electricity sector over the next decade.

			• Summary of Key Challenges and Objectives for Kazakhstan's Electricity Sector and the Digital Energy Platform Project. • Current Challenges in the Electricity Sector. • Objectives of the Digital Energy Platform Project. • Key Tasks to Achieve Project Objectives. • Integration with the Digital Energy Platform. • Strategic Importance. • Plans for two stages of follow-up development efforts after the conclusion of the pilot.
Singapore's Transition To Using the Commercial Cloud for Selected Government Digital Services	4.	Singapore Cheow Hoe Chan	• The challenges of transitioning to cloud and digital are underestimated. • Understanding the origins and progression of the cloud: from infrastructure only to the global ecosystem of software services. • Getting the government to understand the multiple reasons for moving to cloud: scalability, resiliency, ecosystem for software services and application development. • Making the paradigm shift required to transition to the cloud: dealing with the fear of the unknown. • Early low-risk cloud pilots to test and learn. • Cloud vendors needed to be less opaque and make cloud understanding less opaque. • Internally building our government capability to use cloud. • How a relatively small number of cloud technologist and application developers made a big impact across the entire organisation. • Transitioning beyond cloud usage for unclassified information. • All paths lead to using the cloud for many civilian government services. • Why do cloud companies site data centres in Singapore? • For a government to get started using cloud for non-classified data, you do not have to wait until a cloud service provider locates a data centre in your country. • A suggested mindset for getting started with moving government e-services to the cloud. • Keep focused on the real-world problem you are trying to solve vs using new technology for its own sake.
Interview Cluster B: Using AI in the Public Sector			
Big Picture Overviews of Policy Experimentation and	5.	United States, with some other global examples	• Introduction to Ramayya Krishnan's background and involvement in public sector AI. • The importance of policy experimentation and sandboxing for AI applications in public sector settings. • We need better tools for measuring and evaluating *AI.*

Sandboxing		Ramayya Krishnan	• *Pe*nnsylvania creates a governing mechanism for AI-related policy experimentation and sandboxing. • Questions related to due diligence, governance, procurement and the necessary talent required to do these things given AI is moving so rapidly. • A micro-level AI preparedness index for local levels of government would be useful. • Different types of AI governance. • New regulations specific to AI versus use of existing laws and regulations already governing processes or outcomes. • Vertical versus horizontal regulation and different jurisdictional scopes (city, state, national, multi-national). • Broadening of the meaning of "what is AI" and implications for AI governance. • AI governance for a public sector enforcement agency versus a service delivery agency. • Examples of AI use cases to support sensitive types of social services decision making at the local government level. • Example #1: Determining who is eligible for a public assistance programme but not enrolled. • The pilot effort for evaluating the new AI supported approach for determining who is eligible but not enrolled. • The partnership between the local government agency and the university. • The pilot effort is much more than getting the AI predictive model to work: you need to look at the larger socio-technical system. • Example #2: Deciding how to respond to allegations of child abuse. • How do you carefully deploy AI to support a decision that is so hugely consequential. • Concerns with bias in this type of decision making. • Using sandboxing to understand the nature and implication of false positives and false negatives. • Finding the best approach in a given use case setting for creating human augmented systems, and in some cases even automated systems. • The challenge of training data bias and the example of resume screening. • Defining your playbook to use sandboxes as a way of building capability and capacity. • Initial steps and questions for moving forward with a playbook and capability development. • Addressing talent related capability limitations and the possibility of accessing "talent-in-the-cloud". • Data inventory, data governance and cloud infrastructure.

			• Does your country need the capability to develop the largest scale AI models? Or only to deploy and use them? • AI Governance challenges when the software and infrastructure supply chain spans providers in multiple countries. • Concluding recommendations and suggestions. • Recommended steps and questions for moving forward with playbook and capability development for AI. • Suppose there was the equivalent of a global "CERN-like" entity that can provide less developed countries with cloud-based GPU access for public sector AI model testbedding. • Key steps for getting countries at different levels of AI capability and maturity to experiment with AI technologies in beneficial ways. • Suppose international organisations could help enable access to the "digital public goods" that less developed countries need to progress with using AI.
	6.	European Union Gianluca Misuraca	• Introduction to Gianluca Misuraca's involvement in public sector AI and the AI4Gov International Master in AI for Public Service. • What AI related policy experimentation and sandboxing means to me. • Learning to better integrate across policy making, service delivery and regulation through experimentation and sandboxing that harnesses the potential of new technology. • The importance of conceptual reframing as part of policy experimentation and related sandboxing efforts that involve using AI. • The EU's AI Act and its implications for policy experimentation and sandboxing. • The EU and member states are initiating experimentation and sandboxing efforts to test compliance with the new AI Act. • Experimentation and sandboxing for AI applications can be done for other purposes, not just for AI Act compliance testing. • Avoiding getting stuck in the syndrome of never moving beyond piloting. • The rapid increase in public sector AI applications across the EU since 2020. • The 2021 study highlighting several problematic examples of algorithmic or AI-based decision making in the public sector. • Sandboxing as a more controlled and careful way to learn about the issues of complying across a range of public sector settings. • Thoughtfully managing errors and risks and moving forward under uncertain conditions with AI. • Even without algorithmic or AI support, there are known problems with decision making. • Dealing with fears of moving forward given the many uncertainties about the impacts of using AI.

			• Can algorithmic accountability lead to "automated grace" and more compassionate use of algorithms (including AI systems). • Using a sandbox to better understand the nature and consequences of errors and algorithmic transparency when using AI tools. • Public sector use case example: The Italian pension system. • Additional background on the EU AI Act and the current window of opportunity to experiment with compliant AI approaches before enforcement comes into effect. • Healthcare and education would be high potential (and high risk) settings for experimenting with compliant AI applications and policies. • The importance of the conceptualisation phase for regulatory sandboxing as well as for designing the regulation. • Bridging the gap between the content of the AI Act and everyday practice and organising across the EU for enforcement and oversight. • Steps towards frameworks and tools for assessing compliance with the AI Act starting with the ALTI tool. • Organising across the EU for AI Act governance and enforcement oversight. • How can policy makers and civil servants more effectively learn from our ongoing experiences and experimentation with AI and embrace the complexity of these epochal changes to help society.
Country Specific Overviews of Policy Experimentation and Sandboxing	7.	Uzbekistan Dilshat Saitov Nigmatullo Sharafutdinov Jahongir Topildiyev	• Overview of Uzbekistan efforts with digital government and AI applications. • More emphasis on proactive digital services. • Our usual steps for sandboxing and piloting. • Our approach for working with ministries to bring additional services online. • Plans for increasing AI usage in our online government digital services and implications for sandboxing and piloting.
	8.	Rwanda Esther Kunda	• Introduction to Esther Kunda and her portfolio. • Prior Rwanda AI efforts with local language-based chatbots. • Example #1: Ongoing pilot of an AI chatbot system to support community health workers in villages. • Stages of piloting across the language localisation loop (Example #1). • Risks to manage as we proceed with further piloting and scaling of this AI chatbot solution for supporting community health workers in villages (Example #1). • Example #2: Piloting and deployment of an AI chatbot to support customer service complaint response across the banking sector. • The bigger strategic importance of more rapidly resolving the smaller customer complaints (Example #2).

			• The central bank's process of collaborating with the banks to design and implement the centralized AI chatbot for customer complaints (Example #2). • Ongoing refinements to the customer complaint chatbot (Example #2). • The evolution of how Ministry of ICT and Innovation works with the other ministries on digital transformation and AI efforts. • Interweaving the vertical roles of the various ministries with the horizontal role of MinICT for simultaneously driving innovation and coherent digital and AI related policy experimentation. • Rwanda's national AI policy and related efforts. • Commercial Cloud. • "Big Picture" challenges as we continue moving ahead with digital transformation, AI and other emerging technologies. • Making Rwanda a proof-of-concept hub for national scale piloting and "learning-by-doing" policy design. • The role of Carnegie Mellon University Africa in building up Rwanda's manpower and ecosystem for digital and AI innovation. • Engaging the innovation ecosystem through CMU Africa student internships and practicum projects. • Suggestions for other small countries moving ahead with digital transformation and AI. • Suggestion #1: The importance of contextualisation. • Suggestion #2: Utilizing the advantages small countries have with piloting emerging technologies.
	9.	Singapore Dominic Chan	• Dominic Chan's role and background. • My understanding of the purpose and meaning of AI policy. • Sandboxing, piloting and experimentation - including policy experimentation - are a regular part of our product management efforts. • Identifying the risky assumptions and underlying hypotheses that need to be tested and evaluated. • Navigating through the process of policy experimentation. • Knowing when you need to do policy investigation or experimentation and how to frame it. • Retaining human content curation - and augmenting it with AI support tools - to manage the quality control and risk of using Large Language Models for chatbots. • Understanding the technology (including AI) well enough to establish the boundaries of trust and the appropriate risk management measures. • Implications of more AI usage for Product Management within government digital service units – don't lose sight of the basics. • Staying focused on doing things *for* citizens and not *to* citizens.

			• Comments on phases of the sandbox lifecycle. • Conceptualisation phase. • Operations phase. • Evaluation phase. • Advice for learning from the digital services and related AI efforts in other countries. • Concluding thoughts: balancing the ability to try with the discipline to manage risk.
A Domain Specific Overview of Policy Experimentation and Sandboxing In High-Risk Social Services Applications	10.	US and New Zealand Rhema Vaithianathan	• Introduction to Prof Rhema Vaithianathan and her work. • Early career realization of the gap between building models that make "good" predictions and providing useful tools for real-world decision support. • The start of applying predictive risk modelling to child welfare. • The predictive risk modelling partnership with the Department of Human Services in Allegheny County, Pennsylvania, USA (Greater Pittsburgh metro area). • Expanding the predictive risk modelling partnership with Allegheny County beyond child abuse to include homeless housing and other areas. • An example where the use of an analytics-AI based predictive model helped to reduce bias in decision making. • Applying and evaluating our predictive risk modelling work in other geographic locations in the US and internationally. • Our "guard-rail "guidelines for the ethical development and adoption of predictive decision support tools for high stakes social services. • How the illusion of validity influences the initial way case workers respond to the availability of our tool. • The practical challenges of social workers learning how to make better decisions given the context of their work. • As case workers get familiar with the predictive risk tool over time, they learn to appreciate that its data-driven, probabilistic recommendations are useful, even if imperfect. • Why our team focuses on predictions for high-stakes social service decisions where extreme adverse events only occur infrequently. • The end-to-end process for developing and piloting our predictive risk models in high-stakes social services settings. • The four steps related to model development, testing and validation prior to deployment for field piloting. • The two major steps of post-deployment field piloting. • Engaging and informing the community. • The recent field pilot of one of our predictive risk models at the Los Angeles Department of Children and Family Services.

