United Nations | Department of Economic and Social Affairs

Lessons Learned from Sandboxing, Piloting and Policy Experimentation with AI and Other Digital Initiatives: Part 2, Ten In-Depth Interviews

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 2, Ten In-Depth Interviews

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, Junho Lee, Arpine Korekyan, Deniz Susar, Aiai Li and Young 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 moves 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 they 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.

The same Executive Summary used for Part 1 of this report is re-used as the Executive Summary for Part 2.

The Ten People Interviewed

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** **Interview date: May 05, 2024** 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** **Interview date: September 30, 2024** 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:** **Interview dates (via written correspondence): May 03, August 06, November 17 & 18, 2024** 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** **Interview date: May 06, 2024** 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			
	5.	United States,	**Ramayya Krishnan**

Big Picture Overviews of Policy Experimentation and Sandboxing		with some other global examples	**Interview date: April 24, 2024** 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** **Interview date: May 02, 2024** 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** **Interview date for team: May 07, 2024** 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** **Interview date: May 31, 2024** Director General, Innovation & Emerging Technologies, Ministry of ICT & Innovation.
	9.	Singapore	**Dominic Chan** **Interview date: May 10, 2024** 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	10.	US and New Zealand	**Rhema Vaithianathan** **Interview date: September 27, 2024** Professor, Auckland University of Technology; Director, AUT Centre for Social Data Analytics; Consultant to numerous social service agency efforts to use AI-based

Social Services Applications			predictive risk models to provide decision support to social service case worker in areas related to child protection and housing assistance for homeless people.

See Part 1 of this report for a description of

- Project origins and motivations.
- The rationale for presenting the ten in-depth interviews in this order.
- Project methodology regarding the interview protocol and the editing, revision and elaboration approach for creating the full-length interview write-ups.

INTERVIEW 1: Bijon Islam, CEO of LightCastle Partners Consulting, Bangladesh

Date of Interview: May 05, 2024

1. Introduction to Bijon Islam

Bijon Islam co-founded the management consulting company LightCastle Partners and currently serves as the company's CEO.[1]

The consulting firm does market strategy work that helps international corporations come into Bangladesh. They also provide advisory services to help external investors make investments into the Bangladesh ecosystem. As part of the investment advisory services work, they often consult with different development partners, providing them with technical assistance, including World Bank Group, UN agencies, different bilateral partners like the Switzerland Embassy, the Dutch Embassy, and the Japan International Cooperation Agency (JICA). They also work closely with the Bangladesh government's Annual Development Programme (ADP).

2. Background on Bijon's involvement with the Bangladesh project

This project you are interviewing me about is a joint effort with UN DESA and Aspire to Innovate (A2I) which is the innovation arm of the Bangladesh government under the Ministry of ICT.[2] These two organisations bought us in as a technical expert to help set up a project with the goal of developing a solution that would help increase access to capital for micro and small enterprises.[3]

This sandboxing effort started somewhere in 2022, about two years ago and that is when LightCastle Partners started its involvement, in the early part of the sandboxing and pilot effort. UN DESA and A2I started working together on prior planning stages that led to this sandbox and pilot effort even earlier, starting sometime in 2020.

Our role of providing technical assistance consisted of the following:

- First, bring together the different types of relevant stakeholders. To develop a product that improves access to capital for micro and small enterprises, we had to bring in micro and small enterprises, trade associations, financial institutions of different kinds including both banks and non-bank fintech providers, the appropriate national regulatory authorities, and the central bank. We brought all these different stakeholders together for a series of workshops to jointly define the problem, the solution approach, and to suggest how to plan for pilot effort including plans for the technology to support the pilot as well as plans for all the non-technical aspects as well.
- Second, synthesise those findings, develop summaries, and validate our synthesis and summaries with the multiple stakeholders.

- Third, select a set of partners to be involved in piloting the solution.
- Forth, manage the execution of the pilot. This includes all related effort to create and deploy the enabling tech stack/technical infrastructure as well as the administrative and work process required for pilot execution. Currently, we are still in the middle of executing the pilot.
- Fifth, document the learning from the pilot and review these learnings with the key stakeholders and participants.

My role was to work closely with the UN DESA project lead and the Bangladesh A2I project lead to oversee and support this entire end-to-end process and to guide and supervise my employees at LightCastle Partners who did most of the work. They ran the workshops, synthesised the findings, did the validation workshops, the follow up on selecting the partners and liaising with those partners, and set up the tech stack.

We also planned and executed international visits where we took Bangladesh regulators and private partners to see how this same type of effort is done in a few other countries, including Singapore. We synthesized the learnings from these international visits and used that as in input when we set up the pilot (both technical and non-technical aspects).

3. The meaning of policy experimentation and sandboxing in the context of this project

Over the last 20 years or so, there has been a significant uptick of digital technologies throughout the world. For example, in Bangladesh, you have a population of about 165 million people and out of that, you have about 120 million people who have mobile phones, of which 50 million of them have smartphones. Because of this, we have people who are leapfrogging banking. That means they don't have a bank account, but they have a mobile financial service or digital finance account. This leapfrogging makes it possible to build financial services products beyond just basic financial services that include a wider range of services and products that involve financing and exchange of payment, that can be scaled at a rate that was not possible before.

Given people can use their mobile devices to do transactions, when they do this with a financial service and any related service that involves payment, there are risks because someone - a bad actor - can launch a campaign to get a small amount of money from many people through a scam. Such scamming efforts are very popular, especially in Asia and maybe in Africa as well. Therefore, when you come up with a new financial solution of any type, instead of immediately opening it up to the entire mass market, it is necessary to do a small pilot and that's where the sandbox idea comes in. An initial pilot can happen with a small, select group of customers, even as small as around 100 customers, to see if it 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. This is where the sandbox comes in.

Of course, there has to be policy that allows you to do this sandbox, and that is where the digital policy experimentation comes in, the policy that allows people to create and execute this sandbox and that defines the needed guidelines, rules or even regulations for it. Everyone involved in the sandbox effort needs to know at the outset: "OK, if you want to start this sandbox, these are the five (or however many) governance and policy things you are going to need." These guidelines and guardrails must come from somewhere which is why you need policy clarification, and perhaps even policy experimentation, to accompany a sandbox effort. Also, there needs to be clarification at the outset regarding: this is how you apply to participate in the sandbox, this is how the sandbox will be set up and how it will operate, this is how experimentation will take place and how evaluation will take place.

Related to this, it is necessary up front to define the vision of what success would mean within the sandbox effort and how that would be measured and evaluated so participants know at the outset what a successful sandbox would look like. Similarly, how it would be determined if the sandbox is later assessed to be not so successful. These are key aspects of the whole policy framework. This was my understanding of what we needed to do on the policy side of this pilot programme.

A sandbox is something that's more than a concept. You are testing something out. There is an underlying concept that is there for what to do and the sandbox allows you to test out and evaluate the product or service that you have built. Most of the time, there is the tech stack behind the sandbox that is also being tested, but a sandbox doesn't necessarily have to have a tech stack. Whatever it is that you have built, you're testing it out in a closed and safe environment. It's almost like a lab experiment though this experiment may also involve a controlled number of external users. Within the sandbox setting, you are building out a solution for a service offering, testing it in a lab-like setting (which may involve real-world pilot users) within a very controlled environment, seeing the hiccups, and seeing how this is all working and what is not working.

4. Lessons learned from this sandboxing and pilot effort

4.1 Lesson #1: Don't make the scope of the pilot too wide.

We started with too wide of a scope. We started the planning workshops with a total of four streams of work: one stream on access to capital, one stream on access to markets, one stream on access to services, and one stream on access to skills. We thought it was necessary to address all four of these streams in parallel from the outset because helping the micro and small business segments of the economy in Bangladesh requires helping them with access to all four of these aspects.

They need capital. They need market support. They need access to government private sector services. And they need access to skills and to ways of supporting skill development.

However, there is a big difference between only discussing or planning at the concept level and trying to achieve tangible real-world results through piloting within a sandbox setting.

Given the realities of what it takes to get partners involved, create the sandbox setting, start the sandbox, and do the piloting, I can see in retrospect that we started with too wide a scope.

It was good that we had the early workshops where we discussed all four of these streams (per access to capital, access, services and skills) with the relevant participants and stakeholders. There concept level workshop sessions were useful for multi-stakeholder awareness raising, planning and alignment. However, after the workshops, we should have been more disciplined and we should have realized, “We can't tackle everything. Let's focus on just one or two of these key areas.”

When we were just starting, we underestimated the time and effort to plan, organise and execute all the workshops for each of the four areas. For example, just for the one area of access to capital, it took us nine to 12 months because we had to go through two large workshops, 5 medium workshops, and 3 small workshops, and we had to work around the schedules of the government people involved, and they were juggling a lot of other priorities. Similarly for the other three areas.

If we could do it differently, I would still start with the breadth at the beginning and consider the issues related to all four areas. However, after that initial phase, we should have been more focused. My recommendation is that when you start sandboxing and piloting, don't start too wide. Narrow it down and do something more concrete.

Of the four streams of access we considered, given that they were all important and eventually necessary, my lesson learned is that we should have started the sandbox and pilot with just the one stream of access to capital. Technology can make a large impact quickly with access to capital because such a large fraction of our population has mobile phones, this type of access can be scaled quickly, and private sector partners want to participate in this type of effort.

4.2 Lesson #2: Quickly find partners that can move quickly to implement the pilot.

It took for long for us to make the final choices of partners to work with for the pilot efforts.

As part of our initial planning across all four streams, we talked with about 20 banks and about 10 fintech players as well as with regulators. Early on we already had a feel for the partners who can be champions, and we could see that the private sector entities could move more quickly than the government banks. Of course, the government banks still had to be involved and informed as they would need to be involved later on for subsequent scale up efforts.

It took us too long to get to the point where we finalized the choice of a private sector bank and a private fintech as a pilot partner for the access to capital pilot. I wish we would have gotten to that point sooner, and started the pilot sooner with private sector players as opposed to waiting as long as we did to get to that point. My recommendation is to select partners for the pilot effort early on.

A key factor why our partner selection took too long was because our scope was too wide, as per the first lesson learned. We were dealing with four streams, and initially engaging

with 15 or more partners from each of these streams. That is a lot of up front effort to engage with and evaluate external partners before it was possible to narrow down and finalize the partner choice for each one of the streams. And it turned out that we were not able to proceed with all four streams anyway. This breadth of scope slowed things down and made it more difficult and time consuming to finalize the partner selection.

4.3 Lesson #3: Don't underestimate the technological complexity needed to implement the sandbox and pilot.

We underestimated the technological complexity of the effort required to execute our pilots. We thought it would be easier (and hence faster) than it turned out to be to create and deploy the needed tech stack.

A key requirement for our tech solution was to associate a unique identification number to each of the micro and small business entities that we would be working with in order to give them better access to capital and the other areas mentioned above. We thought we could do this by using the mobile phone number as the unique ID. That turned out to be an incorrect assumption. There are too many instances where the actual user of the mobile phone number is different than the person who holds the official registration, where a person or entity has multiple phone numbers registered to their name, and where a small enterprise uses multiple mobile phones but they are not all registered in the name of this business entity. We had examples where the mobile phone user was a female and the mobile phone was registered using the husband's name. We had issues with matching mobile phone numbers for a micro or small business to their trade licence and to their certificate of incorporation. The trade licence number changes periodically and that further complicated things.

When we started to build our tech solution assuming we had a unique way to identify the people and entities we were dealing with, we did not anticipate the level of complexity we would be facing. We thought it would be more seamless to have a SME at one end send information to the bank or to a Fintech platform and have the necessary identify verification done. It was not seamless. There were many complexities, including challenges with location verification due to issues we had with the reliability and consistency of geolocation.

Another aspect of underestimating the complexity of the necessary tech stack is that we assumed we would be able to address whatever challenges we encountered through software functionality and modifications. That turned out not to be the case. We encountered many issues that were physical and therefore hardware-based in nature, and within the scope of this pilot, we did not have the necessary physical digital infrastructure we needed to solve these issues easily.

My lesson learned is to never underestimate the tech stack requirements and related to this, to have more respect for the complexity of the whole process.

We did look at national identify solutions and related credit access solutions from several different countries, specifically South Korea, India and Singapore. We even took some of our partners to Singapore, including representatives from our central bank and from some of our private sector partners, to see their technology solutions and processes for using digital

solutions and related processes to improve access to credit. We were really inspired when we saw what they have done. However, emulating that functionality in Bangladesh has not been easy. Singapore has done this in a very planned way and systematically built up the necessary infrastructure over time. We are trying to emulate them now but it is a lot more complex than I thought.

4.4 Lesson #4: Don't "over-workshop" with large workshops to the extent of causing long delays in starting actual piloting.

The way our project budget was set up, we were required to do many larger, multi-participant workshops. Workshops are important and you obviously have to do them, especially early on, to get alignment and validation.

Knowing what I know now, If I could have changed the design of our engagement approach, I would have reduced the number of multi-partner/stakeholder workshops and increased the number of more focused in-depth interviews with a select number of individual partners and stakeholder because these more in-depth discussions is what really helps the project to progress. Even after interacting with a key partner or stakeholder through several larger workshops, real commitments to participate in the pilot do not really happen unless there is more focused one-on-one discussion to discuss details for mutually acceptable terms of participation.

More general one-to-many interactions through workshops are needed through the sandbox lifecycle, but there must also be the more focused one-on-one interactions both in parallel and as follow up to the workshops to get the pilot moving.

If I could do it again, I would lessen the number of larger size multi-participant workshops and even consider having some workshops with a smaller number of participants. Having 60 people in a workshop is often not so effective, even if you break them out into four groups of 15 participants. Our experience has been that a workshop with smaller numbers of people - say 10 or 15 in total - can get more accomplished. It would help to be more strategically focused with each workshop. We also need more one-on-one interaction. I think this more focused approach would have been very helpful to the overall effort.

Different modes of engagement are needed across the various parts of the sandbox lifecycle: larger-size workshops, medium-size workshops, smaller-size workshops, and more in-depth discussions with just one or several people. It would help if the project budget would have allowed us more flexibility as per which mode of engagement we could use at any given time in the sandbox lifecycle process to meet the needs of the project.

I would also start more in-depth discussions with potential partners earlier on through smaller, more focused workshops and one-on-one discussions. These potential partners need convincing to participate. After all, this pilot project is not paying them to participate. However, we are giving them the opportunity to participate which in turn enables them to gain useful experience through being part of the pilot. But that opportunity is latent, and initially, the partner is not fully convinced of its value. That is why more focused interaction is needed to help them understand the value of participating in this pilot.

4.5 Lesson #5: Maximize the synergy with the few key government entities most relevant to the narrowed pilot scope as early as possible.

We spent a lot of time engaging with a wide range of government entities because we initially started with a plan for pursuing four different streams of improved access.

For example, because we initially planned to pilot a new approach for improving access to markets and to relevant business and government services for micro and small enterprises, we spent a lot of time engaging with the Ministry of Commerce and the Ministry of Industries. Then we ended up eventually narrowing our focus to only the one area of access to capital. Now we see it would have been helpful if we had pursued more in-depth engagements earlier on with several additional key government entities related to capital access and fintech beyond the few we had been engaging with. For example, the Bangladesh Sovereign Wealth Fund has strong links with FinTech companies through its shareholdings. The stock exchanges in the country have strong linkages with the banks, and also have strong interest in promoting financing for the micro and small business segments of the economy.

Early on, because of our breadth of scope, we did not engage with these additional government entities who were strategic to access to capital. Now that we are focused on this one area, we have a better appreciation of how important these additional government entities are to our efforts and how useful it would have been to have involved them earlier in our pilot efforts.

4.6 Lesson #6: Do in-depth knowledge exchange with external experts early on.

- Read about the efforts in other countries that have already achieved what you are trying to do through your pilot.
- Go beyond a superficial understanding of how another country went about their effort. See if more in-depth reports or write ups are available and study them.

Have in-depth conversations with a few key participants and experts from that country to understand the complexities, to learn how they started and evolved, how their local context influenced their approach, and to get their recommendations for how to begin given our own local context.

Have these in-depth conversations with people with relevant experience as early as possible in your sandbox and pilot effort.

Those in-depth conversations should include people from private sector entities who were involved as well as people from the government.

Private sector entities with relevant experience can sometimes tell you very forthrightly why certain approaches are more likely to work or not work given policy/regulatory, operational, technological and user related issues.

Early in our effort, had we done more in-depth experience sharing with players in other countries who had already implemented what we are trying to do, including the private sector partners, it would have saved us a lot of time.

Workshops are needed early on to get both government and private sector partners on board, to get everyone excited about doing the pilot and about learning, and to make sure everyone understands that something important is happening.

However, the earlier the decision makers responsible for your country's sandbox and related pilot effort can have those in-depth discussions with experts from other countries before starting your own pilot effort, the better.

4.7 Lesson #7: Get strong private sector partners with the right motivations involved in the pilot as early as possible.

In this pilot focused on providing access to capital, the government defines and provides the sandbox, but you need the entities to "play" in the sandbox and those are mostly private sector entities. Give the private sector participants the opportunity to experiment in the sandbox. This is an important part of the value proposition of getting them to invest their own time and effort to participate.

5. Other key points

5.1 The role of sandboxing in piloting policies, rules and governance

As we started our access to capital pilot, we saw that many of the relevant rules do not yet exist because digital access to capital for this segment of our economy is something very new. For example, there is no crowdfunding rule. There were some existing rules related to EKYC (Electronic Know Your Customer) which helped us, but in general, many of the rules governing access and use of capital in the physical world did not exist for the digital world in our country.

In our discussions with the Central Bank, it was apparent that if they created what they thought would be the necessary rules before running the pilot in the sandbox, it would get much more complex, and they would not even know what exact rules they would need at this point. We jointly agreed with the Central Bank that it would be best to do this sandbox for digital access to capital and use that to help them determine what kinds of rules to make.

Here are some examples of how we are devising and testing governance as we proceed with the pilot in the sandbox. We knew early on that we had to address data privacy. To address this, we built a token system that we are now testing. If this works, we would recommend that we have a data privacy law where the SMEs will have to allow their data to be shared for a limited amount of time as necessary to use this token for data privacy protection. Also, we need a rule on the maximum number of times in a specified time interval that a CMSME (Cottage, Micro or SME) entity can apply for a loan through this platform.

While we are piloting, we are figuring out the types of rules we need to govern this sandbox, experimenting with what these rules might be, and evaluating how they work. This is why it is so important to execute and finish the pilot, as it informs the relevant government authorities on how to go about rulemaking and governance.

5.2 The importance of having government rules that allow for sandboxing and policy experimentation within the sandbox

The Central Bank already has a rule that allows for this type of sandboxing and policy experimentation. They have something called RFFO, Regulatory Fintech Facilitation Office, which is a fancy name for operating a sandbox under the umbrella of the Central Bank. Because they had this rule allowing for sandboxing, it was easier for us to proceed with our sandboxing effort. If the Central Bank did not already have this rule that allowed for sandboxing efforts, it would have been much more difficult to do this pilot, as each new situation would have required us to go back to the Central Bank to get approval for an exception. Because they have the RFFO and the sandbox rule, we can jointly agree to say, "OK, let's try this particular experimentation within the approved sandbox."

5.3 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

In our access to capital pilot, there are a number of existing rules that apply, but the way they were written, they were not anticipating how this would be done in today's modern digital world.

In response to this, the government has added a number of addendums or made modifications to existing rules, for example, the ability to do due diligence with electronic documents and the ability to use an electronic signature without the necessity to do a physical signature verification.

There have been rule updates pertaining to the financial system infrastructure itself allowing for money to exist in a mobile phone associated with the mobile phone number without the need to have it stored in a physical stored value card or in a bank account. Before this rule change, it was necessary to have a bank account to keep your money digitally. Now in Bangladesh and all over the world, you can keep money "in" your mobile phone number. So those kinds of rules supporting digital updates to the financial system infrastructure are already in place.

Still, we have rules based on the experience of the physical world where you have to physically go and verify the CMSME entity physically exists and is "there." Currently, there is no rule that allows for this verification of existence and presence to be done digitally, but technology allows for it to be done digitally because you have the mobile phone number, and you have geolocation (though with its limitations).

Now it is possibly to verify not just the office of the CMSME entity, you can also track 24/7 where that person using the mobile phone is if you want to. However, the current law states you must go and physically verify the presence of the CMSME entity (or person). When that law was written however many years ago, geolocation technology wasn't as good or as inexpensive as it is now. Years back, you could only do it through special satellite services. Now you can just do it through your mobile network towers. Geolocation has become so easy

and inexpensive that now we have the other extreme of it being so pervasive that I get worried about privacy, but that's a different set of issues. With Google Maps and Street View, you can see a person's home. You can check the address of their business and see via Street View if they do or don't have a shop there, or if their "shop" is actually their home.

We encountered existing rules that apply to the digital setting, existing rules that don't translate to the digital setting, and some situations where there are no applicable rules because the situation only exists in the digital setting.

Sandboxing and experimentation are very helpful in the process of bridging from the underlying intent of existing laws and rules to how they sometimes need to be reconceptualized and applied in the digital realm.

In the digital economy that we are trying to support and expand to help our CMSMEs, I do not always care about the entity's physical office. If they are a digital business, I would care more about whether they have a good website or a good social media presence and what kind of activities are going there. For a digital business, this is more important to me than figuring out what's happening in the office, and I might not even care if the person goes into office or not. I care if they are making money in legitimate and lawful ways, and that they have a good marketing funnel to do that.

6. Risks associated with the digital services solution being tested in the sandbox

6.1 Cyber crime and cyber security

In February 2016, the Bangladesh Central Bank lost a large sum of money when cyber security hackers issued fraudulent instructions via the SWIFT network to illegally transfer close to US$1 billion from the Federal Reserve Bank of New York account belonging to the Bangladesh Central Bank to accounts controlled by or connected to the criminals.

In June 2023, a security researcher accidentally discovered a Bangladeshi government website that was leaking personal information including sensitive details like full names, phone numbers, email addresses, and national ID numbers of millions of citizens. Apparently, this leak wasn't caused by a hacking attempt, but rather was the result of a configuration issue on a government website resulting in the data not being properly secured.