			• Accepting the reality of multi-year periods for the end-to-end AI project effort. • More emphasis is needed on post-deployment evaluations of real-world model usage impacts versus pre-deployment technical studies of the model's predictive accuracy. • Views on the role of using Generative AI with unstructured data vis-a-vis AI-based predictive modelling with structured data. • A transitional Catch 22 situation with getting municipal level social service agencies familiar with using GenAI. • Concluding thoughts and suggestions for public sector agencies. • Understand the reality of how noisy decision making actually is unless front-line workers making the decisions have good support. • Use AI-based decision support for helping human decision makers, not for criticizing past or current performance. • Emphasizing the functionality and use of the support tool more so than emphasizing the use of AI for its own sake. • While there are controversies related to using these types of predictive risk tools, the alternative of not using any type of algorithmic support is also problematic.

2.3 Background on the projects in Bangladesh, Maldives and Kazakhstan sponsored by the UN Development Account 2124B

Interviews #1, #2 and #3 have a common factor in that the project that was the setting and subject for each of these three interviews was sponsored by the UN Development Account for Frontier Technology Policy Experimentation and Regulatory Sandboxes in Asia and the Pacific (DA Project 2124B). The UN website for this overall DA Project 2124B effort[5] summarises the project initiatives undertaken in each of these three countries that were sponsored by this Development Account funding, and also summarises the overall objective of undertaking these initiatives within these three countries as follows:

> *Frontier technologies carry a promise to fast track the Sustainable Development Goals (SDGs) through supporting innovative, forwarding-looking policies and solutions. There are, however, numerous risks and complexities of digital technologies that come along with those opportunities, as well as policy and regulatory challenges….The overall purpose of the project is to enhance the institutional capacity of selected countries in special situations, namely the: (i) least developed countries (LDCs); (ii) landlocked developing countries (LLDC); and (iii) small island developing States (SIDS). The specific focus of the project is to conceptualize, develop and implement policy experimentation and/or regulatory sandboxes on new technologies, as an innovative and catalytic approach to accelerate the progress of the 2030 Agenda for Sustainable Development.*

There were two project initiatives in Bangladesh. One initiative was to establish a regulatory sandbox in order to experiment with how to provide Cottage, Micro, and Small & Medium (CMSME) enterprises with four types of improved access: (a) Access to Market; (b) Access to Service; (c) Access to Finance and (d) Access to Skills. The second initiative was to provide enhanced access to broadband to further broadband equity with the objective of leaving no one behind as a result of the inability to have access to broadband.

The interview for this follow-up investigation of lessons learned only considered the first initiative related to improving CMSME access to market, services, finance and skills. Bijon Islam, the person interviewed for this project, served as a senior consultant to UN DESA for the Bangladesh Development Account project and along with his consulting firm was responsible for the planning, project management and deployment effort for this CMSME access effort.

The Maldives effort focused on doing both the high level and detailed level planning required to define, implement and operate a regulatory sandbox in order to experiment with introducing and using a Central Bank Digital Currency (CBDC). Gordon Clarke, the person interviewed for this follow-up investigation of lessons learned, served as a senior consultant to UN ESCAP for the Maldives Development Account project and had lead responsibility for the planning, project management and report writing for this CBDC planning effort, including briefing Maldives government officials on the status of CDBC pilot projects in other parts of the world.

There were two project initiatives in Kazakhstan. One initiative was focused on capacity development and uplifting in the country's energy sector in ways that would support the decentralization of electricity generation, digitalization of the electric power supply and value chain, and decarbonization. This involved the planning, design and implementation of several different types of demonstration projects for doing the necessary physical, digital and policy related experimentation, which also included regulatory sandboxing. The second initiative was to explore the possibility and feasibility of setting up a regulatory sandbox to experiment with the use of autonomous vehicles in public transport.

The interview for this follow-up investigation of lessons learned only considered a subset of the first initiative related to improving the country's electricity industry. Sayran Suleimenov, the person interviewed for this project, was formerly the Head of the Project Management and Digital Transformation Sector of the Analytics and Project Management Department of the "Kazakhstan Electricity and Power Market Operator Joint Stock Company" ("KOREM" JSC). He played a key role in the effort to design, implement and test a unified digital energy platform across Kazakhstan's electricity generation and distribution industry.

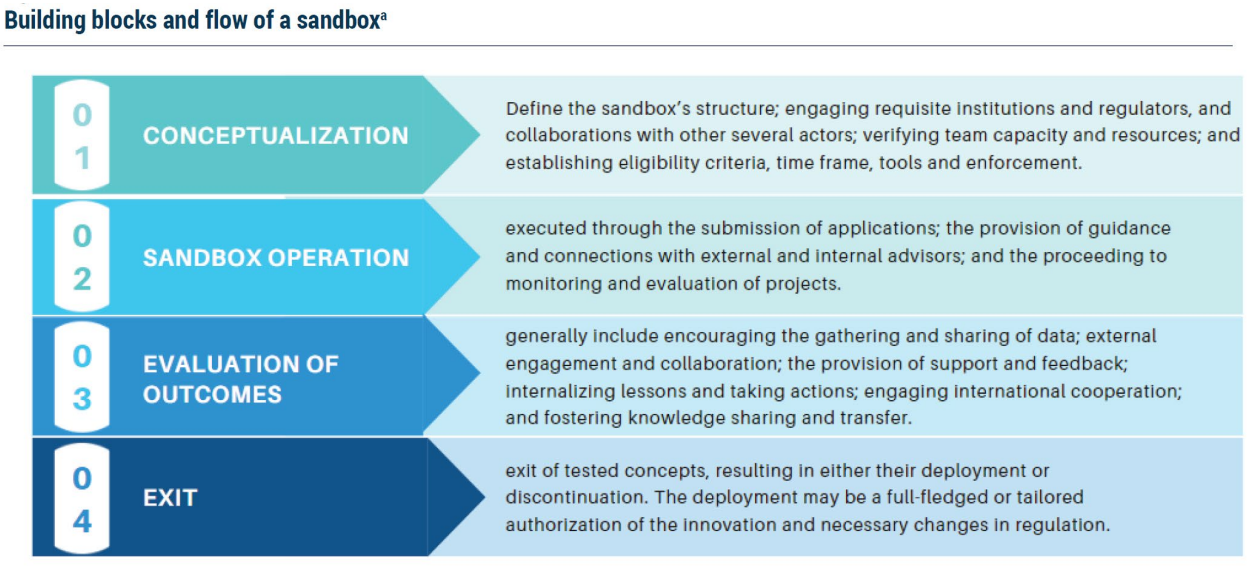
2.4 The definition of a regulatory sandbox from the UN DESA Policy Brief on "Sandboxing and Experimenting with Digital Technologies for Sustainable Development"

The terms "sandbox" and "sandboxing" appear throughout this report. What is meant by these terms? The participants in the projects from Bangladesh, Maldives and Kazakhstan interviewed for this report (Interviews #1, #2 and #3) were formally briefed on the meaning of a "regulatory sandbox" and the process of sandboxing as part of their project effort with their respective country and UN DESA and UN ESCAP. They were familiar with the definition of a sandbox given in UN DESA Policy Brief #123, "Sandboxing and experimenting digital technologies for sustainable development."[6]

A sandbox is a framework or environment that allows for the live testing of technologies, services and business models in the market with real consumers, while benefiting from relaxed or flexible regulatory requirements, often at a smaller scale and on a time limited basis, and with appropriate supervision and safeguards in place. Because of the availability of a conducive and contained space through a more lenient or flexible regulatory environment, a sandbox can facilitate the pertinent deployment of digital technologies and their rapid and effective inclusion in digital government. A sandbox not only insulates innovations from current regulations which may be restrictive or non-existent, but it also insulates against possible negative impacts on end-users.

A sandbox has the potential to meet several objectives, including both regulatory and institutional. While regulatory objectives are most commonly limited to security, inclusion, consumer protection, competition, sustainability and among others, institutional objectives may be wider in scope, such as supporting the digital government ecosystem or encouraging collaboration with the private sector and other stakeholders. Sandboxes, however, are not universally designed for all challenges confronting regulatory and policy-making bodies faced with challenges in disruptive innovations.

They were also briefed on the definition of the building blocks and stages of a regulatory sandbox that were given in that same UN DESA Policy Brief #123, summarised in Figure 1 below.



a Author's adaption from various sources, including (i) Wechsler, Michael, Leon Perlman, and Nora Gurung. "The state of regulatory sandboxes in developing countries." Available at SSRN 3285938 (2018).

Figure 1 Building Blocks and Flow of a Sandbox for Policy or Regulatory Experimentation from UN DESA Policy Brief #123

The other six interviews (#4 through #10) were with people who had no linkages with the UN Development Account 2124B projects in Bangladesh, Maldives and Kazakhstan, and as such, were unfamiliar with UN DESA Policy Brief #123 and the specific definitions of a sandbox contained

within it. They had received a copy of Figure 1 from a pre-interview briefing document sent to them as part of being invited to participate in the interview for this lessons learned project. However, they had not been through training sessions or workshops where they had to pay attention to these specific definitions of a sandbox and the processes of sandboxing as stated in UN DESA Policy Brief #123. This means that when the people in interviews #4 through #10 made use the terms "sandbox" or "sandboxing", they may have been thinking of these terms in their own ways and not necessarily referring to the specific definitions given in UN DESA Policy Brief #123.

3. Key aspects of each interview

Each of the ten interviews is summarised below in the form of narrative paragraphs. This alternative form of summarising each interview (versus the bullet point summary given in Table 2) is also for the purpose of helping determine which of the full interview write-ups you may want to look at in more depth in the Part 2 document of "Lessons Learned From Sandboxing, Piloting and Policy Experimentation with AI and Other Digital Initiatives."

3.1 Bijon Islam, based on the joint Bangladesh-UN digital innovation project to improve access for cottage, micro, and small & medium enterprises

Bijon Islam starts by explaining his understanding of the meaning of policy experimentation and sandboxing in the context of this project. He explains that a sandbox environment for an initial pilot can happen with a small, select group of customers to see if the new digital service works and to figure out how to fix both the technical and non-technical glitches and problems before starting to scale up the new service offering. He highlights that before a regulatory sandbox effort can take place, the country must already have an existing policy that allows for such sandboxing. There needs to be a policy for the relevant sector (in this case, financial services) that allows people to create and execute this sandbox and that defines the guidelines, and governing rules for it.

Much of the content in this interview is Bijon's elaboration on seven key lessons that he learned for this effort:

- Lesson #1: Don't make the scope of the pilot too wide.
- Lesson #2: Quickly find partners that can move quickly to implement the pilot.
- Lesson #3: Don't underestimate the technological complexity needed to implement the sandbox and pilot.
- Lesson #4: Don't "over-workshop" with large workshops to the extent of causing long delays in starting actual piloting.
- Lesson #5: Maximize the synergy with the few key government entities most relevant to the narrowed pilot scope as early as possible.
- Lesson #6: Do in-depth knowledge exchange with external experts early on.
- Lesson #7: Get strong private sector partners with the right motivations involved in the pilot as early as possible.

All seven of these lessons learned are linked to the first lesson learned related to scoping. Bijon reflected that the scope of the effort was too wide. This breadth of scope resulted in a range of complications and delays as explained in the interview summary.

This example illustrates an inherent tension in any capacity development effort. This is the tension between the "good intentions" of wanting to include a wider scope in pilot projects to potentially help more underserved segments of the economy in more ways versus the realities of execution and the complexities of getting things done, which benefit from a more constrained and focused scope.

In this effort, it was useful for the project team to do a high-level, broad scope exploration to map out the big picture view across the multiple interlinked factors that that were important to the goals and context of this project: (a) Access to Market, (b) Access to Service, (c) Access to Finance and (d) Access to Skills for cottage, micro, and small and medium (CMSME) business entities. The challenge was in attempting to do follow on execution efforts across all these fronts in parallel. Eventually, they had to readjust the execution plan, and focus only on one of these pillars, access to finance.

One reason to keep adding more to the scope of an execution plan for a national pilot project may be the concern that there may not be a second or subsequent opportunity to get another follow-on pilot effort approved and funded after the initial pilot effort. This pilot might turn out to be the one and only pilot attempt. With that belief, it is natural to try and include a broader scope of work and more deliverables in the one effort that is going to be approved. However, an overly broad scope, especially for efforts requiring complex execution efforts, can jeopardize the ability to get the highest priority things done within the confines of the project's budget and timetable.

Bijon also shares on how this project used the sandbox to pilot policies, rules and governance. This involved knowing how to navigate across the spectrum of existing rules that can be applicable, identify those rules that cannot be practically applied in the new digital setting, and identify gaps where policies or rules are needed but might not yet exist.

The main risks he identified as being important to this sandbox and piloting effort are 1) cyber security and cyber crime, and 2) exacerbation of the digital divide. He also elaborates on the benefits of the solutions for providing CMSME entities better access to cash and finance that were piloted in the sandbox.

He concludes the interview with his speculations on the potential market implications of successfully piloting and scaling this access to cash platform on existing microfinance providers. He also reflected on key learning points from the conceptualisation, operation and evaluation phases of this sandboxing and piloting effort.

3.2 Gordon Clarke, based on the joint Maldives-UN digital innovation project to plan for a CBDC sandbox effort and on multiple other CBDC national pilot projects

Prior to this interview, Gordon Clarke and other members of the team that worked on this Maldives CBDC planning effort had already published two UN reports with findings and related recommendations, including lessons learned. Following these initial two reports, Gordon Clarke and other team members published a follow-on UN report titled "Global Toolkit on Regulatory Sandbox for Central Bank Digital Currency and FinTech." This was a comprehensive guide that brought together and synthesized their knowledge of lessons learned from multiple CBDC sandbox planning or execution efforts in the Bahamas, Brunei, Haiti, Kazakhstan, Maldives, Nigeria, Norway, Qatar and Saudi Arabia.

As Gordon notes in his interview, "If someone were to ask me to share all of the lessons I have learned and the recommendations I have distilled in recent years about planning and setting up

Central Bank Digital Currency projects, including how to do the sandboxing for the necessary policy and regulatory experimentation, my response would be to read this Global Toolkit document."

Given this foundation for what has already been comprehensively documented and explained (with the references for these three UN reports given in the full-length interview summary write-up in Part 2), this interview with Gordon was positioned to explore and elaborate on other aspects of CBCD and related FinTech experimentation beyond what was already covered in the three previously published UN publications.

In the interview, Gordon discusses the differences between testing, sandboxing and field piloting in the context of CBCD and FinTech efforts, given that these efforts occur in highly regulated settings supervised by central banks. Of the ten interviews conducted for this project, this interview provides the most in-depth and nuanced treatment of distinctions between the experimentation phases of testing versus sandboxing versus field piloting, and why their needs to be a sandboxing phase. He explains that the purpose of sandboxing for FinTech-related digital innovation is to ensure meeting the regulatory requirements and to understand how existing regulations apply to innovative products. He highlights the risks of bypassing the sandbox phase and jumping directly to the field pilot phase when operating in a regulated environment for CBDC or FinTech initiatives. He comments on the differences between the types of users involved in the sandbox phase versus those involved in the field piloting phase.

He explains the origin and evolution of regulatory sandboxing in the context of financial sector digital innovation projects. This provides background on how the concept of "regulatory discovery" evolved with financial sector sandboxing as part of the process of government financial sector regulators and private sector FinTech innovators jointly figuring out how to move forward with innovation in responsible and safe ways.

He emphasises the need for macro level clarity on economic and financial policy goals when moving ahead with sandboxing for CBDC projects. Even if a country attentively follows all the recommendations and guidance in the Global Toolkit document, it will not lead to a successful experiment if they are not clear on the objectives for using CBDC or do not position the experimentation in accordance with these objectives.

This leads to his explanation of retail versus wholesale motivations for using CBDC, along with his observations on some of the ongoing CBDC experimentation efforts across the world that are retail focused versus those that are wholesale focused, including some that are exploring both retail and wholesale use cases in parallel.

From a global perspective, he notes the change of focus away from some of the original thoughts about CBDC as eventually becoming widely used within and across countries for all forms of payment (and therefore becoming very important for retail accounts), towards more wholesale uses of CBDC together with tokenization of commercial bank deposits as part of creating more efficient ways of managing interbank payments, particularly cross-border and in the securities markets.

While this discussion is specific to CBDC related experimentation efforts, it raises the more generally applicable point that any specific sandboxing and field piloting effort done as part of a national level digital innovation experimentation initiative in a regulated sector of the economy must be appropriated aligned with the directions of broader national economic and policy goals.

3.3 Sayran Suleimenov, based on the joint Kazakhstan-UN energy industry digital innovation project

Sayran Sulimenov explains that the primary motivation behind the initiative to create and pilot a unified digital platform for Kazakhstan's electricity generation and distribution sector was the desire to optimize numerous business processes across this part of the energy sector, including reducing unnecessary paperwork, ensuring better oversight of industry stakeholders, improving data quality and reliability, and more effectively monitoring and enforcing the Ministry's instructions and regulations.

There had been prior attempts by the Ministry of Energy to create this type of platform that had not come to fruition. A joint project effort between the Kazakhstan government and UN DESA made it possible to restart and expand efforts to build and pilot this type of platform, and in parallel, to start experimenting with new approaches for regulating electricity generation and distribution. A regulatory sandbox was established as part of the joint Kazakhstan – UN DESA project effort to support both the regulation related policy experimentation aspects as well as the testing of the new unified Digital Energy Platform.

At that time, Kazakhstan lacked a specialized regulatory authority for the energy industry akin to what they had in place for the financial sector. In parallel with this project effort, the Ministry of Energy established a regulatory unit within its structure. The software modules for the Digital Energy Platform become the environment for the regulatory sandbox.

The pilot testing of the platform was done in two phases. The first phase, which was concentrated within a one-month period, only involved users from several Ministry of Energy departments and from the relevant industry-based energy associations. This phase included user education, user testing, and the collection and review of user feedback. The second phase, which stretched over a longer multi-month period, expanded the pilot testing user base to include employees from 10 large power plants. Managers and specialists from various departments within each of these 10 power plants were assigned to test each of software modules being piloted. A software module serving as a discussion forum with messaging groups was added to the Digital Energy Platform to facilitate the ability of pilot test participants to make comments and suggestions, and to facilitate communication across the community of people involved in the piloting.

Saryan provides a brief though incisive summary of lessons learned across the following 11 aspects of the project: 1) Project Definition, 2) Team Formation, 3) Role of the Government Representative, 4) Two-Stage Approach to Piloting, 5) Addressing Legislative Barriers, 6) Constant Communication with Participants, 7) Prompt Response to Feedback, 8) Challenges of Large Projects, 9) Working with Uninterested Participants, 10) User-Centered Design, and 11) Succession Planning. His comments on these 11 aspects, while brief, nonetheless serve as a foundation for improving the efficiency and success of future digital innovation efforts in this sector.

He summarises current challenges in Kazakhstan's electricity energy sector and then states the key tasks that need to be performed by the enhanced, future version of the Digital Energy Platform in order address these challenges. He concludes with a summary of two phases of subsequent development that are anticipated to occur after the completion of the pilot to make the platform ready for full deployment and operational usage. He notes that these subsequent development phases included a change in project management with the responsibilities for further software

development and database management being transferred from KOREM JSC (where he previously worked) to another affiliated unit of the Ministry of Energy.

3.4 Cheow Hoe Chan, based on the Singapore government digital innovation effort to transition to increasing usage of the commercial cloud for selective applications

In a prior series of in-depth interviews with Cheow Hoe Chan sponsored by UN DESA, he provided an insightful overview of how Singapore's civilian government steadily and systematically went about building up the capacity to create digital government over a multi-decade period, with emphasis on that journey during the 2010 to 2021 period.[7]

In this more recent interview with him held just before he stepped down from his dual Singapore government roles of Government Chief Digital Technology Officer and Deputy Chief Executive of GovTech (for Products) in mid-2024, he focused on one topic that he views as the core critical enabler for the ability to continuously enhance and expand government's ability for creating and delivering digital services: *Transitioning to the commercial cloud.*

He provides his perspective on the origins and progression of the cloud, and how it evolved from only being physical computing and storage infrastructure to *the* global ecosystem of software services. One of his big challenges in changing mindsets across the government sector per the use of the commercial cloud was getting the government to understand the multiple reasons for moving to cloud, as most people less familiar with the technology assumed that the value proposition for considering the use of the cloud was confined to infrastructure cost savings.

Cheow Hoe recalls that he had to get Singapore's government, including the Prime Minister, to understand that they needed to make this journey to cloud, and that it was not only about infrastructure. It was also about enabling the government to take a very different and more modern approach to building and delivering applications. One of his key challenges in those early days of considering and experimenting with cloud usage was to get government officials to understand that the benefits of cloud were multidimensional, going beyond the direct cost of infrastructure. He explains why scalability, resiliency, and especially the ecosystem for software services and application development are additional reasons for a government to transition using the commercial cloud for delivering certain types of services.

He explains his view that the challenges of transitioning to cloud and digital are underestimated. He elaborates on the paradigm and mindset shifts needed to make this transition, and explains challenges related to making these changes. He also explains some of the important steps the Singapore government took to build up the necessary background knowledge, capability and culture that enabled them to initiate and manage these changes and make this transition.

He argues for why he believes all paths lead to using the cloud for many civilian government services. He points out that even with the all the restrictions on using the commercial cloud for certain types of government functions and services, there are more "non-classified" and "non-sensitive" applications than most government people realize that can be transferred to the cloud with the proper precautions and cyber security protections. He puts forth suggestions for the mindset needed to get started on moving applicable government e-services to the cloud.

While this interview is not focused on the use of cloud for AI, he does make several mentions on the necessity of being part of a commercial cloud ecosystem to make use of the large-scale AI models that are now hosted by the commercial cloud providers.

3.5 Prof Ramayya Krishnan, based on experiences with US local, state and national efforts with public sector AI and related AI policy and governance

Ramayya Krishnan shares his views on the importance of policy experimentation and sandboxing for AI applications in public sector settings. He crystalizes important questions and relevant observations related to how the public sector can proceed with AI related due diligence, governance, procurement, and the necessary talent build-up required to do these things, given AI is moving so rapidly. He points to ways in which sandboxing and policy experimentation can play a useful role in building up these capabilities.