More recently, in February 2024, there was a cyber hack against the National ID server of Bangladesh. Cyber criminals gained access to the national identity card information system and acquired sensitive personal information which they then used to provide unauthorized government services through fake government websites that they cloned.

Cyber security breaches and cyber crime is an important failure risk for a CMSME credit access platform and puts the confidential information of the CMSMEs at risk. Based on this confidential information, the banks and fintech companies are providing real credit. Suppose the confidential business information of a CMSME is modified to show they are in very good financial shape, better than they really are. Then they could get approval for a higher loan

amount than their actual data would justify which is a higher risk for the lender. Another risk is that when the bank or fintech digitally transfers the money to the CMSME entity, that digital transfer can get diverted and stolen.

There are no easy solutions given the wide range of possibilities for cyber hacks. In our pilot, we have been trying some procedures to reduce cyber crime risk such as limiting the amount of time for which the confidential information would be shared to minimize the time duration of exposure. We are also experimenting with a cap on loan size to limit the amount of money being transferred to the CMSME entity.

Can we eliminate cyber risk? Not entirely. We can only reduce the risk as best we can. There will always be a risk for cyber fraud. In parallel with building up the digital economy to provide opportunities for the CMSME segment, both the platform provider and the government must strengthen their fraud detection units. That's an entirely separate effort outside the scope of our pilot, but something that is essential to do to eventually scale this type of effort to more CMSME entities across the economy.

In summary, cyber security is the largest risk that we see. This includes closely related risk area of data security and privacy risks. There are also closely related misinformation risks, and even things like cyber bullying.

6.2 The Digital Divide

Digital divide is definitely another important risk factor and it is very different in nature than cyber security.

The technology solutions we are building and piloting are benefiting the entrepreneurs who are already digitally enabled to varying extents. This results in further supporting entrepreneurs who are already ok to some degree. Our digital approach to improving access to capital for CMSMEs means we'll be leaving out some of the poorest entrepreneurs in rural areas are do not already have access to digital infrastructure, or who don't or can't use a smartphone and don't have access to, or even knowledge of, the new type of innovative solution that we are piloting.

One must be realistic about the limitations regarding the scope of the market and population segments that can be served with the platform we are piloting for access to cash. Even with our focus on just the CMSME segment, we still cannot address the needs of all the various sub-segments of people and entities within the broader CMSME segment because this population segment is so diverse, and the needs are so different.

Staff from A2I, our government innovation agency and our partner in this sandbox and pilot effort, noted that we are leaving out many rural entrepreneurs, especially poor, female rural entrepreneurs who need help. This digital divide issue was mentioned many times in our workshops.

We also have to acknowledge the needs of our private sector partners for this pilot who are essential for being able to offer this new digital service for access to cash and financing services. They will only go after a market that has potential within foreseeable time frames. It would be impossible to get them to invest in something that will take many years to develop

into a profitable line of business such as serving the most "underserved" rural, female entrepreneurs who currently lack smart phone access and overall digital readiness. That is unfortunate, but we are not able to address those types of issues within the boundaries of this project.

While we do not feel good about this limitation, we had to accept it, and we had to start somewhere. Given the pragmatics of the time frame and scope of this project, we had to make the call of not trying to expand our scope even further to deal with these special segments of the CMSME population (e.g. rural, female entrepreneurs without access to smart phones). To the extent that our pilot will be successful and our platform for access to capital continues to scale, it will unintentionally further accentuate the digital divide. The government needs other parallel initiatives focused on reaching the rural entrepreneurs with much lower levels of digital access and knowledge and helping them in context appropriate ways, even if they do not currently have a smartphone.

We also foresee this capital access platform we are piloting for CMSMEs will work better within some sectors of our economy than in others. For sectors where there is a lot of buying and selling online, we think this platform for accessing capital and financing will work well. For some of the more conventional mainstream physical retail shops, this platform might be less advantageous. We foresee there may be skewness in terms of which sub-segments of the economy benefit from the use of this platform. This leads to the risk of amplifying another dimension of the digital divide to the extent that the services we are providing through our platform end up benefiting certain segments of industry more than other segments.

These three things - cybersecurity, digital divide and the skewness of benefiting certain parts of the industry more than others (which is a different type of digital divide) - are three important types of risks that have to be managed.

7. The benefits of better CMSME access to cash & finance being piloted in the sandbox

The service that our platform provides should be able to reduce the interest rate payments by close to 15% for the CMSME borrowers using the platform. When a CMSME business takes a loan from a (non-bank) microfinance provider, the interest rates usually range between 25% to 34%. Banks are providing interest rates of 12%, and through this platform, the CMSME borrower, who previously may not have been able to obtain a bank loan, could obtain cash or financing through this platform. Even if the banking partners working through this platform were to somewhat increase their interest rate to 15% to cover the associated risks and to provide incentives to service this segment, that would still be an interest rate savings ranging from 10% to 22%, so let's say 15% on average.

To understand the importance of such a large reduction in interest rate, remember that the borrower must make their payments to the lender on a monthly basis starting with the first month. Yet, the borrower may not be able to generate income based on that loan for several months or more. The entrepreneur usually does not make money in their first month

of starting their business. You need 5 to 6 months to incubate your business. If you are growing a crop, you need 5-6 months for the crop to grow. Therefore, having an interest rate payment from the first month onward that is 15% lower than the current alternative will be hugely beneficial to these currently underserved CMSME borrowers who make use of our platform.

There is another benefit to using our platform. In addition to charging a very high rate of interest, most of the microfinance institutions currently serving this CMSME segment also require the borrower to have a savings account with them. Yet, these are people who don't have money which is why they are taking the loan from the microfinance provider that they are not able to get from a bank. They struggle to save any money so the requirement to have a savings account in order to get the loan is an additional burden. Our platform would eliminate this additional need for the savings account as a pre-requisite for the loan in additional to providing this CMSME borrower with a loan at a much lower rate of interest than they currently receive from microfinance providers.

In summary, our platform will greatly improve the efficiency of providing loans to that sub-segment of financially underserved CMSME users who are able to access and use the platform. This is why the sandbox is so important as it enables us to trial and test this platform, the way the services are provided, and the necessary rules for regulation and governance. This provides a pathway to eventually scaling this up so more CMSMEs can access and use it. Banks who participate with the platform will be able to reach a new market segment. The participating CMSME users will benefit from substantially lower interest rates for borrowing, and consumers and the economy overall should benefit from the increased participation of those CMSMEs who are able to make use of this platform to sustain or grow their businesses.

8. Potential implications of successfully piloting and scaling this access to cash platform on existing microfinance providers

If we successfully complete the pilot and move on to subsequent phases of progressively scaling and expanding this effort, it has implications for the existing microfinance providers. One option is for some of the microfinance providers to join as platform partners. They would have to loan at a much lower rate but they would have access to a wider market. Another option is they continue with their current practices, continuing to focus on those segments of the CMSME population who are not able to access this platform and who are also unable to be served by the mainstream banking sector. However, over time, we hope those parts of the CMSME segment not using this platform gradually gets smaller and smaller as this platform becomes more accessible and more people in the most underserved sub-segments of the CMSME market segments become digitally ready. One way or another, I expect to see transitions in our domestic microfinance provider industry as this platform and perhaps other similar efforts expand over time.

9. Summary reflections on the phases of sandboxing

9.1 The conceptualisation phase of the sandbox

The ability to rapidly crowd source a lot of ideas through our workshops for what to do and how to proceed was a big advantage. Our workshops gave us access to a broad cross section of people with relevant expertise who have worked in the private and public sector for decades. We could pick their brains and come up with these ideas and do this more quickly than would have been possible otherwise. This was the best part of the conceptualisation phase. This early-stage idea generation is an example of where the format of the larger size workshops has useful benefits.

We devised a good rating system to vet the ideas to help us select the ones that made the most sense. The large workshop format was also helpful for this step as well.

Harnessing the collective brainpower of our workshop participants for the generation of possible ideas and concepts to consider, and for helping us to apply our scheme for doing a preliminary rating of these ideas were the main successes and contributions of our conceptualisation phase.

In these large workshop setting, we did face challenges in getting agreement on what ideas to select and not select. A key reason for many of these disagreements traced back to the different roles and interests of the individual participants. Private sector participants would view suggestions in terms of commercial feasibility and viability. Public sector participants would view suggestions in terms of national development and public and societal well-being. Regulatory people would have their own distinct perspective, as would people from development agencies, civil society, and other types of backgrounds.

My recommendation for how to address this unavoidable issue of disparate views resulting from different role-related perspectives, interests and incentives is as follows:

- Use the large format workshop setting to do the crowd sourcing of inputs for idea generation and for observations on the advantages and disadvantages of each idea. These large workshops are a good format for soliciting these types of inputs.
- Do not use the large format workshop setting to make the final choices of which ideas to select. During the workshop, the facilitators had to explain why the participating group should accept or heavily weight some inputs and why they might not accept or less heavily weight others. Making such decisions in real time in front of the entire group- or in front of the sub-group focused on one of our four work streams- is too contentious and too political. Also, doing this in real-time at the end of the workshop does not allow enough time to take in, process and reflect on all the inputs.

While we are doing the workshops, we should not attempt to take the final decision on which ideas, concepts, and approach to go forward in the presence of all the workshop participants.

The facilitators and the project secretariat should sort this out after the workshop and take the necessary time to think it through and figure it out. Then, after a decision has been taken, they can follow up and engage with the key participants and strategic stakeholders who were present at the workshop and diplomatically engage with them to explain why some ideas were selected and why others were not, and why the choice was made in a certain way. This approach would lead to better stakeholder engagement, provide a more conducive environment for negotiating as needed, and be more effective for gaining the necessary support for the chosen suggestions.

Finalizing choices in this way would also allow the project facilitators and leaders to more carefully consider the issue of overall project scope given the available capacity for follow on execution and thus better determine how to constrain the scope of the effort.

These reflections on the conceptualisation phase are aligned with the lessons learned summarized towards the beginning of this document.

9.2 The operations phase of the sandbox

The key to the success we have seen to date in the operational phase of our sandbox was the choice of very good partners who agreed to participate in the pilot. Because our objective was to design and pilot a platform to provide improved access to cash and financing for CMSMEs who did not have good access, we had to have private sector partners who were willing to be the lenders in our pilot. As explained in the lessons learned, we would be even further along in the pilot had we engaged in more in-depth discussions with these private sector partners earlier in the conceptualisation and operations phases.

The biggest operational challenges we initially faced and continue to face in the pilot are related to the tech stack for creating the platform.

In our project budget with our sponsors (UN DESA and the A2I innovation agency under the Bangladesh Ministry of ICT), we did not specify the need for a technology consultant or supporting technology specialist. That was a mistake. Now we realize we needed these specialized roles. They are needed for the platform to do the initial concept design, the detailed design, the implementation and deployment, and especially, to handle all of the technology and software modifications that we had to make along the way. We needed stronger technology support.

As highlighted in the lessons learned, do not underestimate the complexities of the tech stack, especially the hidden and subtle complexities that only become apparent after you start doing the pilot. These tech stack and technical infrastructure issues are as complex as dealing with all the policy, rules, and governance issues.

Related to the need for stronger technology support, we did not anticipate all the issues we would have with converting information from one format to another format. For example, we would receive images of text in our native Bengali language and have to convert this to digital characters using OCR. These types of seemingly low-level details turned out to be more difficult than we ever imagined.

9.3 The evaluation phase of the sandbox

Our target for this pilot is to demonstrate that we can use our platform to provide 100 loans. As of now, we have only provided 6, though we expect to complete the provision of the remaining 94 loans soon, over the next few months. As such, we have not yet entered our evaluation phase.

Our North Star metric is to get loan money into the accounts of the 100 CMSME borrowers who are participating in our pilot. Our first evaluation metric is to reach the milestone of getting these loans made for all 100 of these borrowers. It is not a question of how many loan requests are taken in and evaluated, because if we evaluate whatever number of loan requests but our private sector partners do not approve and transfer the actual loan money, then we are not successful. That is why we are focusing on how many loans end up getting made resulting in the loan funds ending up in the accounts of the CMSME entities that made the request.

There are more micro level KPI's that we have set and that are part of our evaluation. We have KPIs for our participating banking partners and fintech providers. But these are lower-level evaluation considerations in support of our North Star goal of getting to those 100 completed loans.

Even before we get to the point of doing our formal evaluation, we know we wanted this entire effort to have moved at a faster pace. The major reasons for these delays were already discussed in the lessons learned. Now we know that for any continuation of this effort, and for any future effort, we need to be more focused with our scope to be able to deliver tangible results.

Endnotes

[1] For background on Bijon Islam, visit the LightCastle Partners team page (https://www.lightcastlebd.com/team) and click on Bijon Islam's picture. Additional details are available on his LinkedIn profile at https://www.linkedin.com/in/bijonislam/.

[2] For more information on Aspire to Innovate (A2I), visit their official website at https://a2i.gov.bd/.

[3] For a detailed description of the project, visit https://www.undp.org/bangladesh/projects/aspire-innovate-a2i.

INTERVIEW 2: Gordon Clarke, Consultant on Payment Systems, CBDC, Banking Technology and FinTech

Date of Interview: September 30, 2024

1. Introduction to Gordon Clarke

I completed a PhD in in medical physics and started my commercial career with the Bank of England in what was then called data processing and did that for eight years. Then I moved into management consulting with one of the international accounting firms, initially concentrating on financial services, and then more specifically on payment systems following from my involvement in the early e-payment system efforts in the UK. I worked for 20 years in top-level management consulting firms with 6 years as a partner in PwC Consulting in Australia. Working in major management consulting companies, I had the opportunity to do e-payments and related financial services technology consulting work in the UK, Eastern Europe, the Middle East, Africa, Asia and Australia.

In 2002, I stopped working for the international consulting firm and started my own consulting company called Monetics Pte Ltd to focus on e-payment and related financial services work. I worked closely with central banks, commercial banks, and technology providers on many e-payment related projects in the Middle East, North Africa, Eastern Europe, the Caucasus, Maldives and Brunei. In the past five years, I have been concentrating on supporting Central Bank Digital Currency (CBDC) projects, including the technology infrastructure related to CBDC and FinTech, in the countries and regions mentioned above, as well as in the Caribbean.

1.1 The Maldives UN ESCAP project: Planning for a Central Bank Digital Currency and supporting sandboxing for policy, regulatory and execution experimentation

Given my background, I was asked in 2021 (via my colleague Dr Emir Hrnjic at the National University of Singapore) to serve as one of the lead consultants on a national study effort to assess the opportunities, risks, and challenges of using a Central Bank Digital Currency and stablecoin digital currency in the Maldives. This study was a collaboration between the Maldives government (through the Ministry of Environment, Climate Change and Technology, and the Maldives Monetary Authority) and the UN (through the Department of Economic and Social Affairs (DESA) and the Economic and Social Commission for Asia and the Pacific (ESCAP)). This effort led to the publication of an assessment report by the UN toward the end of 2022.[1]

The assessment report for this Maldives national study provided a thorough consideration of the following topics:

- The likely opportunities and challenges of digital currencies in the Maldives.
- Current gaps for implementing policy experimentation considering legal issues and monetary policy issues, and necessary conditions for sandboxing.

- Developing CBDC and stablecoin digital currency in the Maldives.
- Role of stakeholders in developing digital currency in the Maldives.
- Forecasting the demand for and use of CBDC and stablecoin in the Maldives.
- Practical policy recommendations for moving forward with CBDC and stablecoin.

After the completion of the assessment report, staff from the UN and the Maldives government followed up by publishing a policy briefing document where they further explored and elaborated on the benefits and risks associated with implementing a regulatory sandbox for CBDC in the Maldives. This policy briefing also summarised operationalization and implementation issues for the monetary authority to consider, and provided additional details on the application phase and other required phases for a regulatory sandbox.[2]

1.2 The follow-on effort to publish a Global Toolkit on regulatory sandboxing for Central Bank Digital Currency and FinTech

After completing the Maldives CBDC assessment project, I continued working with staff from UN ESCAP and DESA to create a roadmap and guide for how a country's Central Bank can establish a sandbox environment to experiment with the policy, regulation and execution aspects of CBDC and FinTech initiatives. UN ESCAP and DESA published this "Global Toolkit On Regulatory Sandbox For Central Bank Digital Currency And FinTech" report towards the end of 2023.[3]

This Global Toolkit document provided a comprehensive explanation of the following topics:

- The nature of a policy and regulatory sandbox for a Central Bank Digital Currency initiative, with supporting information on how this has been done in other countries, and on the processes and key questions for sandbox creation and execution.
- Implementation considerations for all participating stakeholders, including further details on implementation steps.
- Conclusions related to cost, benefit and risk considerations, clarification of circumstances where policy regulatory sandbox efforts are needed, and recommendations for immediate actions for getting started on the sandbox effort.
- Appendices on rules and procedures for sandbox participation for FinTechs and licensed banking institutions.

Creating this Global Toolkit document enabled us to bring together our knowledge of lessons learned from CBDC focused sandbox planning and/or execution efforts that had been done by ourselves and others in the Bahamas, Brunei, Haiti, Kazakhstan, Maldives, Nigeria, Norway, Qatar and Saudi Arabia. We synthesized and packaged this knowledge in a way that makes it generally applicable and easy for any country to make use of.

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.

1.3 The need for clarity on economic and financial policy goals when moving ahead with sandboxing for CBDC projects

As noted above, in recent years, we have had enough experience with planning and trialling CBDC projects to make it possible to codify a playbook in the form of the Global Toolkit document that contains specific supporting guidance, well-defined process models and key caution points. Since this document was published at the end of 2023, as I work with countries on a new or recently initiated CBDC initiatives, I share relevant lessons from the Global Toolkit to aid understanding of the issues and risks. This serves as a good starting point.

However, execution of a CBDC project on the ground in any specific country setting always has its own realities and challenges. Even though the Global Toolkit document now exists, there can still be challenges with getting the relevant people within a given country to absorb the key messages, or to act on them in order to go beyond the early-stage planning discussions.

There are also challenges - in fact, bigger challenges - that go beyond what is addressed by the content within the Global Toolkit document such as gaining the attention of relevant leadership in the country to clarify their economic and financial sector policy goals. Central bank and government financial policy makers need to clearly understand how the CBDC projects they are considering for their country will impact overall economic and financial sector policy goals and what practical measures are involved, including the set-up of a regulatory sandbox regime. This type of policy clarity at the outset of a national level CBDC initiative is essential and is beyond the scope of what is addressed in the Global Toolkit document (which assumes this more macro-level policy clarity).

2. CBDC projects can be retail focused or wholesale focused

From my CBDC project work across multiple countries, I am observing two camps in terms of policy goals for considering CBDC efforts. In one camp, there are central banks whose policy goals are very much to do with financial inclusion. This is often accompanied by a concurrent national policy of taking physical cash out of the economy, particularly in situations where there's a lot of corruption related to cash transactions. From a financial sector policy perspective, this is a retail side motivation for using a CBDC, accompanied by an effort to reduce physical cash in the economy. A number of emerging market countries are focusing on this retail side motivation.

The other camp has been focusing on the wholesale uses of digital currency, especially for interbank settlement, and particularly for interbank settlement in the securities markets for cross-border transactions where slow settlement times introduce a lot of risk into the very high value transactions that may be involved. A number of wealthier countries with already highly developed economies are focusing on this wholesale side motivation for using CBDC.

Some countries are pursuing both retail and wholesale applications for their CBDC efforts, especially giants like China and India.

2.1 Examples of retail focused CBDC projects in developing countries

In developing countries with smaller size economies, the economics of setting up an instant e-payment system as an element of the conventional financial market infrastructures are not good from the point of view of the commercial banks. In some of these settings, the central bank has taken a lead, and provided e-payment capabilities through a CBDC rather than by means of a mobile-handset-based e-payment system using commercial bank money in conventional bank accounts.

Creating a new format for Central Bank money (as in a CBDC), is a much bigger and more complicated decision than setting up a familiar commercially led e-payment ecosystem and infrastructure. For this reason, most of the retail-motivated CBDC initiatives have taken place in the special situations of emerging market economies where there has been an urgent need to deal with financial inclusion problems and concurrently to deal with cash problems, and where the economics of setting up a commercial e-payment system are not favourable.

However, in most of these emerging economy settings where a CBDC initiative has moved forward into pilot or live launch, take-up has only achieved a very limited degree of success in terms of adoption rates. There are a few emerging country settings including Cambodia, China, the Bahamas and Jamaica where the adoption of the CBDC is gaining some traction, but in a number of other emerging countries, the CBDC adoption by the broader public has stalled.

These national level adoption issues, especially for retail motivated CBDC efforts, are due to a lack of clear benefits for users as well as national level economic issues and policy goals. As noted earlier, these types of national level economic and policy issues are beyond the scope of what can be addressed by the type of recommendations we codified in the form of our Global Toolkit document for how to run a sandbox for a CBDC project.

For the emerging economy countries that have focused on the retail motivation for using CBDC, the difficulty is always the level of take up across the everyday population of consumers who are the retail users. Keep in mind that even with the more conventional e-payment instant payment systems implemented by the commercial banks in collaboration with the Central Bank, it usually takes a multi-year period to get widespread adoption. Take the example of Thailand (where I live) which now has a very widely-used mobile phone based instant payment system run by the banks called PromptPay. While this e-payment system now has a high rate of usage across Thailand, it was still a 5-to-7-year gradual adoption process before it achieved such a high rate of usage.

An ongoing effort in Cambodia called Bakong that has been running for about five years is a notable example of a hybrid of a CBDC together with a conventional e-payment system infrastructure. Some view Bakong as not a pure CBDC effort because Bakong is a tokenized version of the Cambodian riel, the country's national currency, and the accounting processes are unconventional. While Bakong is issued by the central bank and represents a claim on the

central bank, wallets are backed by commercial bank money. Commercial banks play a significant role in its operation, integrating Bakong into their systems and providing services to their customers. This is the normal 2-tier approach that most central banks have taken to retail CBDC to avoid competing with their constituency of commercial banks. No central bank is set up suitably to operate retail accounts, and has no desire to do so – that is the job of commercial banks.

While some CBDC experts do not consider Bakong to be a CBDC project because of its hybrid nature, I view it as a CBDC project as there is a claim on the Central Bank, and one that is working quite well as over 30% of the population is using it. I think the main reason that this initiative has been working well is the wish of the people to use the familiar national currency in a digital, tokenized form) and because leadership in both the public and private sector were worried about the economy becoming too dollarized, and they wanted to find a way to create a digital payment system based on their own national currency. This example illustrates there can be creative ways of adapting the pathway towards the direction of a CBDC that meet local needs and lead to higher rates of adoption for retail usage.

2.2 Examples of wholesale focused CBDC projects in developed countries

The other camp who are motivated to experiment with CBDC for wholesale financial purposes tends to be the more affluent nations. For example, in recent years, both Singapore and Canada have expressed on several occasions that there is no pressing reason to introduce to their respective populations a retail central bank digital currency. What has happened in these two countries, and in various other highly developed economies (notably the G7 members), is the CBDC project focus has been on wholesale uses of digital currency for interbank settlement.

As part of my ongoing work with colleagues in the Middle East, we did a benchmarking study looking at a number of the wholesale focused national level digital currency projects. We looked, for example, at Project Jura and Project Helvetia in Switzerland and the BIS (Bank for International Settlements) Project mBridge which is a multi-CBDC cross-border payment system that now includes the central banks of Thailand, Hong Kong, China, UAE and Saudi Araba as participating project members.

The focus of these CBDC projects is on wholesale usage, basically on cross-border settlement between banks and/or DvP settlement of digital securities transactions. Beneath this cross-border institution-to-institution level of interaction comes the customer (actually customer-to-institution) level of the project. What they've managed to do in these projects is to create a technology that enables the safe interchange of different central bank digital currencies. BIS Project mBridge is now operating at what they call the Minimum Viable Product (MVP) standard of performance and the project is now moving into the hands of its members as it transitions towards becoming a live commercial system. The members are inviting more countries to get involved with a view to gradually building up a sufficient critical mass for it to become a real global system that in some ways would rival key aspects of SWIFT. Although SWIFT is a telecommunication system not a settlement system and collaboration with SWIFT

may be necessary for global reach (as discussed at the 2024 SIBOS conference in Beijing 21-24 October 2024). What CBDC-based cross-border payment systems do is eliminate the need for correspondent banking relationships, which is a great benefit for less affluent countries who have difficulty obtaining and keeping such relationships going.

Meanwhile, SWIFT is not going to be pushed off its pedestal very easily as the leading means of sending cross-border messages between banks. SWIFT itself is developing some very interesting ideas about the interoperability of domestic central bank digital currencies as well as tokenized deposits (tokenized commercial bank money).

What we are now seeing in a very big way is a shift in thinking among the richer country central banks away from retail CBDC to supporting the exchange of tokenized deposits for inter-bank and commercial customer transactions, and the idea of regulated liability networks operating within a country to enable transactions in the tokens of different banks, that can be made interoperable cross-border. That is also the kind of thing that SWIFT is looking at.

Other transnational and national level monetary entities such as the BIS and the Monetary Authority of Singapore (MAS) are looking at Unified Ledger or Global Layer One (GL1), a new concept for a unified, interoperable infrastructure for cross-border payments, considering whether there could be a global distributed ledger service in which all countries could participate. This concept of unified infrastructure that banks from all countries can take part in is different from SWIFT's experimental model which is based on interoperable domestic or regional digital currency networks.

What we are seeing is a 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. This shift in thinking amongst the most highly developed economies towards using CBDC for these types of wholesale purposes is aligned with the global financial policy direction of the G20 calling for huge improvements in speed, reduction in cost, and improvements in transparency and access for cross-border transactions. Digital currency in general and CBDC in particular is seen as a way of accomplishing this.

The Central Banks in the richest countries are indicating these types of wholesale uses of CBDC are much more of a priority than trying to introduce a retail CBDC. Also, all these richer, well-developed economies, already have instant payment systems which work well and are already doing a very fine job of taking physical cash out of a steadily increasing proportion of everyday economic transactions.

Ironically, as existing instant payment systems and any digital currency related additions to retail payments continue to further take cash out of everyday economic transactions, including retail transactions, central banks are starting to think, "Well, who's going to have central bank money anymore?" and that leads to new types of challenge for the central bank with regards to managing the economy's money supply.

I would be very surprised if we see any of the G7 countries Introducing a retail CBDC in the foreseeable future, but they almost certainly will work on wholesale CBDC particularly in the cross-border context.

3. The origin and evolution of regulatory sandboxing in the context of digital innovation projects

The Financial Conduct Authority (FCA) in the UK first introduced the sandbox concept in financial services in the year 2015 as part of their "Project Innovate."[4]

The purpose for their creation of the concept and practice of a regulatory sandbox was to create a safe environment in which an innovative company could test a new financial services product that they wanted to offer to the public without the public incurring harm due to all the uncertainties, and without the existing regulations being breached. The idea of the initial FinTech focused regulatory sandbox was to allow certain types of relaxations in the financial regulatory rules which were largely administrative relaxations.

There were no relaxations on rules related to cyber or data security or related to financial accuracy and performance. However, on the administrative side, if a young FinTech company did not have the required lengthy history of having a banking license that was needed at that time to qualify for being in the payments business (which was 30 years), there was a relaxation of that type of administrative requirement to encourage both innovation and competition. In the UK and in a few other countries that were early adopters of the regulatory sandbox concept, the use of this type of sandbox worked to some extent.

As Central Bank efforts with using a regulatory sandbox expanded to other countries, there were a more country specific examples where it was difficult to get companies to agree to participate in the sandbox environment because companies were not going to make big investments in a new product where they could not see a clear path to productization and the ability to sell to the public. If there was sizable doubt about the regulatory environment for the new type of financial product, the company was reluctant to participate in the sandbox.

What started to happen was more Central Banks started to make more concessions about the breadth of new product types that could be included in these sandbox settings in order to increase private sector participation. These sandbox efforts started to expand their scope. Things like new types of lending products, crowdfunding platforms, aggregation products, insurance products and so forth started to get into the sandbox picture as this was all part of what was happening with FinTech.

This broadening of scope of Central Bank sandbox efforts naturally led to the initiation of discussions on the need for a concept of "regulatory discovery." It helps to clarify the situation here. For most of these FinTech innovation experiments, the products were not being put into a sandbox environment within the Central Bank's closed environment. Rather, what happened more often was that a sandbox effort under the supervision of the Central Bank was set up externally – in a contained and carefully controlled environment set up by the FinTech vendors and possibly other commercial banking institutions – that was under the

supervision of the Central Bank and would allow the Central Bank to observe how the product would operate and how the interactions occur across the participants in the sandbox effort.

The FinTech would run their product in this environment that the Central Bank supervised, and the Central Bank and the participating service providers would look at how this would work, and while doing this observation, go through an interaction process that was a part of regulatory discovery. The financial product vendors would say, "For this to be profitable, we have to be able to enable customers to do thus and so, and we need to be allowed to do this, and the customer needs to be allowed to do that." The Central Bank would say, "Well, yes, but in order for that to be safe for customers, we need you to design your product in this way and that way." So there was a meeting of minds to get the thing to work. In areas like lending which have quite strict rules, the back-and-forth negotiation between the Central Bank and the FinTech service providers and vendors could get quite complicated in order to find ways of satisfying the needs of both the Central Bank and the participating commercial entities.

In March 2020, the Bank of England published its first discussion paper on the opportunities, challenges and key design issues related to the use of a Central Bank Digital Currency in the UK.[5]

This discussion paper did not make any specific mention of using a regulatory sandbox to pursue further evaluation of the CBDC concept in the UK. However, it did specifically highlight the need for many questions to be carefully considered through further discussion, exploration and research, especially related to i) impact on payments and their regulation, ii) impact on monetary and financial stability, iii) the functionality and provision of the CBDC, and iv) the technology that would be most appropriate to power a CBDC and that would allow significant further innovation in payments. Hence, they implied the eventual need for a follow-on CBDC experimental program which could involve a sandboxing effort.

It was sometime after the release of this March 2020 discussion paper that the Central Bank in the UK and in other countries considering CBDC as part of their digital innovation efforts began to link follow on exploration of CBDC with their ongoing regulatory sandbox efforts that they had already been establishing for evaluating other aspects of FinTech.

4. Clarifying the terms testing, sandboxing and piloting in the context of CBDC and other FinTech projects

It is important to clarify what is meant by a test, a sandbox and a pilot in the domains of CBDC projects and other FinTech projects. In a lot of communities of people doing digital innovation, including communities outside of the CBDC and FinTech domains, these terms are often used with some degree of interchangeability and overlap. I think a lot of misunderstanding can be eliminated if the meaning of these terms are clarified.

4.1 Testing

Testing is primarily done by the vendor who is developing the new product and is focused on technical and transactional issues to determine if the product works at a strictly technical and functional level. The initial stage of technical testing can be done by sitting at a computer terminal. It does not require a sandbox regime or environment, or other types of participants, although it does need simulation of actual users to determine whether the user experience will be acceptable. This type of product verification technical testing must be done by the vendors prior to initiating a sandbox effort with the Central Bank and other participants.

4.1.1 The importance of including ease of use (usability) testing in addition to technical and functionality testing in the test phase

We recently went through the technical testing phase of one of the CBDC projects we are supporting in a middle eastern country. The vendor put a system together, tested the functionality against the specifications, and smoothed out any technical issues. Then the vendor provided access to this version to several external independent testers who used application programme interfaces (APIs) to send instructions to this test system and get results back. That additional test verification went ok and led to other technical issues being smoothed out.

The development cycle for this version of the CBDC application software did not include iterations for designing and creating “user centric” interfaces that would make it easy for the eventual real users to understand and make use of the CBDC system. The software developers only provided simple interfaces they and the testers could use to execute transactions and verify functionality and performance. They had not been asked to consider the user interface needs from the point of view of those who would later be the real users, either customers using the system to send CBDC from this bank to that bank, system administrators and maintainers, or staff responsible for overseeing the system’s usage and performance from a business perspective or from a regulatory and compliance perspective.

I was asked to be one of the testers in this technical testing phase. Because of my experience of working on many consulting projects related to payments, Real Time Gross Settlement, CBDC, and other aspects of FinTech, I could easily imagine how real users would need to use the system to perform their tasks and how the various interface screens would have to be designed to support different types of users in an intuitive way. I could see the need for many refinements to the interface to support ease of use and good user experience that were not apparent to the developers who were focused more narrowly on the software’s functionality and technical correctness. This is not a criticism of the developers. It is just pointing out the nature of the different types of testing that need to be conducted, and where sandbox regimes can help as discussed below.

For example, I could quickly detect if words used for labels on the screen were hard to understand from a User point of view, or spelled wrong. I could see where a real user would expect some type of supporting pull down menu or other types of task guidance. I could also

test various kinds of use cases, scenarios, and sequences of events that would commonly occur for a real user but might not be apparent to software developers not familiar with the way the actual users would be doing their everyday work with this new system.

In the stages of testing that need to occur prior to using the CBDC application in the sandbox, it is important to also include user interface testing. Initially, this can be done with a tame, knowledgeable "captive user" like the way I and my colleagues did this as described above. As more people representing different User roles make use of the system during the sandboxing, the user interface testing and evaluation needs to be continued. User interface testing and evaluation must also continue throughout the pilot as the user base further expands to include an even wider range of users that may include first time, naïve users unfamiliar with this type of functionality and unlikely to know all the financial sector jargon on one end of the spectrum to the other end of "wild" and aggressive users with enough expert knowledge to try to find ways of making the system fail by thinking of extreme examples, or performing sequences of actions designed to confuse the system.

4.2 Sandboxing

When the technical testing has been completed and bug fixing results are satisfactory, the test software can proceed to merge into an initial experimental sandbox regime conducted in conjunction with the Central Bank. Given that we already know the product works from a strictly technical perspective based on the prior testing, we use the sandbox setting to begin the process of understanding and approximating how the product would function in the context of the outside world and in conjunction with related Central Bank software eg for issue and use of a CBDC Or digital securities. However, we carefully control the sandbox setting and contain the type and number of participant organizations and Users, especially in the earlier phases of running the sandbox.

The broad potential customer base does not participate in sandbox testing cycles. Initially, users within the sandbox are limited to a "captive audience" of people already knowledgeable about the area, and who know they are part of a small "focus group" involved in the early stages of experimenting with the new product. This initial set of focus group users would be small and drawn from staff of the supplier, of the Central Bank, and supporting consultants and project advisors. Over successive stages of the sandbox effort, the type and number of Users may increase though the broader public is still not involved. Generally real-value transactions will not be possible in the sandbox as there may be no integration with the national payment and settlement systems, but there are some circumstances (example below) where this can be done. The important characteristic of this highly contained approach to the type and number of Users and participant institutions in the sandbox setting is that nobody is going to lose any money and everything is safe.

In this experimental sandbox environment, the Central Bank can look at how the product functions and how interactions occur. Other experts can look at it. Users can comment on the user experience and advise on changes needed to the User Interfaces to make the experience not just acceptable but positive. They identify and comment on constraints or

problems. They assess if this seems to be a great opportunity that really needs to go somewhere. If so, then, people involved in this sandbox effort begin to say, “OK, this is great, but what about regulation?”

This leads to using the sandbox setting for regulatory discovery. The sandbox participants go back and forth with the Central Bank and other relevant regulatory authorities to explore, “Does the product as it is, and its vendor, conform with relevant regulations” and/or “If we change the product or supporting process like this, then we could avoid that risk that would be a regulatory problem.” And in parallel, they explore, “But if the regulation is carefully modified like that, though in a way that still safeguards necessary protections, then this particular regulatory compliance issue could be avoided,” though of course, it is the Central Bank and other relevant regulatory authorities (and not the vendors) who make any regulatory related decisions, as well as consider if new regulations might be needed. During this regulatory discovery phase of the sandbox, participants are trying to discover through early-stage exploration and experimentation how best to manage and contain risk and at the same time how to promote innovation and growth in the market.

Some sandbox efforts, in either initial or later stages, expand the types of Users by including “captive customers” who are not part of the broader public. One example is a country recently evaluating CBDC that included staff members who purchased their lunch in the Central Bank’s canteen as sandbox participants. This was a “captive audience” of people who knew what was going on, who were provided with the knowledge and tools to use the CBDC, and who could provide their impressions of what they thought of using the CBDC to do a real task (buy their meal and other items at the canteen) and whether it was more convenient than cash or other types of instant payment. This provided a limited but still useful operational view. A number of the Central Bank CBDC sandbox efforts across the globe have used “captive audiences” of one type or another as part of their sandbox evaluation effort. Under these circumstances, certain types of administrative regulations might be relaxed in order to allow the real-value transactions to be conducted, but no security or financial accuracy rules would be relaxed.

4.2.1 Considerations for the types of participants in the sandbox phase

An important point about a regulatory sandbox is to clarify the nature of the participants. In some sandbox efforts, the participant Users are limited to the product vendors and their supporting technology infrastructure vendors. These vendors may also role-play the assumed actions of non-vendor external customers. In other sandbox efforts, or in the later states of such efforts, Users may also include a small or modest number of non-vendor external customers who already have background knowledge in the product area and who act as representative “live customer” users, though they know they are participating in a pseudo-live experimental sandbox effort to do testing and evaluation from a user’s perspective, and they are protected from losing or making “real money.” While these types of controls on the types of Users involved in the sandbox allows for experimentation with the new product and related services, the limitation is that even after observing the sandbox

results, neither the vendors nor the Central Bank would know how actual customers will respond to and use the new product or service in a market context..

Why not open up sandbox participation to a broader segment of the general public, including naïve users, and also allow for them to actually lose or make money a limited about of “real money”? There is a Catch-22 aspect to this decision. If the regulatory sandbox is to explore how these new (and untested) products will work, and to simultaneously “discover” what adaptations or additions to regulation are required in order to use these products, it is normally viewed as high risk to bring in the broader general public, including naïve customers, as real customers into this situation and ask them to participate in real unprotected transactions, especially before the product or the regulations have been suitably amended. Reputational risk to the Central Bank and the vendors is highly likely if anything goes wrong.

On the other side, if the vendors cannot get an early sense of how real customers will actually respond to this new product being evaluated, it is very difficult for them to make decisions about exiting the sandbox and investing in launching the new product.

One part of the solution to this Catch-22 situation is to move forward into a formal Pilot stage allowing a limited number of actual external “live customers” to use the live product after several iterations of rule and governance clarification have already occurred, and the Central Bank, and participating vendors have a reasonable, understanding of the nature of the risks, problems and potential regulatory hurdles that may end up emerging.

A pilot phase can involve successive stages which include participation from successively broader segments of the general population, though still under a carefully controlled and monitored conditions and supervision of the Central Bank.

4.3 Piloting and examples of recent CBDC pilots

When you start talking about a pilot, that's when the type and number of participating customers becomes more open.

Ghana is one example of a recent CBDC pilot effort. The Bank of Ghana ran pilot trials in three environments: an urban environment, a more suburban, quasi-rural environment, and a deep country environment. They provided groups of people in each of these three settings with the tools and knowledge to participate in the CBDC pilot. As this pilot was a larger scale effort than prior stages of technical testing and sandboxing, they re-verified that there was nothing technically wrong with the way the systems worked, including making sure that the way the payments worked was correct, and that there were no financial errors or security problems.

Beyond these verifications of technical and functional aspects, they used the pilot in these three different types of locations to see how people would react to using the CBDC. Staff from the Bank of Ghana said that this pilot went well and commented that the fact that the pilot was conducted by the Central Bank was important to the success of the effort as participating users indicated they had greater trust in the Central Bank than they have in commercial banks. This pilot provided a very important source of learning from a much bigger

group of users who were not all “friendly users” (as compared to the prior phase sandbox efforts where all the users were part of a friendly, “captive audience.”)

4.3.1 Large scale CBDC pilot efforts

A second CBDC pilot example is China. This is a country where digital payment in terms of electronic instant payment was already ubiquitous, so it was natural for the financial authorities to assume that nearly everybody uses mobile instant payment anyway, so the population won't have trouble accepting the idea of a Central Bank Digital Currency, which to the consumer is just another means of electronic instant payment. The Central Bank chose a number of cities, launched the CBDC pilot publicly, and notified the public mainly via the banks that they can make use of it. They also informed the public that their commercial bank would supply any user with an electronic wallet that will allow the CBDC to be used, and that there were many participating merchants who will accept it. To some extent, the new CBDC became integrated with the mechanisms that were used for Alipay and WeChat pay. Pilot merchants received a QR code at point of sale and also had all the information services and other needs they were already familiar with related to receiving electronic instant payments. It could be perceived as a third method complementing the other two.

By some accounts, over 260 million digital wallets have been opened for use of China’s CBDC, though the number of actual users and more regular, active users is very likely to be significantly smaller. Some estimates I have encountered suggest There are probably half a million people or more who are now regularly using China’s CBDC, so it has significant adoption in terms of numbers of users. At the same time, compared to the size of the population of China, it is a small number. While a lot of money is flowing through this China CBDC pilot, it is a very small proportion compared to the amount of money that goes through WeChat and AliPay. Even so, compared to size of CBDC pilots in other countries, this is a huge pilot trial. The Central Bank can use the pilot to assess what is the benefit over and above what is already available in the country in terms of the existing commercially provided retail oriented instant payment mechanisms. Aside from retail usage, they can also explore wholesale usage in inter-bank settlement.

Another large-scale CBDC pilot effort has been going on in India. This is another large country where electronic instant payment was already very widely used. When the Reserve Bank of India (the Central Bank) demonetised the cash currency notes of large denominations in 2016, that stimulated the electronic instant payment market by necessity. The Reserve Bank of India is now running two large CBDC field pilots. One is a retail pilot that works alongside the existing consumer electronic instant payment systems. The second one is a wholesale pilot which is focused on securities settlement and cross-border institution-to-institution payments.

The experiences from the large scale CBDC pilots is going to teach the global financial community a lot. This is especially the case for large scale pilots in societies where there is more open information flow as the press and other non-government observers are quite active in informing the public about experiences and lessons learned, including the types of issues or mishaps that Central Banks might not be so enthusiastic about publicising. This has

happened with the e-Naira launch in Nigeria, another country with a large population (234 million), where the adoption of the new CBDC has not been so enthusiastic. e-Naira was launched without a formal Pilot.

There are some important aspects of experience, economic and user impact, and lessons learned that can only be observed through a large-scale pilot. For example, are there really positive monetary policy benefits associated with using CBDC? How to create appropriate beneficial incentives for using CBDC or tokenized deposits (digital commercial bank money) that will encourage people and institutions to use these mechanisms over and above existing forms of electronic instant payments? Does the ability to use designated allotments of CBDC as "purpose bound money," as explored in Singapore, result in tangible social and economic benefit? Purpose bound money can only be used in special ways and/or under special circumstances, such as bill payments to utilities, welfare payments; or in unusual circumstances for disaster or emergency aid relief, or handouts under a global health pandemic emergency. "Purpose bound money" can help to combat fraud and to assure government that recipients of special distributions spend the money on everyday living necessities and not on extravagances. CBDC makes it much easier for the government to manage the distribution and usage of special distributions of purpose bound money with low fraud risk.

These are all examples of the types of questions that a Central Bank is very interested in exploring related to the use of CBDC in the economy, and reliable evidence will only gradually emerge through large scale pilots and publicly-launched programs such as those in the Caribbean and Nigeria.

5. Continuing with monitoring, adaptation and regulatory discovery after the sandbox phase and into the pilot phase

The process of monitoring, adaptation and regulatory discovery that occurred during the sandbox phase needs to continue after the transition into the pilot phase. Once the innovation effort enters the pilot phase, and increasingly wider segments of the general public are able to participate in using the product or service, you will inevitably get "wild users" who put a lot of effort into trying to break the system. There will also be well-intentioned users who happen to access the functions in some unanticipated way or make unusual errors that lead to unexpected and possibly problematic system behaviour. Eventually, and more likely with larger scale pilot usage, something will happen that was not anticipated in the design, nor addressed in specifications and not experienced in in prior experimentation during the former phases of technical testing or sandboxing.

Then, the overseeing regulatory authorities (e.g., whatever combination of the Central Bank as the banking and payments regulator, the securities regulator, the insurance regulator) need to come together, understand the unforeseen circumstances, and figure out what to do about it. Usually, they will seek first whether the problem can be addressed through a tweak

or larger change to the product or the supporting system, as changes to the regulation are more difficult and time consuming to implement.