He points out that sandboxing can be used to understand the nature and implication of false positive and false negative results of AI prediction models. He uses several examples in the setting of sensitive social services related decision making to illustrate the importance of why it is so important to have this understanding, especially when using AI-based predictive models to support decisions which have huge consequences. Related to this, he elaborates on the meaning of the term "bias" in the context of using AI, as the term is often used without adequate explanation, and he comments on how sandboxing can be used to characterize the nature and extent of bias.

He highlights the need for better tools to measure and evaluate AI as there is a gap in having the right sets of tools to go from AI related policies to real world practise. This includes knowing how to identify and assemble the data sets that would need to be available to allow for the appropriate testing of AI applications to evaluate them. More fundamentally, he argues that there is a strong need to have a science of measuring and evaluating AIs. This new type of science would lead to better methods to measure and evaluate AIs, which in turn would serve as a basis for creating the standards against which evaluation is to be conducted.[8]

He outlines practical steps a government organization at any level (local, state or federal) can take for creating their own playbook for developing the capability to make use of AI. These steps involve work process understanding, use case clarity, data inventory, data governance, realistic assessment of internal talent availability and obtaining cloud infrastructure access. He emphasizes that sandboxes can play an important role in a government organisation's effort to iteratively expand their own AI capability and capacity.

He clarifies the multiple types and levels of AI governance. He notes important distinctions between i) new regulations specific to AI versus use of existing laws and regulations already governing processes or outcomes, ii) vertical versus horizontal regulation and different jurisdictional scopes (city, state, national, multi-national), iii) AI governance for a public sector enforcement agency versus a service delivery agency.

He raises the question of whether a country needs the capability to develop the largest scale AI models versus only needs the capability to deploy and use them. In this context, he comments on some of the AI governance challenges when the software and infrastructure supply chain spans providers in multiple countries.

He notes that public sector organisations need to realistically assess if the have the necessary internal talent required to do the evaluations that are part of sandboxing along with other aspects of AI related due diligence, governance and procurement. He puts forth some suggestions for addressing AI talent related capability limitations, including the possibility of accessing "talent-in-the-cloud"

One of his overarching observations is that the effort of piloting AI applications for government digital services is much more than getting the AI predictive model to work. The effort of evaluating

a pilot that includes the use of an AI application requires looking at and evaluating the larger socio-technical system.

3.6 Prof Gianluca Misuraca, based on experiences with EU member state and central government efforts with public sector AI and related AI policy and governance

In explaining his own understanding of sandboxing and policy experimentation, Gianluca Misuraca emphasises that sandboxing is an integral part of how the public sector can experiment to learn how to better integrate across policy making, service delivery and regulation. He explains that public sector experimentation, which often involves sandboxing, is not just how to get some new technical thing or method to work. It's testing how to go about it so that the policy, the law, the regulatory aspects and the service delivery can be brought together in such ways that all these facets can work together. He notes that sandboxing is a more controlled and careful way to learn about the issues of regulatory compliance across a range of settings, as well as a better way to understand the nature and consequences of errors and algorithmic transparency when using AI tools.

He highlights that for the public sector, it is especially important to have a mechanism to experiment and fail in a controlled environment. This is the essence of why sandboxing is important. Through sandboxing, a public sector organisation can learn how to thoughtfully manage errors and risks related to moving forward under uncertain conditions with using AI.

He reviews several case studies from a study he published in 2021 that documented several prior problematic examples of algorithmic or AI-based decision making in the public sector as a reminder of the type of public backlash that can occur if there are actual or even perceived problems with the use of AI or other types of algorithmic support for public sector decision making.

He discusses the EU's AI Act and its implications for policy experimentation and sandboxing. He notes that with the passage of the AI Act, the EU and member states are actively initiating experimentation and sandboxing efforts to test compliance. He points out this limited period of time between the AI Act coming into force (August 01 2024) but before enforcement becomes mandatory for most of the act's provisions (August 2026) is a special window of opportunity to do sandboxing and experimentation. He also points out that even with the current emphasis on preparing for the forthcoming enforcement of the terms of the AI Act, public sector experimentation and sandboxing for AI applications can be done for other purposes, not just for AI Act compliance testing.

He expresses that sandboxing and related exploratory efforts are needed because even with all the recent debate about AI, both the public sector and private sector are still not yet clear on the specifics of how to apply the law and on whether the impacts of the law will be positive or not. Sandboxing and experimentation are especially needed now to understand the nature of the obstacles and the problems that will result from applying the law.

He points out that the AI Act as well as Europe's Digital Service Act for online social media platforms establish the fact that there needs to be algorithmic transparency or otherwise the application's usage will not be legal, and those responsible for the application would be liable for non-compliance with these recent laws. Because of this, the issue of algorithmic transparency really needs to be addressed. Through specific types of experimentation in a sandbox, the black box behaviour of the AI or other type of analytic algorithm can be made more transparent and characterized to some degree and this is an important step towards practically realising algorithmic transparency.

He highlights the importance of the conceptualisation phase of sandboxing, noting that if the design is not done in the right way, the sandbox effort can seem to produce very good results that may not be answering the right questions. Across all four phases of sandboxing, including the conceptualisation phase, he cautions against putting too much emphasis on the technological part and not effort emphasis on the social aspects, the legal aspects, and all the other non-technological parts which are also very important to consider. Both the technological and non-technological aspects of the effort have to be tested and evaluated in the sandbox.

He provides observations on bridging the gap between the content of the AI Act and everyday practice, on organising across the EU for enforcement and oversight, and on steps towards frameworks and tools for assessing compliance with the AI Act. He also provides an overview of the AI for Government (AI4Gov) master programme co-founded by European Union that is targeted at preparing public sector professionals as well as professionals from other sectors for how to more effectively use and implement AI-based solutions in the public sector.

3.7 Dilshat Saitov and team from the Uzbekistan Ministry of Digital Technologies based on public sector AI project piloting and deployment in Uzbekistan

Uzbekistan is a medium size, lower middle income, landlocked developing country in central Asia that has steadily improved its E-Government Development Index (EGDI) score and corresponding EGDI international ranking in the prior three UN DESA E-Government surveys published in 2020, 2022 and 2024. This interview provides a glimpse into the focused and determined ways in which that country has set about expanding its scope of digital government applications, including those that make use of AI.

Dilshat Saitov and his teammates from the Uzbek Ministry of Digital Technologies discuss how they have already implemented the following three types of AI capabilities into some of their digital government services: facial identification (face ID) as part of the login and authentication process, a chatbot to respond to common questions, and a voice assistant to help those people who have trouble typing or who have trouble reading the screen more easily interact with their national portal for government services. They describe ongoing efforts to expand their use of AI capabilities to achieve their target of implementing 10 new proactive citizen services before the end of 2024 and even more in subsequent years.

They describe the way they use sandboxing and piloting to test their new digital government services, including those that make use of AI. They make a distinction between sandbox testing in their own Ministry of Digital Technology internal environment with a small number of internal government and vendor test users versus various stages of pilot testing in the setting of their regular national digital government portal using a limited set of citizens as external users.

They highlight the role their team within the Ministry of Digital Services plays in centrally coordinating and executing the process of obtaining external and internal user feedback related to any problems with a new or existing digital service, and for coordinating how to work with all relevant entities across the government to act on that feedback to resolve issues. This team plays the same type of central coordinating role for planning new or enhanced government digital services, including AI usage. They also manage the effort to obtain alignment and approval across all parts of the government involved with each new digital service, noting that this working level approval and alignment is a pre-condition to receive final approval from the Prime Minister and Cabinet to proceed with the plans.

It is notable that the final version of the proposal for providing a new government service as a digital online service is a document that is submitted to Uzbekistan's national cabinet of ministries and must be approved by the cabinet. Such proposals include a statement of agreement to participate from all the ministries involved in providing that service, and a summary of the technical plan describing what will be done and how it will be done. There is therefore strong top level government support for any new approved initiative for a new government digital service (including those using AI) given this process for cabinet level discussion and approval.

They summarise their ongoing strategic effort to study more about AI and to use AI as much as possible as they proceed with their online digital service efforts. As they do this, they will be assessing if they need to take new measures to deal with new types of risk that may result from increasing use of AI in their online government digital services. They will monitor how this situation evolves and respond accordingly.

3.8 Esther Kunda from the Rwanda Ministry of ICT and Innovation based on public sector AI and digital innovation project piloting and deployment in Rwanda

Rwanda is a small, low income, landlocked developing country in eastern Africa that has also been steadily improving in its E-Government Development Index (EGDI) score and corresponding EGDI international ranking in the prior three UN DESA E-Government surveys published in 2020, 2022 and 2024. This interview provides a glimpse into how Rwanda has positioned itself as a proof-of-concept hub for national scale piloting of AI and other digital initiatives and how it goes about "learning-by-doing" for the related policy design.

To provide context for understanding AI usage in Rwanda's public sector, Esther Kunda describes two examples. The first example is an ongoing pilot (as of the time of the interview) of an AI chatbot system to support community health workers in villages. She explains how this current effort built upon prior Rwanda government efforts to build speech-to-text models in their native Kinyarwanda language, and then to link text-based translation models between Kinyarwanda and English. This enabled them to use their native Kinyarwanda language within any type of AI solution that uses language input or output in the forms of either voice or text, and to link the local language text to AI applications based on English language text.

She elaborates on the stages of piloting across the language localisation loop, and on how they used both an off-line desk-based test approach and a real-time field-based pilot approach to stress test the new AI chatbot system with realistic case scenarios. She also comments on the risks that need to be managed as they proceed with further piloting and scaling of this AI chatbot solution for supporting community health workers in villages.

The second example she describes was the piloting and deployment of an AI chatbot to support customer service complaint response across the banking sector which involved a partnership with the private sector banks. She provides insights into the bigger strategic importance of more rapidly resolving smaller customer complaints. She describes the central bank's process of collaborating with the private banks to design and implement the centralized AI chatbot for customer complaints, and ongoing refinement efforts.

With this backdrop, she shares her observations on the evolution of how Rwanda's Ministry of ICT and Innovation has worked with the other ministries on digital transformation and AI efforts. She elaborates on some of the intricacies of interweaving the vertical roles of the various ministries with the horizontal role of MinICT for simultaneously driving innovation and coherent digital and AI related policy experimentation. Related to this, she touches upon Rwanda's national AI policy

and related efforts and their thinking about the use of the commercial cloud for government digital services.

Building on this, she summarises how important it has been for Rwanda to position itself as a proof-of-concept hub for national scale piloting of digital and AI initiatives, and she shares her thoughts on "Big Picture" challenges she faces in her Ministry ICT and Innovation role as the government continues to move ahead with digital transformation, AI and other emerging technologies. She also provides very useful suggestions for other small countries moving ahead with digital transformation and AI.

Developing the locally available human capital needed by both government and private sector to support the technology, management and policy aspects of AI and digital innovation is a critical need and a bottleneck. Related to this, Esther describes the role of Carnegie Mellon University Africa, located in Kigali and operating since 2011, in building up manpower and ecosystems for digital and AI innovation not only in Rwanda but also across other parts of Africa as well. In Rwanda, and in other countries in the region, the private sector as well as government has been increasing engaging the CMU Africa innovation ecosystem through student internships and practicum projects, as well as faculty research projects.