5.1 Elaborating on the purpose of sandboxing

The main reason to deploy a sandbox regime for FinTech-related digital innovation is to ensure meeting the regulatory requirements and understanding how existing regulations apply to innovative products.

In most if not all countries, the Central Bank is responsible for the payment system and for its integrity. If you're going to get into the payment business and offer a device or a service that mediates payments, you have to be licensed and registered by the Central Bank, who may apply certain conditions to the deployment of a product. The Central Bank has requirements related to performance and how the payment system works, for the security of the system, and for making sure that no one is going to lose any money as a result of using the payment system.

Central Banks also have requirements regarding the integrity of the people who are providing such services. Are these people involved in other activities which create a conflict of interest? Or which might give them an inappropriate edge over the banks or over suppliers of other types of services due to the way their payment solution is intermingled with their other lines of business? Do they have sufficient years of prior relevant experience?

Many FinTech companies do not fit the mould of the traditional bank licensing regime. However, that does not necessarily mean that the new payment or other type of FinTech product they have does not have the requisite integrity. Even if the employees of the FinTech do not have the years of experience required as a necessity by the traditional licensing regime, it does not mean that these people do not have the necessary levels of integrity. The regulatory sandbox gives the Central Bank an environment for considering these new people, organizations and products that do not conform with the traditional licensing and regulatory regime for banks.

In a country's financial sector, sandboxing is needed is to think about the governance of the ecosystem, the nature of participants, conflict of interest issues, the rules of interacting, and the various aspects of national regulatory issues. These considerations may vary according to the nature of the financial products and services.

Additionally, there are the higher-level policy goals that we discussed earlier. We are hearing more senior government officials comment that their approach to policy regarding CBDC and other aspects of FinTech is to do policy experimentation. In other words, that their approach to policy regarding these emerging digital capabilities is to experiment to determine the more specific policies. I believe this is quite a sensible thing to say given the degree of uncertainty that exists in any economy and in all aspects of financial services, along with the uncertainty and dynamism of technology development and change.

When we wrote the Global Toolkit document, we were also thinking of sandboxing as a means of supporting policy experimentation for CBDC and FinTech efforts. For example, what happens if a FinTech comes to the Central Bank or appropriate financial authority and

says,”We have this fabulous new idea. It is a new business model that is going to be beneficial for everyone, both borrowers and lenders. However, it doesn't fit the model of the current way that your regulations work. Is there any environment that we can operate this in and show how it works? And that will enable us as the provider of this new product and you as the regulator to both see how to adapt to end up with something that’s within a regulatory framework that protects customers, and that doesn't interfere with the competitive nature of the market?”

This is where policy experimentation and regulatory evaluation and discovery comes in. You can put this type of evaluation request into a regulatory sandbox environment and try out various things to see how different transactions will work according to the different elements of the existing regulation as well as according to possible regulatory modifications. Through iterative experimentation in the sandbox setting, both the financial sector regulators, government policymakers and the FinTech product vendors can determine if they can confidently arrive at a point where there would not be a regulatory pitfall waiting just round the corner if they were to launch this new product into the market.

If those involved with and evaluating the regulatory sandbox effort are positive about the experiences observed, and confident that there is a pathway forward with respect to regulatory issues, they can agree to transition to follow-on pilot efforts that iteratively expands the type and scope of participants, and if that goes successfully, eventually live launch to the public.

Without the mechanism of a regulatory sandbox, there would be no straightforward and relatively rapid way for a country’s financial sector oversight authorities to move forward with evaluating these types of innovation requests. Lack of financial sector regulatory sandbox mechanisms would likely lead to even more FinTech entrepreneurs taking the risk of operating outside the regulatory system. For example, there are already examples in many countries of decentralized finance (DEFI) entities offing loans directly to consumers. In those countries where consumer loans are supposed to be regulated, some of these DEFI entities work outside of the licensing and regulatory regime and just go ahead and offer loans anyway completely out of the regulatory context and often paid for by a cryptocurrency such as Bitcoin. Such unlicensed and unregulated consumer lending has led to a great deal of fraud and can create a wide range of problems for customers, including loss of their money, as well as for the broader economy.

It is important to note that a systematic limitation of experimenting within a regulatory sandbox setting or within a follow-on small scale pilot setting is that a lot of retail financial products only work at scale. If you are experimenting with a new product in the lending business and there are only three people using it in a field pilot, that's not going to tell you anything useful about commercial usage or acceptance. If you get a much larger sample of pilot users, say 3,000 or 30,000 or 300,000 that much larger sample is going to tell you something more useful about how users will respond.

5.2 The risks of bypassing the sandbox phase and jumping directly to the pilot phase when operating in a regulated environment

The big danger is that you release something to a reasonably large number of people participating in the pilot and only then you discover problems that are not due to technical issues but more to do with regulatory issues. Without the sandbox phase, the system might allow the customer to do something that creates a regulatory problem, for example, related to limits on lending. Or the FinTech provider new to lending might not understand how regulatory constraints will impact their ability to access external information or their responsibilities for managing the information they collect.

For example, a new FinTech entity naïve about consuming lending regulations would think that when they make an electronic query to the Credit Information Bureau to get information about a potential customer, they would get back a complete description of that person's financial situation they can use to assess whether or not to lend. It doesn't happen in that way. Rather, what happens is the FinTech specifies to the Credit Information Bureau system that this customer wants to borrow this much against this security. What they'll get back is only a yes or no, without any other supporting information.

The FinTech cannot get more detailed information from the credit bureau because that information is private and a naïve lending product vendor (new to the industry) might not understand what private information is, what rules they must follow related to data protection and similar matters. So by skipping the regulatory sandbox phase, jumping directly to a larger scale field pilot, you run the risk of suddenly having 3000 customers (or 30,000 or 300,000 or whatever larger number of pilot participants) out there and the new FinTech is holding their customer data in cleartext (not encrypted) in some database and a hacker comes along, breaches the database and sells private information to criminals on the dark web.

Then whose fault is that? That type of situation gets very messy. The Central Bank, any other involved public sector entities, and the FinTech vendors are all trying to avoid being attributed as the party at fault for this breach due to regulatory non-compliance because they don't want to be blamed for this accident. Cybersecurity is a huge minefield and these things can very easily go wrong.

You could argue that these types of things related to a lack of understanding of regulatory requirements could all be picked up in the technical testing phase. However, the practical reality is these types of matters tend to surface, or at least are more likely to be noticed, when the CBDC or FinTech product is being used by an ecosystem of users and when there are supervisors monitoring how the entire ecosystem of usage is working and checking for "end-to-end" compliance to regulatory requirements. This happens during the sandbox phase.

Regulatory sandboxing has a special emphasis on getting the rules, the policies, and the governance correctly designed and implemented, alongside the supporting technical aspects. Also, the "customers" taking part in the sandbox setting are "tame users" who are familiar with the innovation effort and helping in one way or another with testing and

evaluation. Nobody participating in the sandbox can be harmed or lose “real money.” In a regulatory sandbox setting, these types of verifications can happen in a controlled and safe way due to the tight controls on access to the sandbox environment and the limitations on sandbox participants.

Once the effort transitions to a live, larger scale field pilot with users from the general public, it is much harder to pull back if you find that you don't have the rules, policies, regulatory compliance, governance, and supporting technical implementation correctly specified and implemented. It is much harder to recover from these types of incidents.

This is why the criteria for exiting from the sandbox (which are discussed in the Global Toolkit document) are so important. When you let a FinTech product out of the sandbox, the Central Bank and other related authorities need to be confident that the regulatory framework and pathway to enforcement is clear, and similarly that the governance framework is clear so that there isn't a big hole waiting for you just round the corner that you haven't foreseen.

5.3 Applying regulatory sandboxing concepts and practices to other industries beyond Financial Services

The need for regulatory sandboxing also applies to other industry sectors where regulation is important and where new innovations are being tested. A good example is the motor vehicle industry - driverless vehicles and driver licensing. You have similar issues about ensuring that the rules work given a new kind of product. The regulation is important because a motor vehicle can kill someone if is not roadworthy or driven with care.

So there need to be rules about how a government ensures that's the case before a vehicle is approved for general public usage. It could also be that the driver is not roadworthy in which case we have to have some kind of rules to make sure that people who can't drive safely are not allowed on the road. These same considerations about vehicle roadworthiness and driver capability need to be applied skilfully to the new product of self-driving cars that have an automated autonomous “driver,” and which may well be safer on average than the current situation. It is vital that the benefits are not lost because of clumsy or ill-informed regulatory decisions.

I highly recommend that other industries that are regulated, that provide products or services that can cause serious harm to people, adopt the same principles and practices of regulatory sandboxing that we describe in our Global Toolkit document. Even though this document was written for the specific situation of sandboxing in the context of CBDC innovation efforts and related FinTech initiatives, the essence of the approach can equally well be applied to other industry settings.

Endnotes

[1] E Hrnjic, G Clarke & UN ESCAP. National study on central bank digital currency and stablecoin in the Maldives. https://hdl.handle.net/20.500.12870/4758. 2022.

[2] UN.ESCAP, et. al. Regulatory sandbox framework for central bank digital currency in the Maldives. https://hdl.handle.net/20.500.12870/6163. 6 June 2023.

[3] S V 'Ofa, Dr Gordon Clarke *et al.* (2023). Global Toolkit on regulatory sandbox for central bank digital currency and FinTech. https://hdl.handle.net/20.500.12870/6522. 28 October 2023.

[4] UK Financial Conduct Authority. Financial Conduct Authority's Project Innovate celebrates first anniversary with plans for 'regulatory sandbox.' https://www.fca.org.uk/news/press-releases/financial-conduct-authority%E2%80%98s-project-innovate-celebrates-first-anniversary. 10 November 2015.

[5] Bank of England. Central Bank Digital Currency: Opportunities, challenges and design Discussion Paper. https://www.bankofengland.co.uk/-/media/boe/files/paper/2020/central-bank-digital-currency-opportunities-challenges-and-design.pdf. March 2020.

INTERVIEW 3: Sayran Suleimenov, Former Member of JSC KOREM, the Centralized Electricity and Power Trading Market in Kazakhstan

Submission of Sayran's written responses to interview questions: May 03 and August 06 2024, with additional updates submitted on November 17 and November 18, 2024.

1. Introduction to Sayran Suleimenov

My name is Sayran Suleimenov, and I hold a Bachelor's and Master's degree in Nuclear Power Plant Design Engineering from Tomsk Polytechnic University (Tomsk, Russia)[1].

After graduating in 2017, I returned to my home country, Kazakhstan, to put my knowledge into practice. At that time, Kazakhstan had no nuclear power plants, nor were there plans to construct any. As a result, I began working at a non-nuclear thermal power plant, where I gained hands-on experience in studying the plant's operational principles and analyzing its technological and business processes.

Later, I was offered a position in the civil service at the Electric Power Industry Development Department of Kazakhstan's Ministry of Energy. In this role, I enhanced my expertise in business-to-government (B2G) interactions between energy companies and the state, particularly in areas related to electricity generation and consumption. I also gained valuable experience in legislative processes for the energy sector and tariff regulation.

2. Origin of the concept of a digital platform for Kazakhstan's electric power industry

In 2018, my colleagues and I, including the then Director of the Department of Electricity, Kairat Rakhimov (who later became Vice Minister), came up with the idea of creating a unified digital platform for Kazakhstan's energy sector, focused on electricity generation and distribution. The primary motivation behind this initiative was the desire to optimize numerous business processes in the energy sector, including within the Ministry itself, reduce unnecessary paperwork, ensure better oversight of industry stakeholders, improve data quality and reliability, and effectively monitor and enforce the Ministry's orders and regulations.

Our platform concept envisioned its use not only by government agencies but also by energy companies, assisting them in digitalizing their internal business processes. Additionally, the platform was designed to serve as a centralized system for collecting data from equipment sensors and for recording and storing real-time operational information from each power plant. Power plants would be able to submit their reports via the platform, and digital passports for all stakeholders could be created. The platform was also intended to streamline

processes such as tariff applications, as well as monitor the execution of investment and repair programs, significantly simplifying interaction among all parties involved.

In 2019, due to personal reasons, I left the Ministry of Energy, but the idea of creating the platform remained in my mind. Between 2019 and 2022, I shifted my career focus to IT, working as a business analyst and product manager.

In May 2022, I received an offer from my former supervisor, Kairat Rakhimov (who by then had been appointed Chairman of the Board of JSC "KOREM"[2]), to join the state-owned company and, with the approval and support of the Ministry of Energy, represented by Vice Minister Ms.Zhanat Zhakmetova, restart the project of developing the "Digital Energy Platform." I had previously learned that an attempt to create this platform had been made in 2021, but unfortunately, the previous team was unsuccessful. The Ministry remained interested in reviving the project and expected JSC "KOREM" to deliver a fully developed digital platform with the necessary software modules tailored for the power sector.

After joining JSC "KOREM," my talented colleagues—Alexey Doronin, Abai Shangitbaev, and Azamat Tamamov, who had previously worked in the Ministry and major energy companies in Kazakhstan—and I collaborated to prepare a series of presentations that provided a detailed overview of the platform and its vision. These included descriptions of business processes for each module and design mockups. This information was presented to representatives of the Ministry of Energy and key stakeholders in the energy sector. We received valuable feedback and suggestions, which were incorporated into the development of the technical specifications for the software.

3. The Innovation Award from UN DESA and the Kazakhstan Ministry of Digital Development for the Digital Energy Platform

In 2021, the UN Department of Economic and Social Affairs (UN DESA) and the Ministry of Digital Development, Innovations, and Aerospace Industry of Kazakhstan jointly developed application submission, evaluation, and implementation procedures for awarding grants for innovative projects that contribute to Kazakhstan's national development. These grant awards supported the implementation of innovation projects, including the organization of expert advisory services, seminars, and study trips to explore international best practices.

JSC "KOREM" submitted a project proposal for the development of the "Digital Energy Platform" and received one of these innovation awards.

Using the grant funds and with the support of UN DESA in planning and organizing events, we conducted our First National Seminar on "Digital Transformation in the Electricity Sector" on October 31–November 1, 2022. The main objectives of the seminar included exploring digitalization trends from international and domestic experts and sharing knowledge and experiences related to policy experiments and "regulatory sandboxes."

At the seminar, we presented and demonstrated the concept of the "Digital Energy Platform" software development and held a panel discussion on the necessary modules for

the electricity sector. During this discussion, several proposals were made regarding changes to the current legislation and regulatory acts.

Two invited international experts—Michael O'Brien and Richard Stace—along with UN DESA representative Arpine Korekyan, recommended launching a "regulatory sandbox" in the energy sector to test new regulations. Based on this experience, the development and publication of amendments to the existing legislation were proposed.

For our study of international best practices, we chose the Netherlands as a case study, as the country has achieved significant success in the digitalization of its electricity sector, actively integrating digital technologies to enhance the efficiency and resilience of the industry. Some key achievements include:

- **Innovation and Smart Grids:** The Netherlands is actively implementing smart grid technologies that optimize energy consumption and integrate renewable energy sources. With a high level of digitalization in the economy (the country ranks third in the EU Digital Economy and Society Index), innovations are easily adapted to the energy sector.
- **Electric Vehicles and Charging Infrastructure:** The country is a leader in electric vehicles and the deployment of charging stations, supporting the integration of electric vehicles into the energy system and promoting sustainable mobility.
- **Cybersecurity:** Significant investments in the protection of information and operational systems within the energy sector have made the Netherlands one of the most secure players in the European market.
- **Integration of Renewable Energy:** The development of wind energy, particularly offshore, and energy storage technologies, confirms the country's role as a leader in the transition to clean energy.
- **Support for International Cooperation:** The Netherlands actively participates in international forums such as Enlit Europe, where it showcases its achievements and seeks new opportunities for the development of technologies, including artificial intelligence and cybersecurity for the energy sector.

4. Getting started on initiating a regulatory sandbox and on building the Digital Energy Platform

In December 2022, we began developing a regulatory sandbox to explore new approaches to regulating the electricity sector. At that time, Kazakhstan lacked a specialized regulatory authority for the energy industry. It turned out that a regulator existed only for the financial sector, but there was no similar structure for the energy sector, including electricity. The Ministry of Energy required an entire year to establish a regulatory unit within its structure. During this period, our efforts within the innovation project were focused on developing software modules for the digital platform, which would later become the environment for the regulatory sandbox.

At the same time, Vice Minister of Energy Ms.Zhanat Zhakhmetova successfully secured grant funding for our project, with the oil company NCOC serving as the sponsor. The company allocated $100,000 to support R&D development. As part of the collaboration, we provided monthly and quarterly reports detailing the work completed.

From January 2023 to February 2023, we conducted a competitive selection process through tenders to choose companies for the development of our platform. Unfortunately, after two rounds of bidding, no one expressed interest in participating for the specified amount. As a result, we decided to develop the platform in-house, using the resources of JSC "KOREM."

In May 2023, with the support of UN DESA, we organized a study tour to the Netherlands to learn from their national experience in structuring and restructuring their electricity sector. We studied their approach to managing the generation, transmission, distribution, and sale of electricity, as well as the roles of the government and the private sector. We met with Dutch experts on the digitalization of the energy sector, presented, and demonstrated the concept of our Digital Energy Platform. We received numerous recommendations and suggestions to improve the concept and specific software solutions.

After returning from the Netherlands, we began forming a team of software developers and analysts. I was appointed as the project manager for the development of software for our Digital Energy Platform. Our team consisted of 1 project manager, 1 tech lead, 1 senior backend developer, and 1 middle frontend developer. Due to the limited budget and large workload, I occasionally hired recent graduates and students for smaller tasks to assist the core developers. At the early stage of development, we encountered many issues related to discipline and knowledge levels, and I had to quickly replace team members to meet deadlines according to the approved action plan.

In October 2023, we held our second national seminar on "Digitalization and Reforms in the Electricity Sector." The event was attended by representatives from the government, state organizations, and participants in Kazakhstan's energy market, as well as Klass Hommes, an expert we invited from the Netherlands. We discussed the goals and plans for digitalizing the energy sector in Kazakhstan, the progress in the development and testing of our digital platform, and the challenges of coordination among industry stakeholders. We showcased the alpha version of the platform software, which received positive feedback on the necessity of its application. Participants, including representatives from the Ministry of Energy, business, and non-governmental experts, discussed the potential and mechanisms for using regulatory sandboxes in the electricity sector.

5. Pilot testing of the Digital Energy Platform

We divided the testing program into two phases:

Phase 1 was aimed at Ministry employees (the Department of Electricity and the Energy Regulatory Committee) and representatives from energy associations. Over the course

of a month, I conducted testing with them and registered the system. I explained how each module worked from the perspective of both the Ministry and the energy companies.

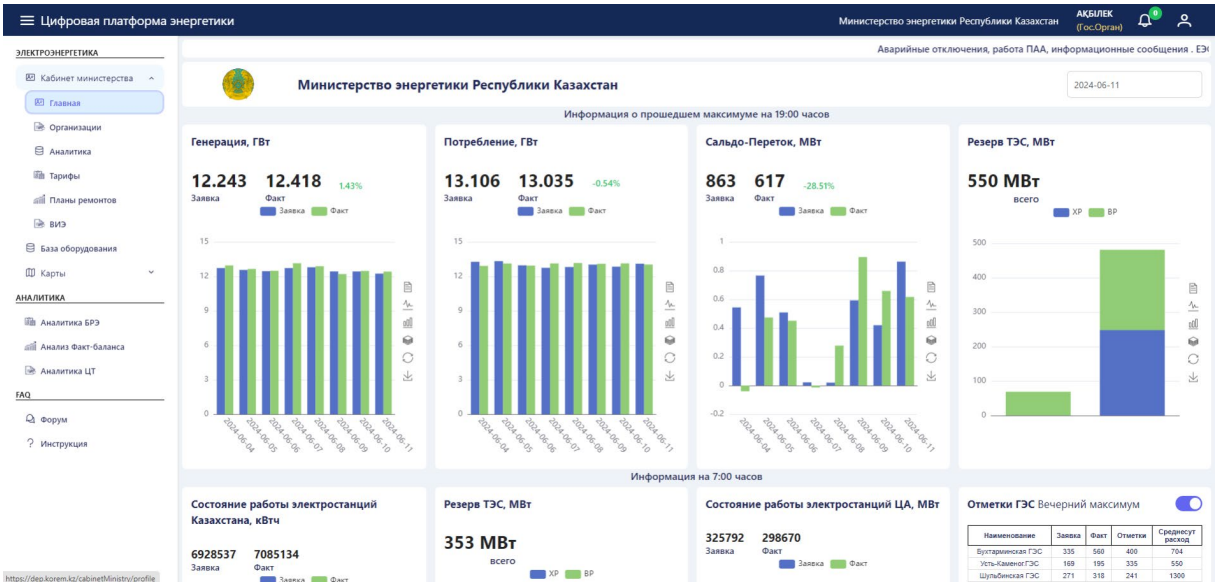

Figure 1. The main profile page of the Ministry of Energy.

We received the final recommendations and prepared the Act of Acceptance for experimental operation, which enabled us, with the support of the Ministry and associations, to begin testing with all energy companies.

In December 2023, we began **Phase 2** of the pilot operational testing of the Digital Energy Platform, involving 10 large power plants. For each plant participating in this pilot project, we assigned managers and specialists from various departments to test each of our modules. To facilitate communication, we created messaging groups for each power plant.

We developed the following modules for the power plants:

- **Power Plant Profile** – This module stored the digital passport for each entity (company information, coordinates on a digital map, equipment data, organizational structure, tariff information, maintenance records, investment programs, etc.) and industry-wide electricity and power indicators: production, consumption, cross-border flows, and losses.

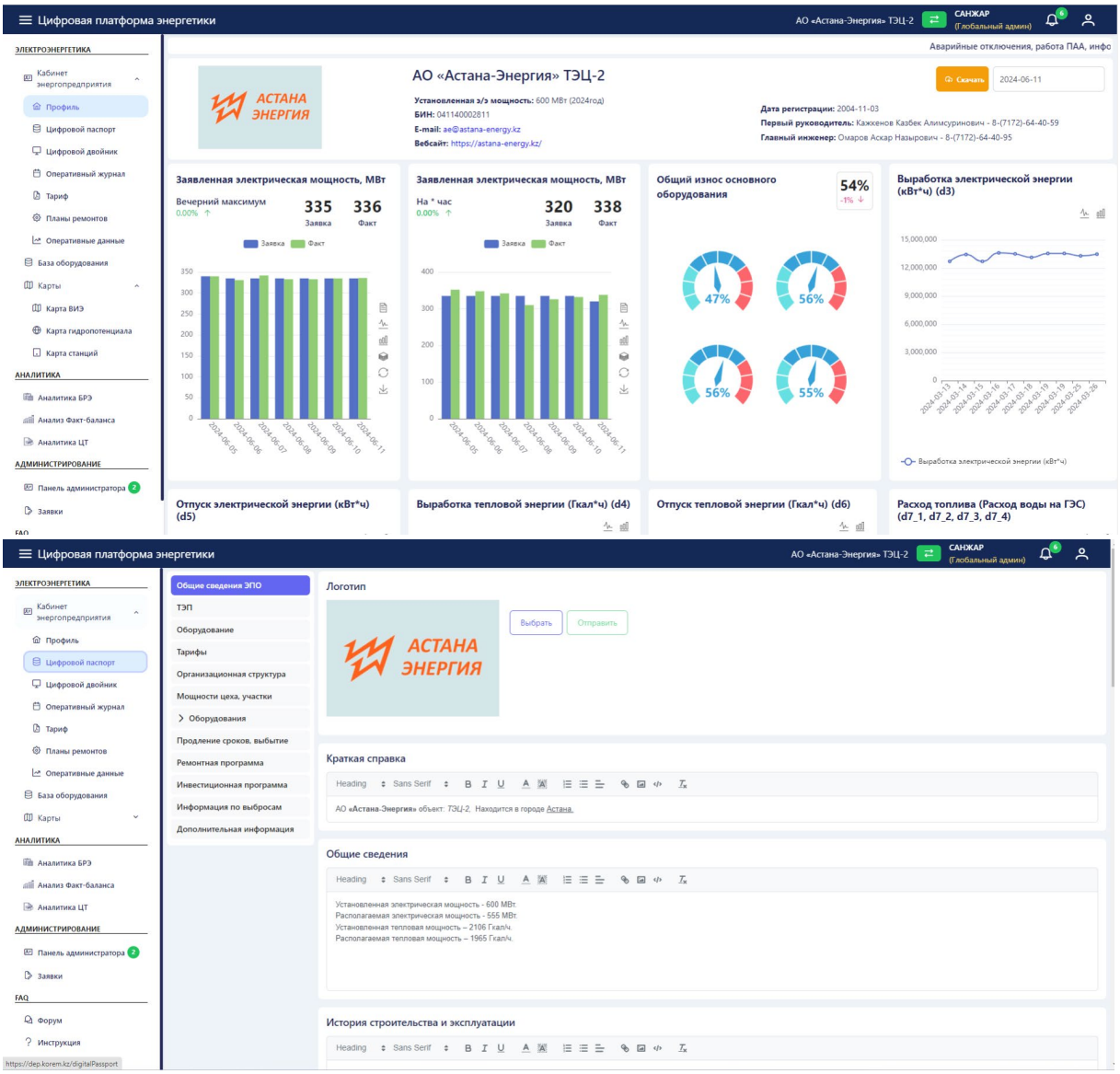

Figure 2. The main profile page of the power plants.

- **Digital Twin** – This module required integration with the power plants' dispatching systems, enabling real-time monitoring of the power plant's operations.

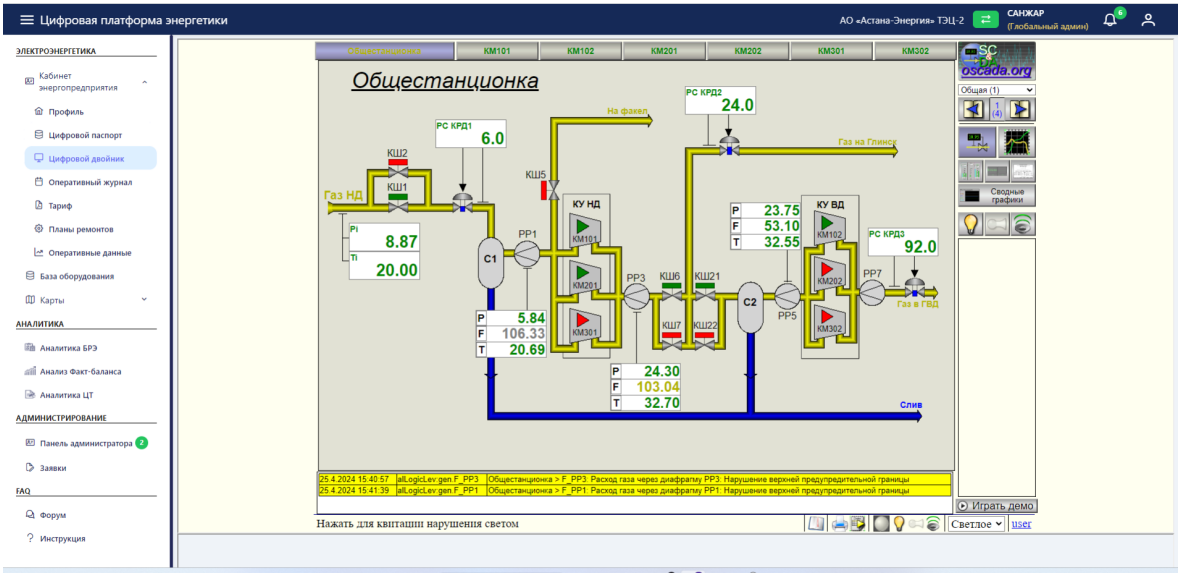

Figure 3. The Digital Twin.

- **Operational Log** – This was used for daily and monthly reporting by shift supervisors and their subordinates, as well as for task assignments. It included a calendar feature and allowed users to view historical events from any given day.

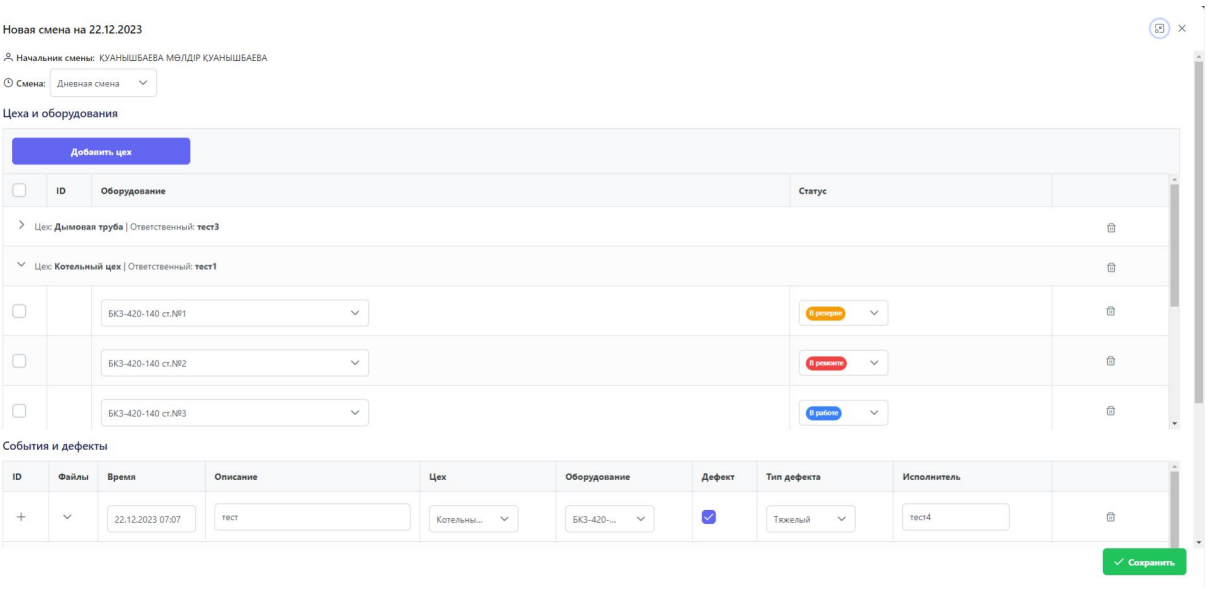

Figure 4. Power plant operational log.

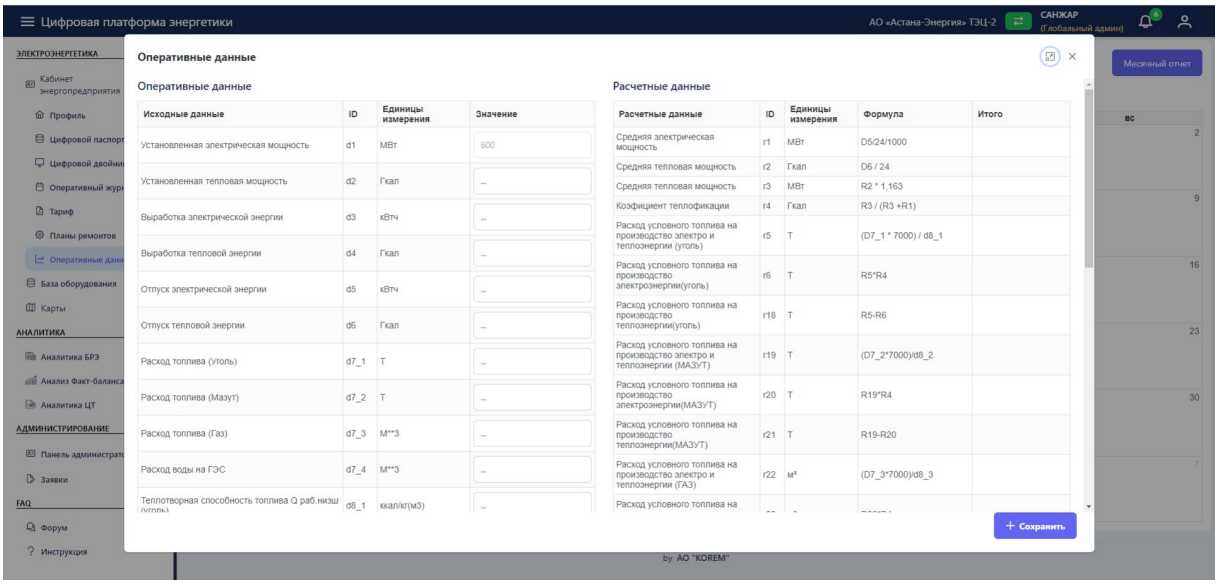

Figure 5. Power plant operational log, additional detail.

- **Tariffs** – We developed a tariff calculator, where the operating costs of the power plant and supporting documents were entered, and the system would calculate the tariff based on the data provided.

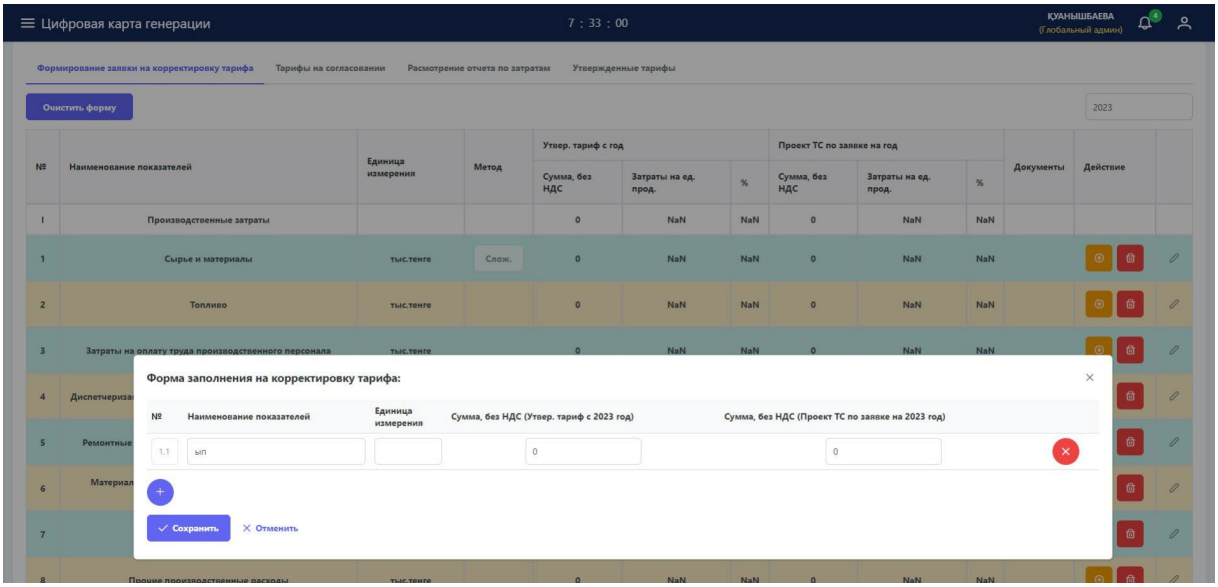

Figure 6. Tariff module.

- **Maintenance Plans** – Power plant employees could create upcoming maintenance schedules and track existing plans on a Gantt chart. It listed plant workshops and equipment, from which users could select equipment and assign maintenance tasks. They filled out the details of the planned work, such as start and end dates, type of repair, last maintenance date, and remarks.

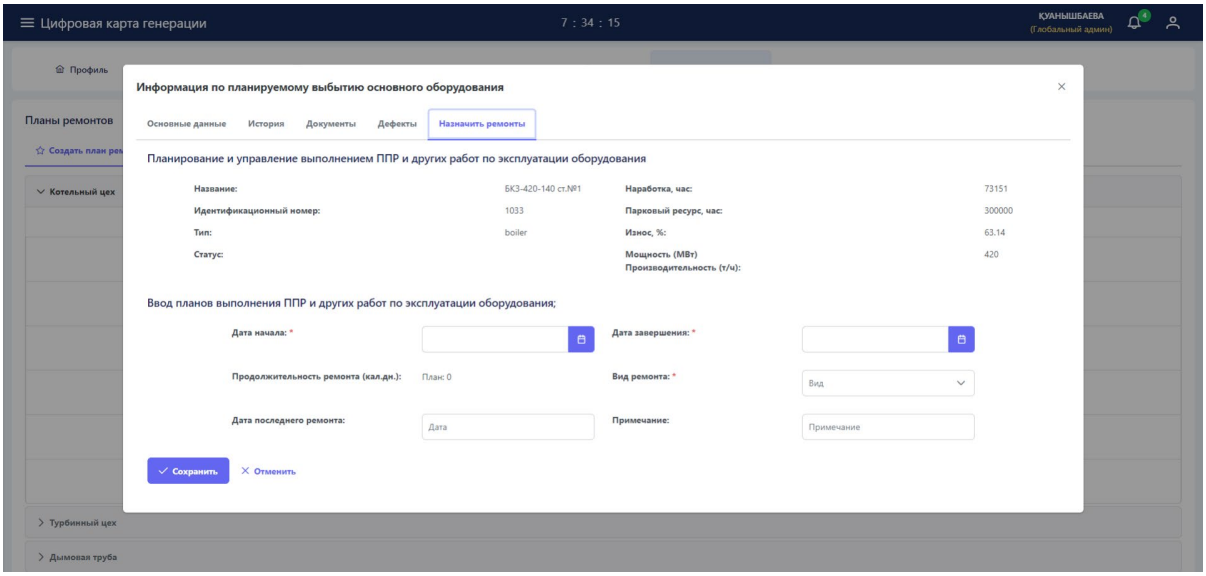

Figure 7. Power plant repair plans module.

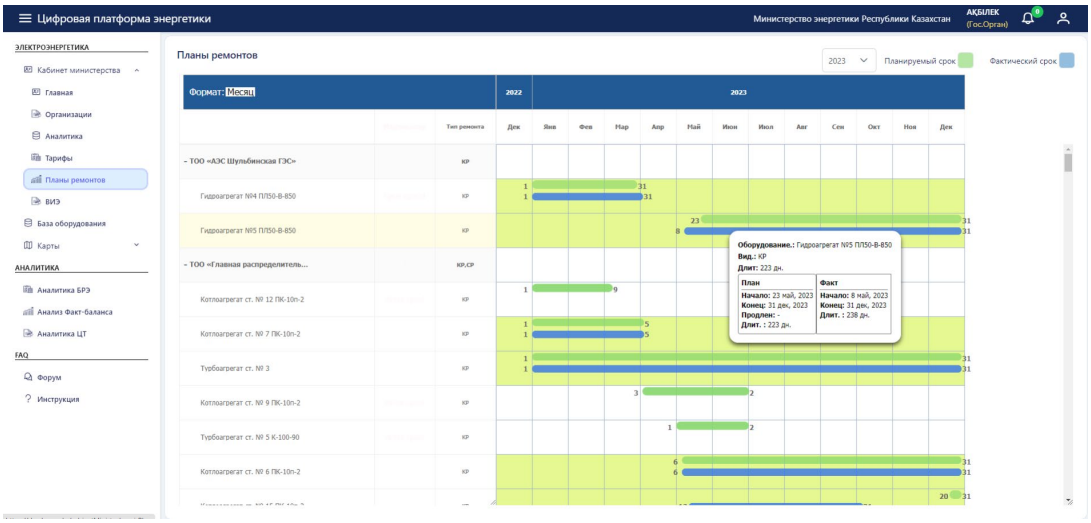

Figure 8. Gantt chart of power plant maintenance plans.

I personally conducted all the training presentations and rarely encountered negative feedback about our digital product. I had to answer all questions, from line engineers in different workshops, economists, IT specialists, to senior executives. For most employees, our platform created additional work, while for some, it simplified tasks. Some of our modules duplicated internal operating regulations at the power plants. We suggested using our modules to complement the requirements of the plants' standards. For instance, our regulatory documents required power plants to send daily reports to three recipients. Previously, this information was transmitted via phone calls, messaging apps, or emails. I proposed digitizing this business process on our platform and changing this requirement in future revisions of the regulatory documents and laws, so that power plants would enter the necessary information in one place, and each recipient would access the relevant information through the same system. Throughout the pilot project, there were many such examples where we aimed to simplify processes for all stakeholders.

During the pilot testing phase, we developed a discussion forum, where all participants could leave comments and suggestions. This forum allowed everyone to view Q&As, preventing duplication of questions. We made efforts to respond to all issues and questions promptly.

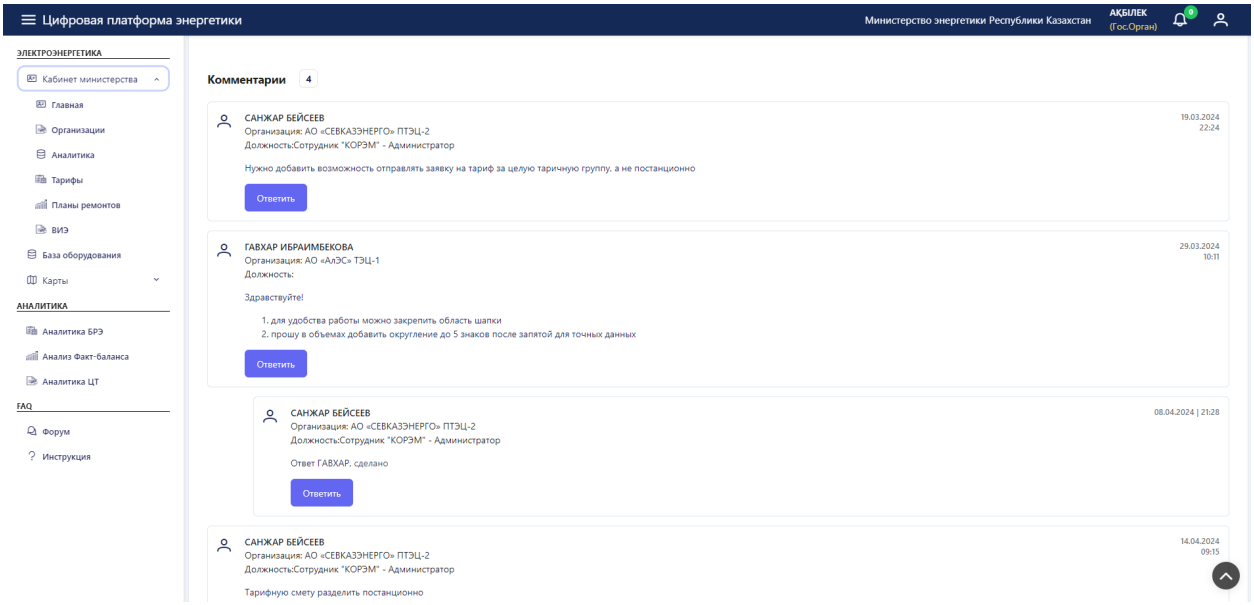

Figure 9. Discussion forum within the Digital Energy Platform used by the 10 power plants participating in the first pilot test to provide feedback on platform's performance and usage. This discussion forum speeded up the process of gathering and responding to feedback.

During the pilot project, the Ministry's Renewable Energy Department (which had not previously been involved) suggested that we develop a new module for submitting quarterly reports and updates on renewable energy sources (RES). Previously, RES entities sent their reports via email, and the Ministry staff spent a week compiling and verifying the data. We quickly developed this module, and the Ministry began using it to generate reports related to RES. The platform instantly calculated and provided a summary report for all entities, showing which ones had not submitted documents.

In one month, we registered 148 RES entities (Solar, Wind, Hydro, and Biogas), grouped them into four categories based on the type of plants, and conducted 2-3 training presentations for each group. Users liked the system because it was convenient, and it stored the history of reports, providing useful content.