3.9 Dominic Chan from the Government Technology Agency of Singapore based on public sector AI and digital innovation project piloting and deployment in Singapore

As noted in the 2024 E-Government Survey, Singapore is a high income, small island state in south-east Asia that has long been recognized as one of the global frontrunners in digital government and governance. The country has consistently achieved a very high level of EGDI ranking in recent E-Government surveys due to advanced digital infrastructures, widespread adoption of cutting-edge technologies, innovative public services solutions, and strong regulatory frameworks and digital development strategies. It had the third highest EGDI in the 2024 survey, behind only Denmark and Estonia.

Dominic Chan joined Singapore's Government Technology Agency (GovTech) in 2018 after 15 years in a private sector electronics manufacturing company. In his first GovTech role, he served as a product manager and project director for the national identity product, SingPass. In his current role as Assistant Chief Executive, he leads GovTech's Technology Management Office and serves as the Product Management Practice Lead for all product managers across the organisation and concurrently serves as GovTech's Chief Information Officer.

His perspective on managing the testing, piloting and deployment of government digital services that make use of AI is strongly influenced by his deep experience as a digital services product manager as well as by the GovTech setting of having a very deep product management culture and a strong internal professional practice of product management.[9] His perspective on these matters also reflects that fact that he works within the context of an public sector organisation with deep experience across the entire lifecycle of conceptualising, designing, testing and piloting, deploying, and operating digital services systems and products, including many which include AI capabilities.[10]

In summarising his own understanding of the purpose and meaning of AI policy, he notes the need to be extremely aware of the direct and indirect downsides of using AI as well as the upsides. He emphasises that it is not a fertile approach to only focus on the downsides in terms of precautionary measures and restraints because of the reality that AI technology is moving forward and being used more widely. His thinking about AI policy is centred on the need for a deep

understanding of the trade-offs in any given use case, considering both potential threats and risks in parallel with considering beneficial aspects of using of AI. In essence, he views trade-off identification and evaluation as the key to managing AI policy and related decisions about design, testing, piloting, experimenting and deployment.

He observes that in the Singapore GovTech context, sandboxing, piloting and experimentation - including policy experimentation - are a regular part of their everyday product management efforts. As such, he does not see the need for versions of sandboxing, piloting and experimentation that are specific to AI. What is specific to using various types of AI is that it may require additional knowledge and expertise for the product manager and product team to Identify the risky assumptions and underlying hypotheses that need to be tested and evaluated as part of the regular steps and processes used throughout the product management lifecycle.

This means that the GovTech staff involved in digital services product development, evaluation and review must understand the technology being used, including the specifics of AI functionality being used, well enough to establish the boundaries of trust and the appropriate risk management measures. On one hand, there is the need to keep up with existing as well as new specifics pertaining to the various types of AI capabilities- including related data usage and management issues- and what AI usage means for the identification and assessment of risks and benefits in any given use case. On the other hand, it is important for the digital services team not to lose sight of the basics of good product management practices, especially trade-off analysis, and of how to apply them in the context of any specific use case, whether or not this involves the use of any type of AI.

He notes that not all government digital services applications that make use of AI require policy experimentation because in many use cases, it is straightforward to use the AI capabilities within existing policies. However, that is not always the case. It takes strong background knowledge of the relevant existing government policies and regulations and of the specific functioning of the AI used as part of the overall product to know whether it will be necessary to do a policy investigation or experiment. If so, there is also a need to know how to frame such an investigation to address the appropriate issues.

He recognizes that because the overall context of the tech environment is evolving rapidly, especially with respect to AI, it is important for GovTech and other parts of the Singapore government to adjust their policy positions from time to time to remain relevant.

He concludes the interview discussion with his thoughts on the different phases of the sandboxing lifecycle. He also shares advice for how one country can be more effective in learning from the experiences of another country per their efforts with digital services and related AI efforts. His concluding thought is that all comes down to balancing the ability to try with the discipline to manage risk.

3.10 Rhema Viathianathan based on public sector AI project piloting and deployment with social service agencies in the US and New Zealand

This interview with Professor Rhema Viathianathan is a highly contextualised and more in-depth elaboration on issues related to building and using machine learning-based predictive risk models to support social service case workers who must make high-stakes decisions every day. The two use cases that provide the main context for this interview are 1) decisions about child protection interventions (whether or not to intervene with a follow-on investigation after a complaint is raised), and 2) decisions about allocation of temporary shelter and longer-term

housing for homeless people (whether or not a particular homeless person should be prioritized for the limited and always insufficient supply of homeless support housing).

She also mentions three additional use cases she and her team are working on for supporting highly sensitive, high impact case worker decision making related to: i) identifying families with new children who were not making use of available social services though their circumstances are such that they likely have a need for such services, ii) making housing allocation decisions for homeless people with mental health problems, and iii) counselling people being released from jail terms.

Based on extensive project experience across all of these types of social services use cases settings, she explains the five types of "guard-rail "guidelines that she and her co-workers adhere to so their collaborations with social service agencies result in what they consider to be an ethical and responsible approach to the development and deployment of machine learning-based predictive decision support tools. In the full interview, she elaborates on each of these five "guard-rail" guidelines:

1. Agency Ownership/Leadership
2. Fairness and Ethical Review
3. Community Voice
4. Audit and Evaluation
5. Decision support and augmentation - not final decision making and automation.

In addition, she also provides an insightful explanation of the end-to-end steps that she, her co-workers, and the collaborating social service agencies go through for developing and piloting the machine-learning based predictive risk models in these types of high-stakes social services settings. This includes four steps related to model development, testing and validation *prior to* deployment for field piloting. It also includes two major steps that occur in post-deployment field piloting. She explains how outreach efforts to engage and inform the community span this end-to-end process.

She provides a rich description of the domain setting in terms of the practical challenges front-line social workers face and how everyday realities that are part of this setting impede or limit their ability to learn how to make better decisions over time. She provides insight into why the predictive risk tools she and colleagues have been building and deploying are a stepping stone in the direction of supporting more systematic and cumulative learning for both the individual case worker and for the agency overall.

She shares her awareness that in these very complex, high stakes social services settings, the ability of these types of models to predict risk are far from perfect. At the same time, she explains the reality that while these models are neither fantastic, nor as terrible or useless as many case workers initially assumed, they have proven to be useful and beneficial when used in combination with the human case worker's overall knowledge and decision making.

An especially noteworthy point she makes is that public sector officials at all levels (local, state, federal) need to understand that end-to-end AI project efforts are typically multi-year efforts. As she points out, although AI technology and methods are obviously evolving very rapidly, a responsible approach to carefully validating and iteratively fine tuning the use of AI based models in real-world public sector settings takes multi-year extended time periods, especially with high stakes and sensitive decision making. She also calls for more emphasis on post-deployment

evaluations of real-world model usage impacts versus the common current practice of putting more of the evaluation effort on pre-deployment technical studies of the model's predictive accuracy.

In wrapping up the interview, she shares her perspectives on the role of using Generative AI with unstructured data vis-a-vis using AI-based predictive models with structured data in the social services case management and decision-making settings she has been working with. She explains why she believes that for the near term, it may be best to target the use of LLM usage toward summarising case notes and case histories using the information from both text-based (unstructured data) case notes as well as from the fields of structured data bases, and not toward making predictions for human case worker decision support.

Her view is that the (non-GenAI) technology for using machine learning methods trained on structured data provides a more reliable way of building and validating predictive risk models. At the same time, she suspect there will be many interesting ways in which the capabilities of LLMs (and their ability to work with text and speech and other forms of data) will be combined with the capabilities of predictive machine learning models trained and validated with structured data to improve the overall workflow support for social service case workers.

She concludes with suggestions to public sector decision makers overseeing AI projects within their agencies, providing her views on why she encourages public sector officials to do the following:

- Understand the reality of how noisy decision making actually is when front-line workers do not have good supporting decision support tools. In addition, understand why the use of well-designed and carefully tested decision support tools (including AI-based predictive risk models) can help to reduce this noise.
- Use AI-based decision support for helping human decision makers, not for criticizing past or current performance.
- Emphasise the functionality and use of the AI support tool using plain language that everyone can understand more so than emphasizing the use of AI for its own sake. And certainly don't just say you are "using AI" without explaining how you are using it and what it is being used for.

As part of her conclusion, she thoughtfully reflects on the reality of existing controversies and a range of opinions related to using these types of predictive risk tools in this type of public sector decision making. She explains why she has come to the conclusion that the alternative of not using any type of algorithmic support in these types of settings is also problematic, in fact, more problematic, and more likely to lead to even higher levels of bias as well as noise in everyday case worker decision making versus not using the support of these types of AI-based support tools.

4. Project methodology

The ten people interviewed for this project were already introduced in Table 1. Each person received briefing material stating that the project was sponsored by UN DESA and explaining the purpose of the project.

For six of these interviews (related to Bangladesh, Maldives and Central Bank Digital Currency, Kazakhstan, EU, Uzbekistan and Rwanda), staff from UN DESA Department of Public Institutions and Digital Government sent out the introductory email requesting their participation in the interview, and I handled all follow-on interactions. For four of these interviews (related to Singapore, the US and New Zealand/US), I sent out the introductory email requesting participation in the interview, as well as handled all follow up interactions.

For the projects in Bangladesh and Kazakhstan, UN DESA Department of Public Institutions and Digital Government staff sent out requests to several people involved in these projects to participate in the interviews. For each of these two country efforts, the one person who responded to the request was interviewed. The several other people contacted for each of these country projects did not respond.

Both the UN side and I attempted to reach out to appropriate experts in China who we respectively judged would be appropriate for the nature of this type of lessons learned interview with public sector sandboxing and field piloting of AI applications. However, we were not successful in getting someone from China to participate.

Prior to each interview, I sent each interviewee who agreed to participate another briefing sheet that provided examples of the types of questions I would ask during the interview. This set of preliminary questions served as a guide and starting point. Each interview evolved in its own unique way depending on the nature of the interactive discussion, and the experiences and interests of the person being interviewed.

Nine of the ten interviews were conducted by video conferencing. Of these nine interviews, the shortest was 74 minutes in duration. The longest was 103 minutes. The average time per interview across all nine was 87 minutes.

One interview (with Sayran Suleimenov from Kazakhstan) was conducted through email-based document exchange as he was more comfortable working with English language written documents than with English language verbal interaction.

The steps to produce the full-length interview summary write-ups for the nine interviews conducted by video conference were as follows:

1. Conduct interview via video conferencing using MS Teams.
2. Auto-generate a verbatim transcript using the MS-Teams AI application for this purpose. This auto-generated transcript include some transcription errors.
3. Using that auto-generated verbatim transcript as a base, I manually created a revised and polished write-up of the interview discussion to capture key points and other important information I deemed relevant to this project.
    - This substantial content editing and revision effort including making editorial changes to the verbatim transcript to make improvements for
        i. clarity, understandability, and logical flow,
        ii. fact-checking,
        iii. elaboration and clarification of the interviewee's intent,
        iv. elimination of redundancies, and filler words and sounds, and

v. the inclusion of supporting context or background information based on a combination of my background knowledge and external information sources accessed online through the web.