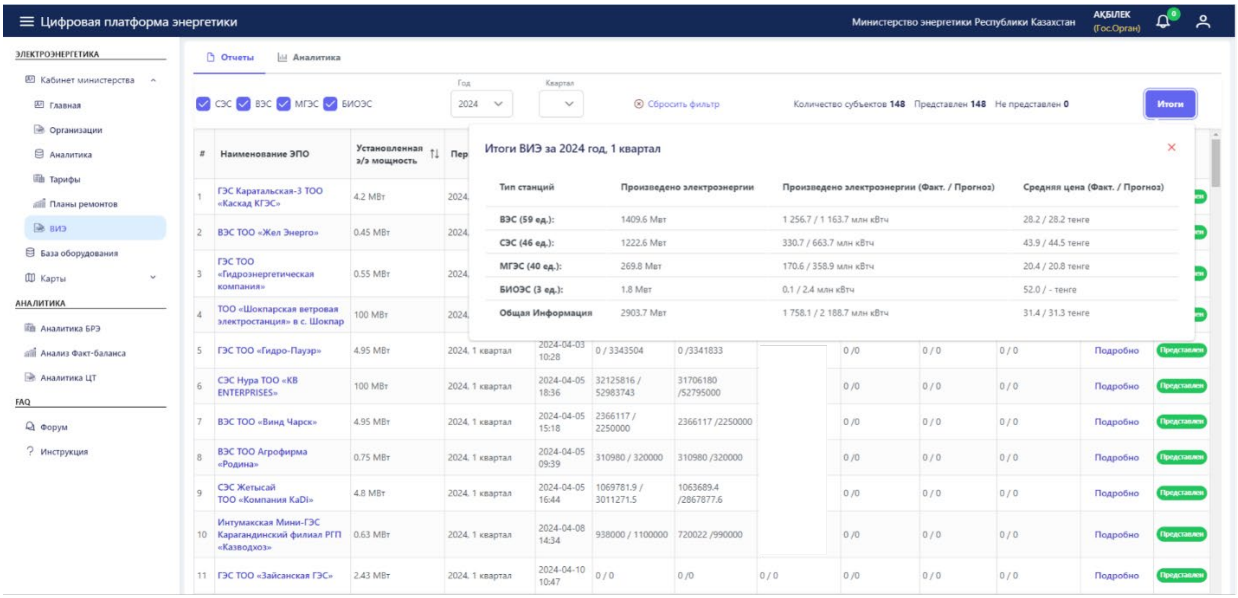

Figure 10. New repository within the Digital Energy Platform for distributing and archiving materials on renewable energy sources (RES). The need for using the platform to distribute training and update materials on RES was identified during the pilot test. This new module speeded up the process of getting the power plants and other participants in the pilot to be aware of additional materials on RES deemed important by the Ministry of Energy and other key electricity sector stakeholders.

During the pilot phase, we worked closely with the Ministry of Energy to clarify amendments to the current legislation (the Electricity and Heat Energy Laws) that were crucial for the adoption and long-term use of the Digital Energy Platform. The key provisions of these amendments were as follows: 1) to mandate that entities in the electricity sector use the platform; 2) to clarify data transfer and reporting requirements to and from the platform. These amendments were approved by the Parliament of Kazakhstan and are expected to come into force by the end of July 2024.

6. Lessons learned from our sandbox piloting and testing the Digital Energy Platform

During the pilot testing phase, we learned many valuable lessons that could be useful for future projects.

- **Project Definition:** A key aspect is a clear definition of the project at the initial stage. This includes describing the project's goals, concept, benefits, business plan, roadmap with milestones, and technical specifications for different project participants. While specific aspects of the project definition can be refined during the process, it is crucial to achieve clarity and agreement from the outset.
- **Team Formation:** The composition of the team plays a crucial role. It should include not only project developers but also stakeholders and end-users of the product. Each team member must clearly understand their responsibilities in line with the project definition, roadmap, and schedule.

- **Role of the Government Representative:** The central figure in the project should be a government representative, authorized to oversee the project, who can act as an executive coordinator and Scrum Master. Such a person, with authority and influence, would be able to address emerging issues and ensure progress. This could be the Minister, their deputy, or an authorized representative.
- **Two-Stage Approach to Pilot Project:** We found it beneficial to divide the pilot into two phases. In the first phase, we focused on a small group of the most motivated participants to quickly identify and resolve issues. In the second phase, a broader audience was involved. This approach helped minimize the impact of initial problems on all users.
- **Addressing Legislative Barriers:** At the end of the pilot, we identified legislative and regulatory barriers. This allowed us to prepare proposals for changes and submit them to the relevant authorities for inclusion in future reforms.
- **Constant Communication with Participants:** Maintaining constant communication with participants and key stakeholders is essential. We regularly participated in parliamentary hearings and conducted presentations for the Presidential Administration, Parliament, ministries, and media. It became clear that many stakeholders needed to be informed about the project to support it.
- **Prompt Response to Feedback:** A quick response to participant feedback encouraged them to share their insights. Most of the new features of the platform were suggested by users.
- **Challenges of Large Projects:** We encountered typical challenges of large projects, such as improper task delegation, procrastination, low engagement from some users, and dissatisfaction among certain participants. We addressed these issues through regular project reviews, applying Scrum/Agile principles, monitoring team morale, and minimizing task delays.
- **Working with Uninterested Participants:** We realized that in large projects, there will always be disengaged or less motivated participants. It is important to closely monitor their performance and replace them if necessary.
- **User-Centered Design:** In the early stages, we focused on the needs of the central authority, but over time, we understood the importance of usability for all participants – producers, traders, distributors, and managers. This required adjustments to functionality and interfaces, as well as the development of new modules.
- **Succession Planning:** We did not pay sufficient attention to succession planning for key team members and mechanisms for transferring accumulated knowledge. In the future, it will be important to consider these aspects for a smooth transition in case of organizational changes.

These lessons will serve as a foundation for improving the efficiency and success of future projects.

7. 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

7.1 Summary of Key Challenges and Objectives for Kazakhstan's Electricity Sector and the Digital Energy Platform Project

7.1.1 Current Challenges in the Electricity Sector

- **Aging Infrastructure:**
    - Over 80% of Kazakhstan's electricity is generated by thermal power plants, many of which are over 40 years old and nearing or exceeding their operational lifespan.
    - Coal-fired plants (70% of electricity generation) are particularly affected, with rising failures due to equipment wear and aging infrastructure.
    - Increased failures in technology, equipment, and infrastructure highlight the urgency of modernization and replacement.

- **Need for Policy Improvement and Investment:**
    - Enhanced state policies and investments are required to ensure long-term sustainability and reliability in the electricity sector.

- **Lack of Real-Time Monitoring and Planning Tools:**
    - A comprehensive digital energy platform and dashboard are needed to monitor the status of all electricity generation facilities in near real time.
    - This tool would support two critical planning cycles:
        - **Long-term planning** for plant retirements, new facility startups, and large-scale modernization efforts.
        - **Mid-term planning** for scheduling repairs and smaller modernization initiatives.

7.1.2 Objectives of the Digital Energy Platform Project

The Digital Energy Platform aims to provide accurate, comprehensive data from all electricity generating facilities to:

- Maximize production capacity by improving technical availability.
- Streamline maintenance schedules across facilities to optimize downtime.
- Enable unified and informed decision-making for modernization and repair efforts.

7.1.3 Key Tasks to Achieve Project Objectives

- **Operational Monitoring:**
  - Reduce emergency downtime through real-time data analysis.
  - Create and maintain industry-wide equipment and entity classifiers.
- **Predictive Maintenance:**
  - Monitor permissible norms of equipment wear and technical condition indices.
  - Rank equipment based on condition for prioritization of repairs and replacements.
  - Predict capacity retirements accurately.
- **Repair and Maintenance Optimization:**
  - Synchronize repair schedules system-wide.
  - Optimize intervals between repairs to minimize downtime.
- **Budget and Risk Management:**
  - Develop precise modernization budgets for production equipment.
  - Assess potential damage and costs from functional unit failures.
- **Best Practices and Standardization:**
  - Harmonize and implement maintenance best practices tailored to equipment conditions.
  - Ensure strict monitoring of repair and investment program execution.
- **Enhanced Reliability:**
  - Improve technical reliability across generating facilities.
  - Reduce gaps between installed and available capacity.

7.1.4 Integration with the Digital Energy Platform

The Digital Energy Platform, with expanded functionality to:

- **Improve Industry Coordination:**
  - Support transparency through the disclosure of key information about the electricity industry.
  - Provide written documentation of Ministry of Energy decisions to relevant stakeholders.
- **Enhance Data Flow and Decision Support:**
  - Act as an intermediary link for operational data from system operators (e.g., SCADA systems, commercial accounting, billing).
  - Facilitate the distribution and archiving of critical transition-related materials for the electricity sector.
- **Strengthen Data Security and Privacy:**

- Implement robust data protection measures for all operational, planning, and policy information across public and private sectors.

7.1.5 Strategic Importance

The Digital Energy Platform are critical for ensuring Kazakhstan's energy transition and securing its electricity infrastructure through 2035. These tools will enable informed decision-making, efficient resource allocation, and alignment of public and private sector efforts.

8. Plans for two stages of follow-up development efforts after the conclusion of the pilot

These subsequent works will be carried out in two main phases:

First Phase: Improvement of the Digital Energy Platform to support both the existing modules and new ones that will be developed. The new platform should:

- Ensure high performance and the ability to handle over two thousand concurrent user connections.
- Include administration panels, services deployed on a microservice architecture, message brokers, and notification features.
- Comply with modern information security standards.
- Enable the acceptance of data from external sources, such as SCADA systems, accounting systems, and other key data sources used in the electricity generation process.

Second Phase: In the second phase, new modules will be developed and existing ones improved. Planned modules include:

- **Energy Transmission Management Module:**
  - Includes energy grid maps, energy flows, electricity imbalances, and associated information related to organizations involved in energy transmission.
- **Energy Companies' Investment Program Management Module:**
  - Supports the interaction between energy companies and the Ministry of Energy on investment matters for the modernization or expansion of capacities and setting related tariffs.
- **Energy Companies' Personnel Database Module:**
  - Includes job search, determining the need for specialists, information on professional development, and the distribution of graduates among energy companies.

Changes in Project Management: In June 2024, the Ministry of Energy of the Republic of Kazakhstan decided to transfer the team of developers from KOREM JSC, the platform's source code, and databases to another affiliated organization performing the functions of the Energy Sector Situation-Analytical Center.

The further development of the platform includes integrating the electricity and oil sectors into a single platform and developing new functionalities and modules for both industries.

To date, the Digital Energy Platform has been renamed "Energy Tech" and is being developed without my involvement.

Acknowledgements

I would like to express my sincere gratitude to everyone who participated in this project and wish them continued success.

Special thanks to:

- **UN DESA staff:** Vincenzo Aquaro, Arpine Korekyan
- **Ministry of Energy of the Republic of Kazakhstan**: Vice Minister Sungat Essimhanov, former Vice Minister Ms. Zhanat Zhahmetova, former Vice Minister Zhandos Nurmaganbetov, Asylzhan Musin, Alexey Anisimov, Ilya Rozhkov, Gulnara Baktybaeva, Zhalgas Shogelbaev, Alma Zhukenova, Sanjar Takenov.
- **Ministry of Digital Development, Innovation and Aerospace Industry of the Republic of Kazakhstan**: Ayazhan Mukanova, Zura Kamzinova.
- **Experts**: Bas Kruimer, Klass Hommes, Michael O'Brien, Richard Stace, Almaz Saukhimov, Telman Shuriyev, Bekzhan Mukatov, Bauyrzhan Uteuliev, Saniya Idrisova. **KOREM JSC** – former CEO Kairat Rakhimov, CEO Aidos Daribaev, Alexey Doronin, Abai Shangitbaev, Mirjan Asrepov, Azamat Tamamov, Askar Sergaliev, Akbilek Sabyr, Tolybek Islamov, Janar Kaukhenova, Assel Berdymbaeva, Aliya Otebek, Savely Ermolayev, Baitemirov Esenzhan, Bapakhin Saken, Ilyas Ibraev, Nurlybai Seitkassymov.
- **Developers:** Ilias Kabden, Kuanishbaeva Moldir, Aibol Orazbekov, Rymbek Kimanov, Kanat Baibatyr, Medet Shayakhmetov, Nurlan Baitassov.

Endnotes

[1] Background on Sayran Suleimenov: https://www.linkedin.com/in/sayran-suleimenov/.

[2] Information on Joint-Stock Company "Kazakhstan electricity and power market operator," referred to as JSC KOREM: https://www.korem.kz/eng/ . JSC KOREM is a centralized trading market operator and auction organizer for conducting electronic trading of electric energy and power in Kazakhstan.

INTERVIEW 4: Cheow Hoe CHAN, Former Government Chief Digital Technology Officer, Singapore

Date of Interview: May 06, 2024

Cheow Hoe CHAN served as the Deputy Chief Executive Officer of Singapore's Government Technology Agency (GovTech) from February 2014 through February 2023. From February 2016 through May 2024, he also served as the Government Chief Digital Technology Officer for the country.[1]

1. The challenges of transitioning to cloud and digital are underestimated

People underestimate the complexity of the journey to digital and to cloud in both the government sector as well as the private sector.

People talk about agile, about cloud, about new technology including AI. But most governments and large organisations struggle with actually making the transition and doing it. They really don't know where to start because they are plagued with legacy systems. That is where the biggest challenge is. Do you rebuild? Do you try and make the change incrementally? Do you try to buy something externally this is in essence a competing system? What do you do?

These are very challenging questions that many organisations, including governments, struggle to answer today because technology is expensive. If you have already spent a large amount of money like $50 million in the past on building a system, are you going to abandon it and buy something different? And if you don't abandon your legacy effort and buy something simpler, are you going to struggle with your legacy system for next 7 to 10 years to stretch out the use of those assets and just make the best out of it?

These questions are very, very difficult to answer. The answer depends on the organisation, and especially on how important technology is to the organisation. This issue of technology's strategic importance to the company or government organisation is where the I think the most important point comes in.

"What do you want out of technology?" is a very important question to ask yourself.

For a private sector enterprise, if I am able to move to something new and better, I can ask the following types of business questions: Will I likely save money for my organisation? Will my business get better? Will my revenue increase? Those types of indicators are quite clear. If there is good evidence for a positive Return on Investment (ROI), the company will usually invest and either go greenfield, build a new system, buy software-as-a-service, or whatever form that investment might take.

Public sector has a bigger conundrum. If I stay with my legacy system today, can I just tweak it a little bit and still serve my public purpose? Or do I really have to go out there and

get something totally new, which might be going to cloud, going to AI, going to all these new things. Even if I do this, what incremental benefit do I get in terms of what I do every day as a government? Will the citizens and businesses of the country even notice or appreciate these types of changes in the first place? With government, the measure of success becomes more difficult to define. This leads to government officials needing to ask themselves, "Am I willing to invest the next $100 million on transitioning to digital/cloud/AI? Or, should I invest the next $100 million on building a new hospital? These are very difficult questions to answer.

2. Understanding the origins and progression of the cloud: from infrastructure only to the global ecosystem of software services

Let's go back 15 to 20 years ago. For the most part, the people most actively using cloud in those earlier days were start-ups. You know why? Prior to cloud, when you wanted to deliver IT services or create an IT application, the first thing you had to do was to go and buy servers and build a server room. That was horrible. It cost a lot of money just to do this set up. The upfront capital cost was often insurmountable for a start-up. You would need to spend something like $100,000 to build and equip a server room. That was $100,000 bucks down the drain even before you could start something.

In those early days of cloud, it enabled start-ups to get up and going with an application quickly, often within three months, and with only spending a very small amount on infrastructure costs, less then hundreds of dollars, and sometimes with only tens of dollars. This was very appealing to start-ups who were still in the nascent or earlier phases of acquiring customers and generating revenue streams. With cloud, they could get their application and their website out to the market, and they could run that that application for a very small monthly fee (e.g., for only $50 month.) Imagine having that fast and low-cost alternative versus the conventional approach of having to invest hundreds of thousands of dollars up front for the IT infrastructure? That was a very big impetus for start-ups to start using the cloud in those early days.

But that was not the impetus for governments. Governments already had their own data centres and IT capacity and were already investing large sums of money annually in IT infrastructure and applications.

Over time, the impacts and benefits of cloud evolved and expanded. Cloud ceased being just infrastructure. This was a significant shift of the landscape and of mindset.

When cloud initially started and in its earlier years, it was basically infrastructure-as-a service. Just pure infrastructure and nothing else. It was just replacing a data centre. Then over time, as more start-ups kept building their applications on cloud infrastructure, and as more and more departments within established companies started to make use of cloud in selective ways, the cloud gradually transitioned into being far more than just infrastructure. It evolved into becoming the cornerstone of a global ecosystem of IT libraries, services, and applications. This resulted in a big change in how services and applications were developed.

In past decades, most software applications were developed from scratch, from the ground up. All the necessary functions and many of the supporting software utilities and services were built from scratch for each project, especially for larger projects including large government software application projects. For many larger projects, each of those applications was bespoke in the sense of being very specific to a particular setting and the details of the application.

Over time, more and more libraries, components, and services were built on the cloud and were made available for reuse and sharing. In parallel, the concept of Application Programme Interfaces (APIs) came into play in a new way based on the emerging modern environment of internet, web and cloud. APIs made cloud services more accessible and usable, driving cloud adoption. Cloud platforms provided the scalability and infrastructure needed for APIs to thrive.

Once this started to happen, it was no longer necessary to build an application from scratch from the first line of custom code. It was possible to take a building block approach in the spirit of creating a complex shape out of simple, pre-existing Lego building blocks. As this ecosystem on the cloud evolved, it was no longer necessary to take two years or more to build the initial release of a custom application. Today, the initial iterations of some types of applications can be built in a short period of time, sometimes even as short as one week. Why? Because of how components and services can be sourced through the cloud ecosystem.

For example, if the application requires a means of enabling payments, I can get pre-existing cloud-based payment tools and gateway access via Stripe or from PayPal or from many others. If I need to make use of a Customer Relationship Management (CRM) application, I can get it from Salesforce and from other companies offering these types of cloud-based offerings. With this evolution of the cloud ecosystem, application developers can get many of the software services they need from the cloud ecosystem, piece them together like Lego blocks, and build whatever they want to build much more quickly.

That's a big differentiator of cloud. I was familiar with this and with other advantages because I had been using the cloud in this way for many years prior to joining the government. That's why I knew the cloud quite well.

2.1 Getting the government to understand the multiple reasons for moving to cloud

I had to get our government, including the Prime Minister, to understand that we 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. I also needed to get government officials to understand that the benefits of cloud were multidimensional and went beyond the direct cost of infrastructure.

Other important benefits related to scalability, resiliency and software ecosystems are as follows:

2.1.1 Scalability

Cloud makes it much easier and economical to deal with peak demands over key time periods such as weekly peaks, monthly peaks and yearly peaks. That eliminates the need for the government to build additional data centre capacity just to deal with the infrequent “peak-of-peaks” while for most of the year, this large amount of buffer capacity remains idle. Because the commercial cloud providers are averaging demand of large numbers of customers across their data centres, they can easily manage our periodic and seasonal peaks, and even our “peak-of-peaks.” This helps the government to manage their infrastructure costs.

Another dimension of scalability is the ability to scale very rapidly. If there is a very sudden surge in usage because some new government service or website is introduced, and suddenly 100,000 people come in at one time, I can just flip a switch or adjust a parameter and I can scale up by 200% or by however much is needed. My application will never crash due to a sudden spike in usage because I'm able to scale it up very, very quickly, not in days or weeks or months, but in seconds.

2.1.2 Resiliency

If you only have one data centre and anything goes wrong with the power supply or with cooling or with many other infrastructure details, there is a crash and that leads to big trouble. To achieve high availability there needs to be redundancy, including having a second and sometimes even additional data centres for backup that are synchronously replicated. That costs a lot of money because increasing availability by an additional decimal point (e.g., going from 99.0% to 99.9% to 99.99% to 99.999% availability and so on) results in non-linear cost increase.

The infrastructure cost doesn't go up by just a few percent to get that additional decimal point of reliability, it escalates by a much larger amount, which could be as large as 50% or 100% or even more in order to get that higher degree of resiliency and reliability. The commercial cloud is designed to have multiple availability zones. You will (almost) never get caught if one of their data centres has a crash because they have multiple availability zones for each data centre.

2.1.3 Ecosystem for software services and application development

I already mentioned above how the cloud had become the cornerstone of a global ecosystem of IT libraries, services, and applications, resulting in a big change in how services and applications were developed, making it possible for application developers to build much more quickly, sometimes even in days or weeks rather than months or years. Applications can also be deployed (put into service) more quickly in the cloud environment.

There are additional considerations where there are benefits and cost advantages of transitioning to the cloud, including many aspects of cyber security management. This is why

I thought going to the cloud was essential for the Singapore government, at least for some parts of our application portfolio.

3. Making the paradigm shift required to transition to the cloud: dealing with the fear of the unknown

One of the biggest challenges we faced when we first started on the cloud journey was that the IT and data policies in the government were never written for cloud. They were written for on-premise services. If you take that checklist and you check it against the cloud, everything doesn't work.

It's really a significant paradigm shift. You are moving from one type of paradigm and related assumptions to another paradigm and its related assumptions. That is tough because it's no more an evolutionary type of change that can be addressed with familiar and small incremental steps. It's a revolutionary type of change that requires making new and unfamiliar changes that are quite different from prior practice. It means that you have to rewrite many of the existing policies and that's where the biggest challenge is.

Governments are inherently conservative. Many government officials are reluctant to try something new that they are not nearly 100% sure of, and this is especially the case for IT security people in government who are usually even more conservative. We had to spend several years, about three years or so, working with government staff and government cyber security experts, to rewrite our IT policies to be aligned with cloud usage.

During these initial 3 years of starting our cloud journey, we sent many of our government cyber security people and some other types of IT professionals for cloud training. If we don't know something or don't understand it, we often fear it because we all have a natural fear of the unknown. As you get to know something better that was completely unfamiliar, you become less fearful as you understand what it really is and what it is not.

Eventually, we ended up retraining a generation of our government cyber security people about cloud. We sent them for very deep training that took a few months where they could learn about the security aspects of cloud, how data is protected, how encryption is done and about other aspects of cloud operations. Our staff needed to understand these things.

Take the example of cyber security for an on-premise data centre versus for the commercial cloud. These are really two different things altogether. The prevailing concept for on-premise cyber security at that time was perimeter protection. Because of the distributed and dynamic nature of cloud computing resources, perimeter protection cannot be relied on in the same way, and cloud security is based on the concept of Zero Trust Security which involves approaches such as "never trust, always verify" and "least privilege access." These were totally new cyber security concepts and practices that our people had to learn about and accept, and that was a difficult transition for some of our people. These new technical concepts for cloud cybersecurity are quite hard to accept at first, and implementing zero trust is so much more difficult than implementing perimeter security.