4. During this editing and revision effort, I would often replay segments of the video recording to make sure I understood the interviewee's comments and context and to review and correct as needed questionable instances of auto-generated transcription content.
5. I inserted URL links to references mentioned during the interview, and to relevant supporting background information I identified after the interview. In subsequent stages of reviewing and formatting these full-length interview write-ups appearing in Part 2, these supporting references were moved from the main text to the endnotes.
6. The first two interviews conducted were with Professor Ramaya Krishnan and Professor Gianluca Misuraca. For these first two interviews, I retained the speaker identification in the revised write-up to convey the sense of the back-and-forth dialogue between the interviewee and myself, and the nature of my probing into the interviewee's comments.
7. For the other seven interviews conducted by video conference, which occurred after these initial two interviews mentioned above, I remove the speaker identification from the interview summary write up, even though these additional seven interviews also had that same type of highly interactive back-and-forth interaction. I incorporated relevant content emerging from my questions and the back-and-forth interaction into the summary of the interviewee's comments. This was part of the content editing and revision effort.
8. For each of these interviews, after I completed this intensive editing and revision process, I manually constructed and inserted headers and sub-headers to help the reader more easily identify topics and follow the flow and content of the interview summary write-up.
9. I sent this substantially edited, revised and polished version of the interview summary write-up back to each interviewee for review to obtain comments, corrections, and suggested revisions. In this communication, I also clarified that this revised version of the interview write-up (which would incorporate any of their corrections or suggested revisions, if any) would be the version of the interview I would use for drafting my report and that the report would also include the publication of the full version of the revised write-up.
10. Any corrections or edits made by the interviewee were accepted exactly as given without further change (except for proof-reading corrections of typos, or errors with spacing or punctuation).

For the one interview conducted by email-mail based document exchange, the identical process steps listed above, from steps 3 through 10, were followed. The only difference was that that the base document I started with was the interviewee's initial written description in response to the questions I had sent him. This one interview done by email exchange of documents went through three rounds of review and revision, whereas those done by video conference went through one round of review and revision by the interviewee.

For nine of the 10 interviews completed, the interviewee acknowledged receipt of my substantially edited, revised and polished version of the interview summary write-up, and

responded back with either whatever corrections or edits they wanted to make, or with their approval to use the content in the document "as-is."

The only interview where I did not receive the interviewee's acknowledgement of receipt of the interview summary write-up or any response to the document was the interview with Esther Kunda, Director General, Innovation & Emerging Technologies, Ministry of ICT & Innovation, Rwanda. Multiple attempts were made to resend the document to the interviewee asking for acknowledgement of receipt and for any suggested changes. No response was received to these follow up contact efforts.

5. Concluding comments and recommendations

5.1 Concluding comments

5.1.1 Fast per AI technology development Versus Slow per testing, validating and deploying

AI technology and other aspects of digital technology have been rapidly progressing.[11] In parallel, an increasing number of public and private sector organisations across the world have been mobilising to experiment with using these new AI enabled capabilities and other forms of digital innovation in the context of their specific use cases, with a steadily growing, though still limited, number of these trials transitioning into ongoing operational usage.

In terms of the global public sector landscape, the ten in-depth interviews that form the foundation for Part 1 and Part 2 of this report provide a highly contextualised and more detailed view of how this happening across a small though diverse sample of countries. This provides a complementary view to the results of the 2024 E-Government Survey that provides evidence that these types of activities are taking place more broadly across many of the world's countries.

What is often much less visible to public sector officials and the broader public are all the behind-the-scenes efforts required to carefully test and validate that these new types of AI-based systems and other digital solutions for delivering government digital services are assuredly working per their intended use and in acceptable, beneficial and non-harmful ways. The ten interviews in this report provide a rich description of this less visible aspect of making the transition to using AI-based systems, and to some other forms of digital innovation, in the public sector.

The R&D and new technology aspects of AI and other digital technologies moves at a fast pace. The process of carefully testing and validating the performance of these systems in the context of real-world use cases and conditions necessarily moves at a much slower pace, as does the effort to empirically evaluate the broader domain and societal related impacts (benefits, costs and risks). This creates an inherent and ever-present tension that will not disappear. In fact, as AI capabilities continue to be enhanced, and as public sector organisations continue and further expand efforts to apply existing and new AI capabilities in their digital services, this gap between the fast "clock-speed" of the rate at which the frontier of AI capabilities keep advancing and the much slower "clock-speed" of the time required to carefully test, validate and evaluate such systems in real-world contexts will only widen.

This widening gap will likely induce increasing pressure. There will be increasing pressure on public sector units to "keep up" with the pace of AI R&D and technology development and to move ongoing and new AI projects into operational deployment at a faster pace. This will likely create pressure to short-cut or even bypass some of the steps necessary for careful testing and validation, for the needed policy experimentation, and for related broader and longer-term impact evaluation.

Public sector officials at all levels need to grasp that the effort required to do careful and reliable validation and policy experimentation through a combination of sandboxing and field piloting requires persistent and patient effort over extended time periods, with longer (even multi-year) timescales required when the AI applications are being used to support more consequential and higher impact decisions.

This contrasts with the fact that the initial steps of technically testing AI models using only historical data sets and other available information sources can sometimes proceed much more quickly because this type of testing does not involve any type of user testing or trials in the real domain context. The fact that technical benchmark comparison testing in a lab-like isolated setting can sometimes be done in much shorter time spans can be misleading as it may lead public sector officials to believe that the follow-on steps of carefully conducted sandboxing and field piloting efforts can proceed at the same speed.

Higher fidelity, more realistic sandboxing and field piloting cannot move as fast and be completed as quickly due to all the complexities involved with working with live use cases and live users, complex real-world domain requirements and constraints, and the time required to do validation, experimentation and evaluation. While requiring longer time periods, these higher fidelity means of testing, validation, experimentation and evaluation produce more reliable and useful results pertinent to safe and responsible real-world usage and subsequent operational deployment.

Public sector organisations are also under pressure to demonstrate they are experimenting with many innovation initiatives, especially those involving the use of AI. This is similar to this same type of pressure now faced by private sector organisations. For the very first steps of technical and benchmark testing, it may be possible to conduct a larger number of tests in the spirit of "letting a thousand flowers bloom" and to try many possible new AI models. However, no major public sector digital innovation initiative, and especially one making use of AI, should be allowed to proceed directly from this first phase of the initial steps of technical testing to operational deployment, bypassing sandboxing (when needed) and field piloting.

This creates a management challenge for public sector officials within any given organisational unit. They need to be highly selective about which early-stage new AI model possibilities, or other digital innovation possibilities, they allow to move forward into the subsequent, and much more time and resource intensive testing phases of sandboxing and field piloting. The practical need to do this type of project filtering may seem antithetical to the spirit "trying a lot of different things" that is part of the early stage of innovation exploration.

However, without a disciplined approach to filtering and prioritized selection that substantially limits the number of initiatives moving forward to sandboxing and field piloting, these next stage trials with users in the real domain contexts cannot be done effectively. Not only must the number of projects moving forward to sandbox testing and field pilot testing in a given time period be

carefully constrained and paced, the scope of each of those projects must also be sharply defined and reigned in from being overly broad in order to be able to execute the project as needed in the sandbox and/or field pilot setting.

In any one public sector organisational unit, within a given time frame, realistically limiting the number of new AI efforts being tested in a sandbox or field pilot trial, and carefully managing the scope of each of those efforts, will substantially improve the quality and usefulness of testing, validation, policy experimentation and evaluation results that can serve as a basis for determining whether to continue moving the project further through the testing and validation pipeline towards operational deployment. The smaller number of projects that end up making it through these more intensive, user-based trials are more likely to be refined to the point where they make it to the phase of operational deployment.

In parallel, maintaining a steady flow of projects into the first phase of technical testing, and a also maintaining a steady though filtered flow of promising project candidates into the subsequent testing, validation and policy experimentation pipeline phases used to assess as well as enhance the feasibility of innovations, will help the organisation to try more things over time in a manageable way.

A public sector organisation's capacity to manage and execute this pipeline can be built up and expanded with supporting cloud infrastructure, supporting platforms for the development and execution of AI models and supporting software applications, strong data management, and standardised processes. These capabilities will not only help with maintaining the flow through the testing, validation, and policy experimentation pipeline, but will also increase the overall capacity and speed of this pipeline.

Hence, public sector decision makers overseeing these AI and digital innovation projects need to pursue the strategy of disciplined selection and filtering to limit the number and scope of initiatives being user tested at any one period via sandboxing and field piloting. Simultaneously, over time, they also need to increase the cumulative quantity of efforts moving into sandboxing and field piloting efforts by maintaining the steadiness of project flow through the validation pipeline while also implementing the supporting efforts to increase the capacity and speed of this pipeline.

5.1.2 Clarifying the progression from technical testing to sandboxing to field piloting to operational deployment

Throughout the ten interviews, there are numerous mentions of terms related to "test", "sandbox", "pilot", "field pilot", and "deploy", though the meaning of these terms sometimes differs across the contexts of the various interviews. The fact that these terms are used so often in discussions of digital government applications and solutions, and often used to mean different things in different situations, can lead to misunderstanding.

Table 3 provides a framework for understanding these terms in the context of a systematic progression of phases used to test and validate a new AI-based application or any type of complex digital solution. It also provides a basis for better understanding the comments made above as per why user-based testing through sandboxing and field trials is relatively slow compared to the fast rate of R&D and technology development for AI and other types of digital innovations.

The four main test and validation phases highlighted in this progression are:

- Technical testing and validation (steps 1, 2, 3) done by the development team.
- Sandbox testing and validation (steps 4, 5, 6) done by the sandbox participants and evaluators.
- Field pilot testing and validation (steps 7, 8) done by field pilot participants and evaluators.
- Operational usage (steps 9, 10), which also involves its own types of ongoing testing and validation by the operations support team.

The types of people involved in each of these phases and key aspects of the various types of testing and validation done are noted in the table.

This table is meant to provide a simplified view, highlighting the progression across these four phases. As such, many details related to the typical testing steps for AI model testing and for application and system testing are purposely omitted. Also, all the complex feedback loops occurring in response to finding, fixing and iterating on addressing problems, are omitted. The real-world sequence of steps seldomly occurs in the nice clean and orderly linear progression suggested by Table 3.