3.1 Early low-risk cloud pilots to test and learn

We started our pilot efforts with cloud with something very simple. We started by pushing unclassified websites to the cloud, starting with the websites of our public schools (K-12). We chose this as our very first "learning point" for piloting a government service on the cloud because school websites are mostly just informational. There are no national security implications or other high risk aspects. Even if that type of website were to be hacked and infiltrated, it could easily be shut down without major consequences.

The key is to start small both in terms of scope as well as risk. You should start piloting the use of cloud for government services with something that that's not going to blow up in your face and cause complicated problems. That is why we started with what websites that used unclassified government data. This type of piloting allowed our internal government IT and cyber security technical people and related policy people to start appreciating what the cloud could and could not do, and this gave us a much better sense of how to move forward with cloud.

The combination of the in-depth training and running these pilots with the public school web sites and subsequently with some other unclassified websites enabled our government staff to learn. Our people came to realise that using cloud was not as bad and not as onerous as they initially thought and feared before they really understood it. With the training and the early pilot experiences, our people were able to make up their own mind and assess whether what the cloud service providers were claiming actually made sense, and whether the cloud provided the required protections for that type of workload and data. And they saw that it did.

3.2 Cloud vendors needed to be less opaque and make cloud understanding less opaque

When the Singapore government started its cloud journey, at that point in time, the major cloud service providers themselves did not have a good narrative for the government. They were familiar with selling to people in the start-up space who knew all these things about how cloud worked and how to use it, and who did not have the same type of risk management concerns that a government must necessarily have. When we initially started talking to the cloud providers, they don't really appreciate nor empathise with the special concerns of the government regarding security affairs of the state, and why we were so much more concerned about potential crashes, hacking, privacy and other matters than a typical start up.

The cloud providers and their sales executives were coming from one type of culture based on their prior experiences of how to communicate with start-up companies. We were coming from an entirely different type of culture based on based on the government's natural tendency to be very conservative regarding the operation and protection of our data and IT services. We had two groups of people with very different points of view and mindsets talking to one another, and that was a challenging situation. It took time for both our side and the

cloud vendor's side to truly understand one another, and to learn how to communicate to one another about addressing our concerns.

The cloud is like an opaque blank wall. You can't see underneath it. You don't know what's behind that wall. Everything is hush hush. If everything is opaque, then I don't know what is happening behind those doors, and of course I get nervous because of fear of the unknown. So you need to unlock these doors and remove the opacity, and that includes the need for the government to understand the paradigm shifts underlying the way cloud works (e.g., understanding the difference between perimeter security and zero trust security) as cloud requires thinking about trust, transparency and control in new ways.

The government needs to be able to see and understand these new cloud concepts and technical implementations more clearly. While the government side must do its own work to prepare for this, the cloud service providers are partly responsible as well because in the past they kept things opaque, and they used to say to us, "trust me." Having the vendor ask for blind faith in our setting of public sector technology usage and IT procurement was a very scary thing. The government customer must be able to see and understand before we can make the appropriate decisions.

Over time, as the cloud providers engaged with more governments and more large corporations, they changed their approach. They realised that the communication style of saying "trust me, everything is fine" and keeping the details opaque just does not work with these types of organisations. So this situation has evolved and improved in a big way over the years since we first started on our cloud journey.

3.3 Internally building our government capability to use cloud

I brought in people with very, very deep cloud capabilities and built our first cloud team in the government. These people became the evangelists because they truly understood the technicalities behind the cloud and the resulting capabilities. The devil is in the details, so we needed our own internal technical expertise so we could address the concerns of our internal senior policy and decision makers and business users, and so we could work more effectively with the vendors. With this new internal capability for technical cloud depth, we could go beyond just talking about high level concepts.

I also hired software developers and application people from some of the FAANG (Facebook, Amazon, Apple, Netflix, Google) companies and other progressive high-tech companies known to have good software application development practices, and from start-ups. As these types of people joined our government technology group, they demanded having the type of software application development environment they were used to having previously. Our earlier hires of this type would say to me, "Hey, I can't go on the cloud. So how do I how do I do the application you asked me to do?" because they joined us before our cloud-based software development environment was eventually set up.

Once you start bringing in the right kind of technology and software application people over time, the organisation's culture has to change. The culture has to change because you are now dealing with a growing number of your technology employees for whom this new

cloud-based way of working is bread and butter. It's not something esoteric, it's the everyday bread and butter way they do their work.

This goes back to the realities of how organisational and cultural change happen both within government and in other types of organisations as well. Initially, there may only be a single voice advocating for a change (in this case, transitioning to using the commercial cloud for selective use cases). Then others begin to whisper for this change. More participants come on board. The voice gets louder, and very soon there'll be a tipping point. It starts with a single voice. It starts with a whisper, but it gets louder to a crescendo and it becomes a big change in terms of how things are done. That is what we went through.

3.4 How a relatively small number of cloud technologist and application developers made a big impact across the entire organisation

Even though we made a number of new hires, it was still a relatively small number of people in proportion to the size of our entire government IT/digital technology organisation. Yet, this small number of people had an incredibly powerful influence that propagated across our larger organisation and across other government units who were our customers.

We were able to demonstrate through a small but growing number of examples where we could create an application quickly, in about two months, and inexpensively, with a budget of about $100,000. Previously, this type of effort typically would have taken much longer, perhaps about 2 years, and cost much more, for example, $2,000,000.

When the higher up bosses and business users saw that, they really sat up and took notice!

The difference was so stark when we were first able to do this that it could not be ignored. Then we started to hear positive comments from our internal government business users and from their leadership about the fact that they asked this new (cloud-based) IT team to do something, and they did it in three weeks. This was a big surprise as the business users were accustomed to much longer turnaround times for any type of new application or service.

And then, we had the Covid-19 pandemic which was a great example of how we made use of the cloud-enabled capabilities we had built up over time by early 2020. The many facets of our government IT efforts rapidly deployed in response to Covid would have been impossible without the cloud. Changes and new requirements were being thrown at us every few days. We would have to turn around those requests in days, not weeks or months or years. This situation was changing so quickly that that we could not call a contractor and do a tender to get some aspects of this done. We had to get a small squad of own internal people to build some of these things.

You think this would have been possible if cloud was not there? It would have been impossible without cloud (including the cloud-based approach of rapid application development and deployment). During the Covid-19 response period, we had to then demonstrate the art of the possible. All of a sudden, a much wider range of people inside the government realised that cloud becomes indispensable because it's so important. It becomes part of the new normal and that's how people move forward.

4. Transitioning beyond cloud usage for unclassified information

We started our cloud efforts with what we were initially allowed to use which was unclassified systems and data. Unclassified means that this information can be public. People can see it anyway. That was an easy starting point.

A big hurdle arose came when we later moved to consider a cloud application that used information in the next classification category, restricted data (more protected than unclassified information, though still not considered as “classified” secret information). This was a big and painful effort because of all the additional policies, including cyber security policies, that apply to restricted information that did not apply to unclassified information.

Working through all of these challenges led us to a very interesting revelation. The information classification system of most governments, not just Singapore, came from the military. There is a reason for why the military classifies information the way it does. They are focused on threats to the state, threats to the country, and as they are the country’s military, they have every right to do it that way. In scenarios related to the military and to homeland security, when something (including information) gets compromised, it affects the state. This is a very serious matter which means that you must take a lot of caution about how you classify information and how you determine the trust level of people both inside and outside the government who can access that information.

However, if you take that same military mindset regarding information classification and you extrapolate it to the rest of government, you'll find that a large proportion of government agencies don't deal with national security. Rather, many government agencies deal with providing services to citizens and businesses. Is that national security related? No, it's not. It’s more like the situation of commercial entities that provide services and support to a large customer base.

When the government is providing services to citizens and businesses, the protection of personal data is very important as are other aspects of cyber and data security. But this is a different than dealing with threats to national security. If a citizen or business entity can give their data to Google, or to Amazon, or to Facebook, or to any other service provider that is in the cloud, it raises the question, “So what can’t the government make use of the cloud to provide certain types of services to citizens and businesses?” Of course, there are some caveats. This requires the government to have a highly secure, trusted and validated set up on the commercial cloud, and there would naturally be some limitations on what types of government services could be offered via the cloud.

As we go forward with government usage of the commercial cloud, this leads to the realisation that we need to come up with a different way of classifying data and specifying security requirements when national security or homeland security is not involved. Otherwise, the possibilities for using the commercial cloud for developing and delivering government services will be quite limited.

That is why I told you at the outset of this interview that this is actually a very complicated journey. It's not something that you can do over a short time period of just

months. It is a multi-year transition. There are many elements to this this journey of using the cloud for government services.

5. All paths lead to using the cloud for many civilian government services

For a government to transition to the cloud, it is much more than changing the physical IT infrastructure. It requires changes that are instrumental to every aspect of what you're trying to do with digital services across every part of the government.

It is a tough journey. However, for providing civilian services, if you want a better, more resilient system, with better scalability, with lower cost, and with better customer experience, you need to be in the cloud.

If you truly want to be digital and be agile, you need to be on the cloud. You have no choice because taking two years to make a change in your system is not agile. Taking several million dollars just to make one small change on some application is not agile. Without being agile, the government is not able to keep up with the customer expectations, so your customer experience evaluations will fall.

Let’s go one step further. If I want better customer experience and other capability enhancements, I need data. No organisation has all the data in the world, which means that you need to go on the cloud and find other public repositories of data that will help you. The thing about digital is it's data-driven. Digitalization works because of data. Where do you get supplemental sources of data? You get supporting data from the ecosystem and that ecosystem resides on the cloud.

Then you take the next step and you say, ”OK, now I am going to be more ambitious. I don't only want small data. I also want big data. I want to use the large language models and the other types of foundation models that are part of Generative AI.”

How are you going to host all these humongous things? Are you going to build your own data centre that is big enough to host thousands of GPUs? In most cases, government cannot do this. You don't have the skills for it, nor the budget for the investment. So to run these large AI models, you need go to the cloud. All these things are step by step that you need to layer on if you're serious about being digital, being agile and using the cloud. Otherwise, it doesn't work.

6. Why do cloud companies site data centres in Singapore?

Singapore is very small, and on the world scale, its economy and domestic market is small. We are also constrained in the amount of land we have, in the amount of electricity available. The weather is always hot, which requires a lot of cooling.

Yet, a number of cloud service providers have put regional data centres here. The reason is the country’s governance and rule of law.

Data and your workloads are precious assets of a company and of a government. When you have precious assets, you put them in a place where they are protected and where you know they are secure. You also want to put them in a place where there's no natural disasters.

You want to put those precious data and workload assets in a place where if something happens, there's recourse. You want a place where the cloud service providers are held to the strictest standards. Where if these service providers muck it up, there is the necessary legislation in place like our Computer Misuse, Act, Official Secrets Act, and Personal Data Protection Act that makes it possible for the government to enforce compliance to strict standards and to hold business entities accountable if there are violations. These are examples of important considerations for protecting these crown jewels in the form of your data and your workloads in the cloud.

7. 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

Some of the cloud service providers are now so big that they are putting data centres (of varying scales) in a large number of countries.

Even so, the largest cloud providers do not have a large data centre in every single country, but they have one in most of the major geographic regions. No matter where you are located, there will be a regional cloud data centre that you could make use of, even if there is not one in your country. Over time, If there's enough business volume to justify building a cloud data centre in a country where one does not yet exist, or in an additional geographic region, it will eventually happen.

For a government to get started with using cloud, you do not need to wait for a data centre to be located in your country. Suppose you want to get started with a pilot of hosting a few of your e-government services on the cloud, and you select a few applications where your data is not so sensitive (e.g., unclassified information). From my experience, every government has a lot of unclassified information that is used for providing a wide range of everyday e-government services. A surprisingly large fraction of the government's data, especially for civilian sector everyday matters, is unclassified (where data and privacy protection is very important, but where the data is not directly linked to national security or domestic security issues and in this sense, it is unclassified).

Even if there is not a cloud service provider located in your country, why can't you use one of the regional zones outside of your country, or even outside of your continent? In fact, this would be a lower cost way to get started, as the larger regional zone cloud data centres would be less expensive to use because they have higher scale, and they would also provide a wider range of cloud services. You could get started using the Singapore zone, or the Dubai zone or the Dublin zone, or several other large regional zones. What's wrong with starting that way, especially if your data is unclassified and not so sensitive? You do not have to delay

getting started with using cloud for your government e-services just because there is not yet a data centre in your country.

When you are just getting started with cloud, do not get distracted with insisting that one of the major cloud providers put a data in your country before you are willing to start using cloud for government e-services. In the early phase of getting started, you are better off using the established cloud zones and focusing your efforts on your use case and on getting the experience of working with cloud.

If you start small and steadily increase the number of cloud applications based on unclassified data, it makes a big difference very quickly.

In your cloud journey, especially in the earlier phases, keep focused on the 80% of basic things that are easier to do, many of which use unclassified data. Don't focus on the other 20% of things that are more complicated to do, and until you are ready for it, avoid applications and use cases that use more sensitive data.

8. A suggested mindset for getting started with moving government e-services to the cloud

Start small. Thing big. Act now.

Figure out how to start small, and how you learn by getting started on the journey.

Start that first step now rather than taking a few more years to get started.

Start it now. Try it out. Experiment. Learn. Modify. Experiment. Learn. Modify. That's what Agile is all about.

Then hopefully you'll get to the big picture that you initially defined. But the irony of it all is that the big picture that you think is what you will want to be doing in four or five years' time will change. That big picture plan will not be the same a few years from now because government changes, decision makers change and the needs and priorities change.

In my view, a mistake that many people in government (and industry) make in their digital/IT/cloud effort is to spend too much time on the big picture plan, and to delay getting started while waiting for that big picture plan to be completed. That results in the government saying they will spend more time planning and then get started in three years' time, and based on the plan, complete the work over the next 7 to 10 years. In my view, ten years from now, the whole government will have changed in so many ways that today's ten-year plan for the areas of digital/IT/cloud and AI will not be relevant.

They key point is that agility is important. You need to start small and start now.

Yes, you also need to think big and anticipate the future. You need to have a vision for achieving lofty goals and objectives. But you have to start now, start small, and develop and demonstrate the real capabilities you need for moving in the direction of your ever evolving big picture goals.

9. Keep focused on the real-world problem you are trying to solve vs using new technology for its own sake

As part of selecting the right starting point for starting small, you need to understand what problem you're trying to solve. You really need to go back to that fundamental thing. Referring to the famous expression from the book “Alice in Wonderland”, if you don't know where you want to go, then it doesn't matter which direction you take. So you have to first figure out where you want to go, and this leads you to clarify what problem you really want to address and solve.

If you are very mindful about what probably trying to solve, a lot of things will follow suit, and much of the rest are supporting details. This will lead to a good pathway for starting small. Also keep in mind that there are many innovative ways to solve a problem, and sometimes you don't even need to use new or advanced technology to achieve the desired outcomes. So look for the simplest ways to go about addressing the problem you focus on. Try not to get enamoured by technology for its own sake just to say you are using the latest and greatest. Don’t get seduced by technology just to play with it. Keep focused on the fact that you are using the technology to solve a real problem.

Endnotes

[1] For background on Cheow Hoe Chan, visit his LinkedIn profile at https://www.linkedin.com/in/cheow-hoe-chan-92646215/. For background on Cheow Hoe Chan’s work prior to his relinquishing his full-time role at GovTech and Smart Nation Group, visit https://blogs.worldbank.org/en/team/c/cheow-hoe-chan, November 2022 and https://www.itnews.asia/news/chan-cheow-hoe-takes-up-advisory-role-in-edb-590501, February 2023.

INTERVIEW 5: Prof Ramayya Krishnan, Carnegie Mellon University, USA

Date of Interview: April 24, 2024

1. Introduction to Ramayya Krishnan's background and involvement in public sector AI

Steven Miller: Please self-introduce yourself to state your title and position and the way you like to be referred to professionally.

Ramayya Krishnan: I go by Krishnan. I'm Ramaya Krishnan.[1] Officially I am the Dean of the Heinz College (Information Systems, Public Policy and Management) at Carnegie Mellon University and the founding faculty director of the Block Centre for Technology and Society.[2]

Steven Miller: Can you briefly and broadly explain how you have come to be knowledgeable about the use of AI capabilities in public sector settings? In other words, specific ways in which you've been close to public sector organisations where you've come to develop substantive knowledge of how they're using AI capabilities in Public Sector use cases.

Ramayya Krishnan: I'm fortunate to be at a college that is home to two schools, a School of Information Systems and a School of Public Policy. This combination is a fairly unique co-location of these two types of schools. Also, the Heinz College is located at Carnegie Mellon University with very low boundaries (barriers) between the School of Computer Science, the School of Engineering and the Heinz College. So, just by virtue of location, I am constantly exposed to the use of AI and advanced technology in public sector settings.

The capacity to understand not only technology, but also how it's operationalized and put to use in public sector settings is very much a part of what Heinz's core competency is. That's the first point. The second point is that we are embedded in the city of Pittsburgh in Allegheny County. When the Heinz College was founded back in 1968 by Professor William Cooper (and at that time called the School of Urban and Public Affairs), there was a strong focus on connecting to problems of society, public sector problems which could be city level and county level in the case of Pittsburgh, but also state and federal level.

The founding purpose of the Heinz College (including the prior School of Urban and Public Affairs) was to try and bring the power of analytics technology and computing technology (and now more broadly digital technology) to supporting consequential public sector decision making. So this purpose goes back to 1968. In fact, Professor William Cooper's vision as the founding Dean was for the Heinz College to educate men and women to be capable of "intelligent action." Intelligence was about the use of data and analytics and action was to engage in public sector and societal decision making that actually change the world for good.

So, with that history in mind, from very early on, from when I first arrived at Carnegie Mellon, which is now over 34 years ago, I've been engaged in the use of various kinds of decision support and AI technologies in public sector settings. My PhD thesis completed in 1989 dealt with the use of expert systems for decision support. Very early on in my faculty research career, I started doing project work with the Census Bureau on issues of privacy and confidentiality. Later on, I also started collaborating with local city and county teams to help them better understand how to use data analytics to think about better ways of providing services to support the public good.

My many years of combining my faculty research work and public sector engagement at Carnegie Mellon helped the university to establish large externally funded applied research centres with government entities such as Metro 21[3] and Traffic 21.[4]

So, the idea of using digitization, data and analytics to support public sector decision making has been a core part of my everyday work for decades as a faculty member, researcher, Dean, Research Centre participant and Centre Director. This has given me in-depth exposure to many examples of using AI and related aspects of Data Science and Decision Science in public sector settings to address such issues as how best to provide transportation, or how to better understand and respond to underserved communities, or how best to support individuals who are eligible for public sector services but who are not enrolled in local or state government programmes designed to support them with the intention of understanding why that was the case -- these are just a few illustrations of common examples that arise and that I am regularly exposed to.

In the last five years I have had a heightened engagement with the public sector in three regards. One was a joint effort with the state of Pennsylvania during the height of the Covid pandemic where I led a team of faculty from CMU to support Governor Tom Wolf's policy making and decision-making efforts related to when and how to reopen the economy after the initial lockdowns. Using data science, decision science and AI, we developed a dashboard and related models to support state decision makers with recovery and reopening policy.[5] This was a significant effort to use data-driven approaches, including supporting AI models, to address many of the issues of how to balance public health while maintaining economic vitality during the pandemic.

Second, I have been serving on the Government Accountability Office's (GAO) Educator's Advisory Board that is chaired by the Comptroller General of the United States government. The GAO has a unique function in the federal government in terms of auditing and dealing with compliance. I've been actively involved in the GAO's Educator Advisory Panel for nearly a decade and a half.[6]

Third, I serve on the US National AI Advisory Committee (NAIAC) to the President.[7] That's been a very active role since my appointment in May of 2022.

As part of this, I also co-chair the NAIAC's Working Group on AI Futures (Preparedness, Opportunities, and Competitiveness). I also serve as chair of the DOD's Responsible AI academic advisory council.

These three examples are US centric. I've also been involved in international efforts related to the use of AI in public sector with the Asian Development Bank where I serve on the Presidential Advisory Board.

In summary, I have both wide and deep range of engagements related to the use of AI in the public sector.

2. The importance of policy experimentation and sandboxing for AI applications in public sector settings

Steven Miller: As a prelude, before we dive deep into specific examples, I'm going to read you a few summary extracts from the Terms of Reference document that defines this project I am now doing for the UN's Division of Public Institutions and Digital Government (under the Department of Economic and Social Affairs). I want to make sure that you understand each one of these summary statements and solicit your comments.

The first statement is as follows: "In recent years, relatively new approaches of policy experimentation and regulatory sandboxes have emerged amongst countries and have proven to be effective in creating a more conducive and contained space where governments, in partnership with relevant stakeholders, can experiment and trial with digital technologies and innovations at the edge of, or outside of the existing policy space or regulatory framework."

Do you have any comments on this statement with respect to what's now going on with AI usage in the public sector?

Ramayya Krishnan: Absolutely. In fact, this same point is part of one of the recommendations from the "Year 1 Report" of the ÙS National AI Advisory Council.[8]

We specifically talked about regulatory sandboxes. The UK has been at the forefront of articulating the need for such regulatory sandboxes particularly in terms of understanding how to come up with operational ways of understanding and doing risk management with AI.[9]

The EU AI Act, which recently been endorsed by the European Parliament (but has not yet been formally adopted by the Council of the European Union) has a risk tier framework associated with AI use cases.[10,11]

This act defines use cases considered to harmful practices where AI should not be used at all. This includes use of 'real-time' remote biometric identification (surveillance) in public spaces using sensitive attributes (political party, sexual orientation, etc.) and use cases that involve evaluating or classifying individuals or groups based on social behaviour or personal characteristics, leading to detrimental or disproportionate treatment in unrelated contexts (social scoring) are prohibited. There are exceptions for law enforcement (specific necessary objectives defined in the act).

High risk AI applications include those that impact a person's employment or healthcare. Then there's medium risk AI applications and then low risk applications of the sort where you might get a recommendation on Amazon or Netflix and the consequentiality of that type of output is not as significant then when the risk tier is higher.

2.1 We need better tools for measuring and evaluating AI

While these policies are useful, the important question is, "what are the right sets of tools to go from AI related policies to real world practise?"