Table 3: Progression of phases used to test and validate a new AI-based application or any type of complex digital solution

		Testing and Validation Progression	Testing and Validation Focus	Additional Testing and Validation Focus
No user involvement; Only technical testing with development team	1	**AI Model validation testing with internal data**	Technical evaluation of AI Model performance, e.g., model output accuracy or quality	Adherance to applicable internal or external Ethical AI/Responsible AI guidelines or requirements
	2	**AI Model validation testing with internal and external data**		
	3	**Application validation testing to achieve required functionality and workflow**	Technical aspects of application functionality, performance, security, data protection, and usability	Technical aspects of data and system integration, system level performance, system level security, data protection
Carefully pre-screened, controlled set of trial users invited to participate in sandbox	4	**Sandbox participants limited to project team and related project members (external and government)**	Same as above in sandbox setting with added emphasis on application functionality, usability and user behaviour within context of target use cases and workflows	Comparison of application usage and outputs (including AI model usage and outputs), user behaviour, sandbox user ecosystem behaviour, and sandbox usage impacts and consequences VS existing applicable regulations and laws to determine non-conformances and gaps PLUS policy experimentation to explore how to address non-conformances and gaps per existing regulations and laws
	5	**Sandbox participants expanded to include "friendly" support network of domain and use case experts acting as end user surrogates**		
	6	**Sandbox testing further expanded to include controlled sample of invited real-world end-users** *** Sample expansion step 1** *** Sample expansion step 2….**		
Available for use by restricted samples of real-world end users	7	**Field pilot testing with initial sample of real-world end users**	Same as above in field pilot context that includes unconstrained, real-world users, including untame and malicious users PLUS scalability, reliability, response time, resiliance, robustness, costs of operation PLUS re-visting of AI model technical performance with increasing field usage and broader user base	Same as above in field pilot context that includes unconstrained, real-world users, including untame and malicious users PLUS evaluation of impacts of application usage (including AI model usage) on desired outcomes as per project justification and business case, evaluation of overall benefits vs costs
	8	**Field pilot testing with expanded sample of real-world end users** *** Pilot sample expansion step 1** *** Pilot sample expansion step 2…**		
Available for use by full population of potential real-world end users	9	**Initial operational usage, first 12 months** *** Operational upgrade or revsion step 1** *** Operational upgrade or revsion step 2…**	Same as above PLUS Managability of system and user behaviour with increasing scale, increasing diversity of users, and increasing variety of user interactions	Same as above PLUS Continued evaluation of impacts of system and AI model usage on desired outcomes with increasingly large and diverse user base, and continued evaluation of overall benefits vs costs
	10	**Ongoing operational usage beyond initial 12 months** *** Operational upgrade or revsion step 1** *** Operational upgrade or revsion step 2…**		

It is important to understand that testing of various types take place across all four of these phases. An important distinction made in several of the interviews is the difference between the technical testing phase done by the development team (steps 1, 2, 3) versus the other aspects of

testing that occur in the subsequent sandboxing or field piloting phases where different types of users get involved. It is useful to refer to this very first phase of testing as “technical testing” (which is often done without user involvement). When one talks about “testing”, or “the test phase”, this could be occurring within any of these four phases, so it is best to clarify which phase of this progression we are referring to when we talk about “testing.”

In a broad sense, both sandboxing and field piloting are forms of “piloting”, so it is easy to misunderstand what someone means when they say, “we are piloting.” This leads to the question of, “How is sandboxing different from field piloting?”, especially as sandboxing may sometimes proceed to step 6 that includes the inclusion of real-world end users, and both sandboxing and field piloting are forms of “trials.” The distinguishing features of sandboxing (vs field piloting) emphasised in several of the interviews are as follows:

- The sandbox is a protected and controlled setting. Related to this, the people and companies invited to participate in the sandbox setting as either product/service providers, surrogates for end users, or actual real-world end users, are carefully pre-screened and selected for suitability for the sandbox effort. Users are supposed to be protected against experiencing real harm or losses by participating in the sandbox trials.
- The relevant government authorities overseeing the sandbox effort often relax or suspend certain types of regulatory requirements within the controlled confines of the sandbox setting to allow for the trialling and testing of products, services and scenarios that would not be permissible under existing regulations.
- The sandbox is often used to perform experiments with government regulations and policies in conjunction to performing experiments with new types of products, services and scenarios. This may involve regulatory exploration and discovery as described in the interview with Gordon Clarke.
- A formal sandbox effort follows well defined processes for conceptualisation, operation, evaluation and exit as explained in the previously mentioned UN DESA Policy Brief on “Sandboxing and Experimenting with Digital Technologies for Sustainable Development”. This is especially the case for sandbox efforts with financial service innovation overseen by either a country’s central bank or financial services regulator.

As pointed out in the interview with Gordon Clarke, Ramayya Krishnan and Gianluca Misuraca, the controlled sandbox setting is conducive for doing regulatory experimentation and discovery and related policy experimentation because all participants in the sandbox effort can be notified and trained in a controlled way about changes to the “rules of the game.” This allows for relatively rapid rounds of iterations that provide the feedback needed to co-evolve the design of the new digital innovation solution (including an AI model) and the regulations and policies used for governance. This type of experimentation and co-design becomes much more difficult or even impractical to do in much less restricted or unrestricted field pilot setting.

In contrast, field pilots involve real-world end users, including potentially un-tame, un-friendly and even potentially malicious users who are not pre-screened and selected for participation in the same controlled way that is done for a sandbox. The sample size of the end users participating in a field pilot is restricted initially, though that sample may be progressively expanded as the field pilot effort proceeds (e.g., adding users from additional government departments, or residents from additional geographic locations). However, the setting for the field pilot is the actual real-

world setting, “in the wild” so to speak, without constraints as to what may happen or who within the allowable sample may participate.

The advantage of the field pilot is that it is possible to observe real-world end user behaviour and impacts under real-world conditions. It is much harder or impractical to do real-time experiments with making regulatory changes within a given field pilot setting, though it may be possible to do comparisons across field pilot settings where there may be different regulatory policies being applied within each separate setting by intentional experimental design or by natural occurrence.

We see from the interviews that there can be multiple pathways through the four phases and their respective steps shown in Table 3. In the technical testing phase, a more careful and robust model development effort will involve both steps 1 (with internal data only) as well as step 2 (additional validation with a broader range of external data) for technical testing. Not all AI model technical testing project efforts include step 2, even if they should.

Some sandbox efforts, even if they follow the formality of the four stages previously discussed (conceptualisation, operationalisation, evaluation, exit), may only take the sandboxing through steps 4 and 5, and either eliminate step 6 or only do it to a very limited degree, and save the larger scale testing with real-world end users for the field pilot phase.

Some projects may proceed from the technical testing phase (steps 1, 2 and 3) directly to field piloting (steps 7, 8) without having a formally declared sandbox phase. Even if this is the case, they may end up doing the same type of controlled testing as occurs in steps 4, 5, and possibly even step 6 of sandboxing, as a pre-cursor to doing their field pilot testing. In other words, they are informally doing sandboxing without formally referring to it as such. Some of the project examples in the interview with the Uzbekistan team seemed to proceed in this fashion. The Rwanda example of testing the chatbot for the community health workers combined aspects of sandboxing and field piloting in parallel. Dominic Chan from Singapore noted that not every digital services application making use of AI requires changes to existing regulations, and as such, regulatory sandboxing and efforts for regulatory and policy experimentation are not always needed for every new digital services project, even if they make use of AI (depending on how they make use of AI).

Durations for field pilots can vary over a wide range of time scales. Some may only run for weeks or months. For other situations that involve much higher degrees of risk and where the observation time window needed to observe effects and impacts is longer, the duration for field piloting may need to extend over a multi-year time period as was the case for the examples given in the interview with Rhema Viathianathan.

The term “deployment” also needs to be clarified in the context of the framework shown in Table 3. People in some public sector settings, and in some private sector settings as well, reserve the use of the term “deployment” for all the implementation efforts that occur during Steps 9 and 10 for transitioning to operational usage (sometimes referred to as “production” usage). In other settings, people commonly refer to the deployment effort required to proceed with a field pilot, and even the deployment effort required to proceed with a sandbox effort.

Preparing for a sandbox effort or a field pilot effort does indeed require a supporting deployment effort. However, it is a deployment effort of lesser degree and scope than what is later required for the final phase of deployment necessary to support the full scale and scope of operational usage. As smaller though still complicated deployment efforts occur for sandboxing and field piloting, and much larger deployment efforts occur for the transition to operational usage,

when one talks about “deployment”, it is best to clarify which type of deployment one is referring to, as not everyone will assume the term only refers to the final stage of deployment for operational usage.

Concluding points related to this discussion of the pipeline of the four testing and validation phases shown in Table 3, and incorporating other related inputs from the ten in-depth interviews, are as follows:

- Public sector officials involved in reviewing and overseeing AI efforts and other digital technology innovation efforts need to understand the nature and importance of all four phases of testing and validation, and the special importance of sandboxing and field piloting for policy experimentation.
- In particular, they must understand that even if the results of technical benchmark testing done in the very first step of the first phase of technical testing is promising (which it often is), it is a slow, complex, and effort intensive road ahead to appropriately do the needed sandbox testing and field pilot testing, especially for new AI models and other digital applications that involve more consequential types of risks and impacts to users.
- They must learn to cope with the dissonance and pressures associated with the gap between the fast clock speed of AI and digital related R&D, technology and commercial product development versus the slower clock speed of user-based and real-world contextualised testing, validation, experimentation and evaluation efforts.
- Project budget and timetables must be aligned with the reality of the need for substantial amount of follow-on user-based testing, validation, experimentation and evaluation that need to be done during the sandbox phase and field pilot phase before the new (increasingly AI-enabled) digital services application can be responsibly transitioned to operational usage. And even after transitioning to operational usage, the effort for ongoing testing, validation and evaluation needs to continue.
- While sandboxing and field piloting are both forms of “piloting”, there are important defining characteristics which distinguish the sandboxing phase from the field piloting phase.
- While not all AI-based projects or other digital innovation projects for new government digital services require formal sandboxing, some do, and it is important for public sector officials involved with these efforts to know when formal and more comprehensive approach to sandboxing is warranted.
- Even when formal sandboxing is not required, some of the steps of sandboxing still need to be conducted as a precursor to, or as the initial part of field pilot testing.
- The productivity, quality, and capacity of moving a portfolio of projects through this pipeline of technical testing, sandboxing and field piloting and into operational usage can be increased through gradually building up supporting cloud infrastructure, supporting platforms for the development and execution of AI models and related supporting software applications, strong data management, and standardised processes.
- However, this type of capability and capacity build up is itself a large and multi-year digital innovation effort which also must go through these same testing and validation phases over extended time periods.
- A public sector organisation must be able to realistically assess its internal ability at any given point in time to accomplish these four phases of testing and validation through a

combination of using internal staff and through procuring external vendor and consultant services. The public sector organisation may even need external help to do this type of assessment of their internal ability to the necessary testing, validation and experimentation and how to get the work done given the capability gaps.

- As was demonstrated during the Covid-19 pandemic and in other special circumstances, these timetables for sandboxing and field piloting can sometimes be substantially compressed, though such exceptional circumstances require very serious risk-benefit considerations and a willingness or necessity to tolerate much higher degrees of uncertainty and associated risks.
- As discussed in several of the interviews, a piloting effort is so much more than just getting the technical aspects of an AI model to work as it is necessary go beyond that and observe, test, validate and evaluate the larger socio-technical system involved. Also, as pointed out in one of the interviews, the challenges for a government to transition to cloud and digital, as well as to AI-enabled applications, are usually underestimated. The discussion above provides more context for understanding "the wisdom" of these observations from several of the interviews.
- The interviews with Professors Ramayya Krishnan, Gianlucca Misuraca and Rhema Viathianathan and with Rwanda Ministry of ICT and Innovation official Esther Kunda highlight ways in which public sector organisations have successfully partnered with universities to support public sector AI efforts. Examples are given where the university partnership was important for either developing AI models for specific types of decision support, testing AI solutions and evaluating them in terms of technical performance as well as in terms of broader impacts, or for developing human capital and manpower for the technical side of creating AI models and solutions as well as for the managerial and operational side of organisational oversight and usage.

5.1.3 Where does policy experimentation fit it?

Sandboxing and field piloting are the key mechanisms for different levels of government to do both technical and policy experimentation related to the use of new types of AI and other digital applications. Through sandboxing and field piloting efforts, experimentation, especially policy experimentation, becomes more practical, reliable and scalable.

A key point to remember is that a sandbox is not primarily for technical testing, though it includes a continuation of technical testing. Gianluca Misuraca reminds us that sandboxing is an integral part of how the public sector can experiment in a controlled way to learn how to better integrate across policy making, service delivery and regulation. Public sector sandboxing efforts are not just how to get some new technical thing or method to work. It's testing how to go about it so that the policy, the law, the regulatory aspects and the service delivery can be brought together in such ways that all these facets can work together.

As such, policy experimentation is an integral part of sandboxing and provides the basis for how to continue with field piloting.

5.2 Recommendations

5.2.1 Recommendations for UN Development Account projects involving sandboxing and piloting with AI and other digital technology applications

The Development Account is a capacity development programme of the United Nations Secretariat aiming at enhancing capacities of developing countries in the priority areas of the United Nations Development Agenda.[12] Future Development Account projects which involve efforts to build digital technology applications, including AI enabled applications, as a means of demonstrating the ability to deliver new types of services for supporting sustainable development goals must carefully consider how to realistically manage the scope of project deliverables.

The digital technology applications, which are increasingly AI-enabled, need to go through the various testing and validation steps discussed above (Table 3): from the first phase of technical testing to the second and/or third phases of sandboxing and field piloting as well. This is an intensive and complex effort, not only in terms of testing and validating the technical and functional aspects of the digital (and let's assume AI-enabled) solution, but also for doing the necessary policy experimentation, and for evaluating if the use of the technology solution in conjunction with any new policies are providing the desired economic or social development outcomes.

Project deliverables associated with the technology demonstration should have a carefully constrained and focused scope that can realistically be executable within the limits of the project timetable and budget. The initiation of project work efforts specifically related to creating and using the technology deliverable should not be deferred to the very later parts of the project or else it will not be possible to do the required work within the time schedule.

Within a Development Account project involving technology solution creation, demonstration and evaluation, it may be useful, even necessary, for other parts of the project to identify and address a broad scope of issues without the need for a technology demonstration. This can be done through the regular mechanisms of desk research, stakeholder engagement, workshops, and surveys, with the summaries and results reported as typically done in presentations and documents.

The key issue to manage is the sequencing dependencies between those parts of the project that require building, implementing and testing a technology demonstration and doing the related impact evaluation related to the technology solution, and those parts that involve broad stakeholder engagement and alignment efforts and supporting information-gathering, analysis, and summary report generation across multiple issues and stakeholder segments. As illustrated with the Bangladesh project examples, if the same people serving as project team leads are responsible for both the technology-based depth deliverables, and the information-summary breadth deliverables, these overlapping team lead roles are likely to cause a problem with meeting simultaneous delivery requirements on both the breadth and the depth fronts.

It may be better to have these two types of delivery efforts run in parallel with different team leads. Perhaps some staggering in start times is appropriate so enough information gathering can be done to make sure the technology demonstration related effort has sufficient clarity needed to define performance objectives and requirements needed to proceed with design efforts. The situation to be avoided is having the majority (e.g., 50 % to 60% or more) of the total project timetable of a relatively short-term (1, 2 or 3 year) effort spent on the broad scope, information

gathering and analysis across a wide range of issues, and deferring the start of sizable and complex technology demonstration and related sandbox effort, and possibly even a field pilot effort, till the very back end of the total project timetable.

All of this comes down to adhering to the project management basics of realistically aligning project scope, deliverables, timetables, staffing and budget. This is well known and seemingly obvious need. Yet sometimes the enthusiasm of wanting to "do more" to address a wider range of complex issues, or the pragmatic political need to include some type of consideration of a broad range of social and development issues along with efforts to also do a substantial trial involving a complex technology-based (and likely AI-enabled) demonstration into the same Development Account project can easily lead to a misalignment in these project management basics.

5.2.2 Recommendation for the UN DESA E-Government Survey related to sandboxing and piloting with AI and other digital technology applications

The addendum on Artificial Intelligence included in the 2024 edition of the UN E-Government Survey conveys that the UN DESA team involved in designing and implementing the survey and publishing the results has a realistic, sophisticated and up-to-date overview-level understanding of issues related to the use of AI in the public sector across the population of UN member states.

The 2024 E-Government Survey already included the following types of survey questions related to AI:

- Does the country have legislation or regulations on the use of new/emerging technologies such as AI, robotics, blockchain, 5G, and the Internet of Things?
- Has the country adopted legislation or regulations on the ethical/responsible use of AI in public administration?
- Does the national e-government strategy make specific reference to the use of artificial intelligence (AI)?

Given this starting point, the recommendation based on this project effort is to consider the addition of some or all of the following types of questions for each member state to respond to:

At the national level:

- Does your country's public sector have defined processes for how to pilot new e-government services through whatever combination of sandboxing or field piloting?
- Do these processes include frameworks, approaches or measures for testing and validating the use of AI as part of the e-government service?
- Do these processes provide guidelines or tools that employees involved in AI-enabled projects can use to make it easier and more practical for them to assess risks as well as benefits associated with using AI?
- Are there defined processes or defined cross-government coordination approaches for evaluating the broader social and/or economic impacts of using new or improved e-government services, including those that are AI enabled?

Adding questions of this type to future editions of the E-Government Survey (however edited or condensed or possibly incorporated into existing survey questions) may motivate member state countries participating in the survey to move in the direction of the practices referred to in these suggested questions. To the extent that scoring well on responses to these types of questions

enables a country to incrementally boost its EGDI score, this would motivate countries to put more emphasis on the quality of how they implement AI-enabled government digital services versus only the quantity of how many they implement.

It would also signal to UN member states participating in the E-Government Survey the importance of making it more understandable and practical for public sector staff members to do the everyday, complex work of executing and supporting the piloting, evaluation and deployment of new digital services. Working level public sector staff and their upper management need usable, practical and understandable tools to determine the adequacy of AI testing, to do AI related risk assessment, and to do broader impact and outcomes assessment of the digital services being deployed.

An emphasis on making these types of assessment and evaluation tasks more practical and easier to do is especially important given the increasing number of AI governance related policy documents being published by international organisations and new AI related laws and regulations being enacted by national governments. If public sector staff doing the everyday work related to creating, testing, validating and evaluating these systems cannot translates these polices, laws and regulations into easy-to-use everyday practice, these higher-level intentions will stay at a high level, removed from realities of the everyday work efforts of specific projects.

Future E-government survey reports could also consider case studies or vignettes that highlight how some countries are going about doing the things referred to in these suggested questions, even if these types of suggested questions are not added to the E-Government Survey. Such case study descriptions would provide useful examples to the broader community of UN member states participating in the E-Government Survey.

5.2.3 Recommendations for public sector institutions implementing government digital services that use AI

My one recommendation is a request to the public sector officials and staff looking at this Part 1 report document to read at least a few of the full-length interview summaries compiled in Part 2 of this report. Which ones would be most relevant for you to look at? You can make this determination by looking at the summary of who is interviewed (Table 1), the topics covered in each interview (Table 2), and the brief narrative summary of each interview (Section 3).

These full-length interview summaries in Part 2 are the “gems” of this project. They capture a rich array of descriptions, lessons learned, and insights from years of accumulated public sector project experience related to sandboxing, piloting and policy experimentation with AI and other digital initiatives. With apologies to today’s remarkable though still limited and sometimes flawed AI-based Large Language Models, no simply distilled, condensed summary of the full-length write ups, or briefly stated recommendation derived from them, can entirely substitute for the richer experience of actually reading some of these interviews. You can be assured that you and your colleagues will find something in one or more of these interviews that will be helpful to you.

Endnotes

[1] https://www.un.org/development/desa/dpad/wp-content/uploads/sites/45/publication/PB_123.pdf
[2] https://publicadministration.un.org/egovkb/en-us/Reports/UN-E-Government-Survey-2022
[3] https://publicadministration.un.org/egovkb/en-us/Reports/UN-E-Government-Survey-2024
[4] See National Academies of Sciences, Engineering, and Medicine. 2024. *Artificial Intelligence and the Future of Work*. Washington, DC: The National Academies Press. https://doi.org/10.17226/27644 for an excellent comprehensive overview of recent AI technical progress, the mechanisms through which AI can impact an economy's productivity and labour market structure, ways in which AI is impacting skills and labour requirements, and risk related to increased usage of AI.
[5] https://publicadministration.un.org/en/Capacity-Building/Projects/Frontier-Technology-Policy-Experimentation-and-Regulatory-Sandboxes
[6] https://www.un.org/development/desa/dpad/publication/un-desa-policy-brief-123-sandboxing-and-experimenting-digital-technologies-for-sustainable-development/
[7] See "Creating the Capacity for Digital Government," Asian Management Insights magazine, March 2023 (Vol. 10, Issue 1). This article was based on interview content from the following four interview sessions: Singapore's Digital Government Story **Episode 01**, Singapore's Digital Government Story **Episode 02**, Singapore's Digital Government Story **Episode 03**, and Singapore's Digital Government Story **Episode 04**.
[8] On 24 September 2024, the US National Institute of Standards and Technology (NIST) announced they awarded $6 million to Carnegie Mellon University to establish the CMU/NIST AI Measurement Science & Engineering Cooperative Research Center (AIMSEC) https://www.nist.gov/news-events/news/2024/09/nist-awards-6-million-carnegie-mellon-university-establish-ai-cooperative. CMU's Professor Ramayya Krishnan serves as the lead research coordinator for this joint CMU/NIST research effort https://www.cmu.edu/news/stories/archives/2024/september/nist-awards-6m-to-carnegie-mellon-university-to-establish-ai-cooperative-research-center.
[9] For an in-depth example of product management within GovTech, with background on how they built up their internal capability to prototype and deploy digital services and products, see both Part A and Part B of the teaching case, "Digital Product Management Under Extreme Uncertainty: The Singapore TraceTogether Story for Covid-19 Contact Tracing," February 2022.

[10] See "Singapore's AI Applications in the Public Sector: Six Examples," Management and Business Review journal, special issue on AI for Customer Engagement, Spring/Winter 2023 (Vol 3, Issues 1&2, July 2023), for an overview of six Singapore government digital services that were making use of AI as of the end of year 2022. More current information on GovTech products and digital services are given at https://www.tech.gov.sg/products-and-services/overview/.

[11] Two excellent and highly reliable sources of information that provide annual updates on AI R&D, technology, application and impact trends are 1) The AI Index published around March of each year (since 2018) by Stanford University's Human-Centered AI Institute (HAI), https://hai.stanford.edu/research/ai-index-report and 2) The State of AI Report (in the form of a powerpoint presentation) released around October of each year (since 2018) by Air Street Capital, https://press.airstreet.com/p/state-of-ai-report-2024.
[12] See the UN Development Account website at https://da.desa.un.org/.