For instance, think about a generative AI model. How do you audit the data and the outputs to assess its degree of toxicity and the nature and extent of its bias? For a classifier AI model, how do you determine the false positive and false negative rates it commits? And the nature of and impact of those misclassifications? How do you assess if the AI model has the right thresholds set that determine false positive and false negative rates, especially for more consequential use cases?

What should be the data sets that should be available to allow for the testing of these AIs to evaluate them? These types of questions are very much tied to this idea of saying **we really need to have a science of measuring and evaluating AIs.** It is still very much early-stage work-in-progress especially for evaluating the more recent generations of AI models (deep learning models, reinforcement learning models, large-scale pre-trained foundation models for language or for vision, multi-modal models, generative AI models for creating language, computer code, image, video and audio content). Better methods to measure and evaluate AIs need to be created. Also needed are the standards against which evaluation is to be conducted. This is also work in progress. As they are created, they can provide the necessary guidance for policy formulation and policy implementation. Sandboxes are a useful device (though the exact meaning and specification of a sandbox is required) to be able to do this.

In the examples that I just mentioned, where I asked about the data sets that should be available to allow for the testing of an AI to evaluate it, consider the situation of a public sector organisation considering whether to purchase a commercially available tool from an AI product vendor. The public sector unit (department, agency, ministry) would have a particular intended use case in mind. They need their own sandbox, or access to a trusted external sandbox, that they can effectively think of as being like an "Underwriters Lab" for being able to test the AI to determine whether it meets certain thresholds that the agency needs it to meet (standards it needs to comply to) prior to acquisition and deployment. There are a lot of important, unanswered questions about how such regulatory sandboxes should be designed. For example, how should it be done in a way where you can't have adversarial gaming of the sandbox by the vendor? The main point I want to convey is that we need ways in which we can assess and evaluate AI in consequential settings.

2.2 Pennsylvania creates a governing mechanism for AI-related policy experimentation and sandboxing

In the state of Pennsylvania, Governor Josh Shapiro signed an executive order in September 2023 expanding and governing the use of Generative Artificial Intelligence Technologies within the Commonwealth of Pennsylvania.[12, 13] This Executive Order called for the creation of a Generative AI Governing Board for the Pennsylvania state government to provide guidance and direction on the design, development, procurement, and deployment

of Generative AI in government decision making, and on policies for the utilization of Generative AI. The executive order also specified that state agencies shall proactively and iteratively evaluate and present to the AI Governing Board for consideration potential salient use cases of Generative AI. In essence, this state of Pennsylvania Executive Order creates a governing mechanism for AI-related policy experimentation and sandboxing across the state government.

In January 2024, Pennsylvania Governor Josh Shapiro announced a new partnership between the state government and the company OpenAI to launch a Generative AI Pilot for state employees using OpenAI's product, ChatGPT Enterprise.[14] This pilot effort will enable state employees to test where and how generative AI tools can be safely and securely leveraged in their daily operations. The lessons learned, evaluation results and other findings emerging from this large-scale pilot programme will be used to guide the wider integration of Generative AI technology into state government operations.

This pilot effort resulted from a recommendation made by Pennsylvania's Generative AI Governing Board. The partnership announcement also noted that throughout the duration of this pilot, the state of Pennsylvania will collaborate with Carnegie Mellon's Block Centre for Technology and Society (housed within the Heinz College) as a knowledge partner. Coordinated through our Block Centre, Carnegie Mellon faculty and staff will provide expertise and guidance on how to plan, design and evaluate pilot efforts. Hence, this pilot is a three-way collaboration involving the public sector institution (the state of Pennsylvania), a private sector industry partner (OpenAI) and a higher education academic partner (Carnegie Mellon University).

From my involvement with both national and state-level public sector AI efforts, I see that one of the things that government officials are keen on understanding is how to procure AIs. How do you assess and evaluate these systems as part of a procurement process? How do we "Red Team" them prior to deploying them in consequential settings? You could imagine a regulatory sandbox being a special type of "evaluation device." What exactly is this regulatory sandbox evaluation device? You have to define what it is. How do you design it? All of those are still open questions, but as a concept, it's about trying to address these sets of issues.

Steven Miller: When you talked about sandboxes, you used the term "regulatory sandboxes." Do sandboxes for this kind of experimentation always have to be regulatory focused? Are there other aspects of public sector policy experimentation that might not specifically be regulatory with respect to testing and understanding how to get stakeholders to work with outputs of AI enabled systems?

Ramayya Krishnan: At the risk of being too abstract, consider two ways AI can be used by an agency – a) it could be used by the agency to support its own decision-making processes and b) it could be used to regulate other private sector or non-profit organisations. In both cases, what the agency can do is determined by the law or by rulemaking. So sandboxes can be used to support either type of AI use. The adjective "regulatory" is being used in the public sector context because almost everything the public sector does is set within the framework that it's

authorised to do something because of either a law or some rule making or some executive order.

3. Questions related to due diligence, governance, procurement and the necessary talent required to do these things given AI is moving so rapidly

Steven Miller: Here's the second high level statement, which is an assumption and you can agree or disagree or elaborate:

"In today's hybrid digital error AI's complexity, speed of development, broad applications and dual use potential presents inherent challenges for governance in today's slow and siloed policy making context."

Does your experience align with that statement? Or do you have a different take on that?

Ramayya Krishnan: There is much truth to that statement. By the way, both in the private sector and the public sector, that there's a general sense that technology is moving so rapidly. This a multi-dimensional problem – one has to address organisational and governance challenges as I will illustrate using some examples.

An organisational version of this question is that I have a business process which could be, for instance, some aspect of how do I best serve citizens in a public sector type of setting? Consider the following example. Let's say that an agency is trying to help displaced workers find new jobs.

As part of this effort, could workers use a chatbot, put their resume in that chatbot, and determine what roles they're well suited for? Even though, I'm making this up as an illustrative example, a lot of people actually do this. They put their resume into either Gemini or ChatGPT and ask, "How good is this resume for this job description? Is this resume a good fit?"

These publicly available chatbot tools are not just used by individual workers like you and me. They are also being used by recruiters without fully understanding what the technology is capable of doing well and what it's not capable of doing well. So this is the governance question. How do we come up with rules that determine when does it make sense? And this goes back to my earlier remark about what kind of use case is being considered. If this is really a consequential use case, then the use of the AI technology and the limitation of the technology has strong implications. In this type of CV screening example, a false negative has a strong implication as in it tells the recruiter that (for example) Steve Miller is not a good fit for this role when the actual situation is that you would be a good fit for the role. You should have been included in the interview process, but you were excluded due to the AI screening of the CV. You've borne the cost. The harm that's been caused is you bear the cost of being excluded. Society has borne the cost of you being excluded.

For those false negatives, what are the tolerable levels of false negatives? That's a governance kind of question. If the errors are at a level that are not tolerable, then what? What is not tolerable has to be determined both societally and by that particular public sector agency. That's a governance challenge now.

There are the challenges of dealing with all of the AI model and related product releases. There was the initial ChatGPT (based on GPT 3.5), and then there was the next version of ChatGPT based on GPT 4, and perhaps there's a new GPT 5 coming and there would be yet another new and more capable version of ChatGPT. And there is Gemini with a very large context window. In addition, there is the method of retrieval augmented generation for combining pre-trained language models with an organisation's specific set of documents.

With all of these technology developments, what is a public agency supposed to do? And by the way, this is not just a public sector problem. There's a private sector version of the same thing. Should I wait? Should I create my own version of something? Or should I buy? Will Microsoft in Office 365 (and similarly for AI-based product offerings from other companies) soon make this available six months down the road? Or should I build it in house? This is today's version of the classic make-buy decision-making problem. Should I wait and see what the market and external vendors provide? Or should I start now and make it myself?

Another key question is talent. Do they have the talent in house to build an AI-based proof of concept system? And to make the assessments of the proof of concept? Even if they're going to procure, do they have the capacity in terms of knowledgeable staff to procure? Can they do a responsible AI procurement with their existing talent? Do they have the talent to know the ways in which they are going to test if something meets their criteria?

By the way, having an AI testing sandbox can be handy there as well. Having a sandbox would be helpful to determine if the AI that I'm procuring meets the relevant thresholds. Is it compliant on privacy? Is it compliant on security? The organisation might already have these types of standards in some cases. One of the big concerns is that these standards are still being developed, and in some cases they are just recently emerging. ISO 42001 (Artificial Intelligence Management Systems) is an example of an international standard for AI systems that was recently released in December 2023.[15]

And then you have US federal government programs closely related to standards (but which are not standards themselves) such as FedRAMP which stands for the Federal Risk and Authorization Management Program which defines the process for evaluating private sector cloud services against US federal security and privacy standards which may be higher for the public sector than it is for the private sector.[16]

In summary: First, there are a set of issues complicated on the one hand by the fact that technology is moving rapidly. Second, there are questions of due diligence with regard to how do we govern a technology that's moving rapidly? And third, do we have the talent in house to be able to do proofs of concepts, figure out procurement approaches, and do the required assessments?

We should ask what support the public sector needs in terms of doing procurement of AI-based systems, and in terms of building literacy and education, so that it has that capability.

3.1 A micro-level AI preparedness index for local levels of government would be useful

A public sector organisation needs something like an AI preparedness index and a checklist to help guide procurement. Such an index would be a combination of tech maturity but also talent maturity that is needed and would indicate how prepared a public sector agency is per using AI.

The IMF has an AI Preparedness Index that they use to assess a country's macro level AI readiness by considering the four areas of digital infrastructure, human-capital and labour-market policies, innovation and economic integration, and regulation and ethics.[17]

You could imagine doing this type of AI Preparedness Index at a more micro level so that it would be useful for a public sector agency to assess and determine its degree of preparedness for making use of AI capabilities.

4. Different types of AI governance

4.1 New regulations specific to AI versus use of existing laws and regulations already governing processes or outcomes

I want to come back to the points about the availability of standards related to AI usage and the capacity to decide how to meet a requirement in the current state of AI governance. In settings like the EU, there is the new EU AI law. In the US, there is the President's Executive Order. It is important to keep in mind that to a large extent, in many settings, AI is just a means to an end within some pre-existing area where the government already has laws governing various aspects of processes and/or outcomes.

For example, in the US, related to employment, there's already a lot of existing regulation overseen by the Equal Employment Opportunity Commission (EEOC). Related to the collection, dissemination, and accuracy of information in credit reports, there is already the Fair Credit Reporting Act overseen by both the Consumer Financial Protection Bureau and the Federal Trade Commission. There are pre-existing laws and regulations related to granting loans. These are just a few illustrative examples of the many laws and related governance requirements that already exist.

For these pre-existing laws and regulations, there is no "AI exception." You can't say, "You know what? I was using AI to do loan granting. Therefore, I get an AI pass", as in, you cannot argue that because you made use of an AI model to do a loan assessment, you are not bound by the already pre-existing laws and regulations related to doing this. No, you don't get a special "AI pass". You still have to meet the requirements of what the relevant laws and regulations are. (For example, related to lending, the requirements of the Fair Credit Report Act and Equal Credit Opportunity Act and other laws related to consumer lending).[18]

In settings where there's existing vertical (as in, domain and use case specific) law and regulation, and there are consequential use cases where a public sector organisation is considering using AI, I think those existing laws or regulations will provide the thresholds and

provide the basis for governance including AI governance. For examples, existing laws and regulations will already define the governance aspects in terms of not being able to discriminate based on race or ethnicity, and what will be needed to meet various reporting requirements (e.g., the principal theories of liability based on disparate treatment and disparate impact as discussed in the US federal Equal Credit Opportunity Act (ECOA) as explained in the URL linked to the endnote above). So a lot of existing public sector governance requirements are already in place regarding things like when you grant loans, admissions to schools, and so on and so forth. There are already many of those areas where there are reporting requirements and there is existing law that sets thresholds that would be relevant to evaluating the use of AI (or evaluating the use of any approach to decision making).

4.2 Vertical versus horizontal regulation and different jurisdictional scopes (city, state, national, multi-national)

There are also horizontal requirements and that's where jurisdictions differ. For instance, Europe might have a different set of requirements than the US or other parts of the world, for instance, about the use of very large and highly versatile AI models (e.g., for what they refer to as General Purpose AI models). There may end up being EU specific special requirements in terms of transparency, what data were used, data privacy, and using copyrighted data and information without permission. So there are a set of requirements that have to be met for the type of AI being used, and that area is still in flux.

Some things are complicated just by virtue that any organisation is finding it difficult to figure these things our right now, be it private sector or be it public sector, because the rapid pace of tech change, the inherent slower rate of organisational absorptive capacity, and the various issues of preparedness that I pointed out.

This distinction between vertical regulation within an established sector within a jurisdiction and horizontal regulation across jurisdictions is an important distinction for AI governance.

Then, the third piece has to do with the differences across the different levels of government: city and county level government versus state level government versus federal level government. There are very different levels of resources available- both in terms of money as well as in talent as well as in terms of technology- across these different levels of government.

So, all of this has to come together.

Steven Miller: You gave examples where you discussed the need to do testing of AI capabilities to determine performance, such as the need to determine rates for false positive and false negatives. You also said that in many domains, there are already existing laws, regulations, and rules governing the approach for a specific public sector function, and that these apply whether or not AI is involved, as in, there are certain guardrails there already that have been previously defined by various parts of the public sector.

There are some who may ask, “Why is there a specific need in the public sector for AI related sandboxes and policy experimentation? Why not just govern what we want for the outcomes, such as employment outcomes or housing outcomes or loan outcomes? And get that clear and then just hold the AI systems accountable to that governance for those outcomes.” Related to this, you can use data science and analytics to support this type of decision making in a number of algorithmic ways, with or without so-called AI methods.

4.3 Broadening of the meaning of “what is AI” and implications for AI governance

Ramayya Krishnan: There are two separate issues here. One is organisational, and one is technological. Also, you are getting at the heart of “what is AI?”

Suppose we are doing bus scheduling using a mixed integer programming model. This is a well-established operations research (OR) optimization methodology that has been used for decades in transportation and many other areas of application. And now, even the use of long established OR methods are currently billed (by some people at least) as using “AI”. So AI has become this broad moniker. It is interesting and important to recognise that the use of the term AI has broadened. If you look at the world’s most widely used AI textbook, “Artificial Intelligence: A Modern Approach” (now in its 4th edition released in 2020) written by Stuart Russell and Peter Norvig, they include some language and concepts from operations research.[19]

They talk in terms of “intelligent software agents” with objective functions and constraints, and methods of constraint satisfaction, and they apply these concepts very broadly to describe a wide range of AI methods. In fact, bus scheduling, deep learning prediction models, large language model querying and content generation, retrieval augmented generation, and other areas of AI are all a consequence of some kind of optimization. So the use of software algorithms that are based in one way or another on optimization is indeed very broad.

Steven Miller: My sense is that in these current times, the general public and even many technically literate people who are non-AI specialist broadly associate any use of data analytics and data science as being part of “AI.”

Ramayya Krishnan: That's right. The term AI in common usage has become as broad as that.

I don't think we should regulate the technology because today's algorithms are going to evolve and grow. And what healthcare needs per AI technology and governance is not what autonomous vehicles are going to need.

So context matters, usage matters, and we shouldn't be regulating the technology or the specifics of the technology. New AI methods will continue to be developed and correspondingly lead to AI capability increases. Will deep learning with attention based on the transformer model (which was first announced in 2017 and which dramatically improved the state-of-the-art of language model and chatbot performance) be the way in which large

language models five years or 10 years from now still work? I don't know. At that future time, the best language models and chatbots might be based on something else altogether.

4.4 AI governance for a public sector enforcement agency versus a service delivery agency

There's an organisational question which is implied that relates to the mission, purpose and nature of execution of a particular public sector organisation, or to units within that organisation. Is it the setting of Health and Human Services? Or Transportation? Or the setting of the Securities and Exchange Commission? Or the Consumer Financial Protection Bureau (CFPB)? The primary role of some public sector units is to enforce compliance, whereas other public sector units have an emphasis on providing, enabling or overseeing service delivery. CFPB has a strong emphasis on compliance. If the CFPB determines that a private sector vendor (loan provider) is not granting loans to consumers properly, they may step in to ensure that the loan provider does so appropriately. That's a very particular kind of public sector agency and the capabilities, including AI capabilities, that it needs to work with private sector organisations- that may be better funded than the CFPB- to assess and ensure that there is safe use of AI by a private sector vendor in what clearly is a societally consequential application (loan granting or housing allocation, or other things of that nature) is one type of situation.

A very different kind of situation is the public sector agency responsible providing public transportation, for example running the local bus services or train services. In addition to managing everyday operations, they must make policies on the extent to which routes that are not covering their operational costs are still served in order to provide access to the people who live or work in those areas. AI capabilities can be used to both support everyday service operations as well as to help with the policy decisions related to route coverage. However, false positives and false negatives with respect to AI-based predictions in these types of transportation settings have very different types of implications than in a loan granting or a housing allocation setting.

So, what kind of public sector organisation are you? Are you primarily an enforcement agency? Or are you primarily a service delivery agency? Now, in reality, many public sector organisations are involved in both enforcement and service delivery to varying degrees. Or there may be units within a larger public organisation specialized on the enforcement side versus the service delivery side. Even so, this is a good way of characterising an important difference in the nature of the public sector organisation or its sub-units. Both types of public sector organisations need AI capabilities and the ability to govern those capabilities. However, these two different types of public sector settings require different types of AI capabilities and different approaches to the risk management related to the governance of such capabilities.

5. Examples of AI use cases to support sensitive types of social services decision making at the local government level

Steven Miller: We are now going to bridge into some mini case studies. You've already made references to situations and examples that are very much in the spirit of the UN's sustainable development goals (SDGs) which apply to all countries wherever they are on the development spectrum, because within any country there are always people across all parts of the wide range of the socioeconomic spectrum. You are very familiar with public sector applications of using AI to enhance the ability to give access to people who might not have access for transportation or education or health.

Ramayya Krishnan: Or broadband!

Steven Miller: I think you understand the spirit of the SDGs. The reason I raise this is as a primer to segue into some applications and examples of public sector usage of AI that has the spirit of SDGs.

I want to ask about how some public sector institutions have devised ways to test and work out relevant policies associated with their use of government digital services that are AI enabled, or AI enhanced.

Ramayya Krishnan: And the examples you'd like are examples of government or public sector agencies using AI in ways that serve their communities?

Steven Miller: Yes, and within that, a subset of those types of examples.

5.1 Example #1: Determining who is eligible for a public assistance programme but not enrolled

Ramayya Krishnan: Let me give you a couple of examples and then you can tell me if I'm in in the right direction. There is the eligible but not enrolled problem. In the US, when the federal government has enacted some policy, in many situations, the way in which it actually makes its way down to the citizen is by being routed through the state, which in turns routes it down to the county.

Pittsburgh is in Allegheny County, and Allegheny County has a Department of Human Services.[20]

This county level Department of Human Services is responsible for operating and overseeing a collection of services, some of which are federally funded. Food stamps (the Supplemental Nutrition Assistance Program, SNAP) is an example. Think about the scenario of food stamps for a low-income family that needs supplemental nutrition support. Suppose this county-level Department of Human Services had the ability to determine, or predict with a high degree of confidence, the households that are eligible for receiving food stamps. They could use their food stamp programme participation data to cross-check which of those

households eligible for receiving food stamps (for example, a low-income family consisting of a single-parent with young children) is not enrolled in the food stamps programme. This is a very concrete example of a real and common social needs problem.

The challenge is that the local Department of Human Services did not have a way to determine the households that were eligible for the food stamp programme but not enrolled. They only had the data on which households were already enrolled. So they could not identify the gap of "eligible but not enrolled."

Faculty members at CMU's Heinz College are working with the Allegheny County Department of Human Services to develop a predictive model of who is potentially eligible. Of course, we made use of AI capabilities (machine learning based prediction models) in combination with other statistical and analytic methods to do this. There was another critically important step to this model building effort, and that was to arrive at an understanding of the reason for why the person was eligible but not enrolled. We developed analytic means to understand (to the extent possible) causality.

Why is it that someone who was eligible was not actually enrolled? Was this due to a lack of information availability? Was it the case that the head of this household did not know about this food stamp programme? Or did they know about the programme, but they did not know how to go online and enrol using either an internet connected mobile phone or a computer? Or rather, was it that they did not even have convenient access to the internet?

In order for staff at the Department of Human Services to follow up on the assessments of which households are eligible but not enrolled, they need to know the root causes of this gap, and it helps them in a big way to have an initial best estimate of which root causes apply to which households. The follow on support interventions that were devised had to based on an understanding of what the root cause was. If the head of household was aware of the programme, and had mobile phone internet access, but she or he was not able for whatever reason to deal with the steps involved in going online and getting registered, then the support solution would be to assign a caseworker to visit and sit with the person and to help them navigate the process and get enrolled. A different root cause (e.g., lack of internet access, or a lack of awareness of the existence of the programme) required different types of follow up interventions.

5.1.1 The pilot effort for evaluating the new AI supported approach for determining who is eligible but not enrolled

The whole objective of this type of pilot programme- which is an example of a policy experiment sandbox involving a new use of AI capabilities- was to determine the most cost-effective way of identifying people who are eligible but who are not enrolled. And in addition, to identify to the extent possible (through data-informed estimation) the root causes of why they were not enrolled, and based on that, figure out the feasibly optimal way to do the follow ups.

Steven Miller: That is a great example for a lot of reasons. This issue of people who are eligible but not enrolled, yet need the help of available social support programmes, is a common situation across almost any country setting you can think of. It is also a good example of taking a more proactive approach to providing social services to those most in need.

Can you comment on the nature and extent of the policy experimentation in this example you just described? Also, this example raises a number of issues, such as how proactive should the public service agency be.

Ramayya Krishnan: This entire pilot project was an experiment that was devised by the county-level head of the Department of Human Services as an alternative to the traditional survey-based approach which did not use AI-based prediction models. They knew there was this problem with the data and results they were getting from the prior survey-based approach, but they didn't know the extent of what they were missing, nor did they have a way of identifying the households that were likely to be eligible but not enrolled.

Then once these likely households were identified, the prior approach did not provide any support for understanding the root causes of why that was the case, and therefore, previously, they did not have a basis for planning how they might provide follow up support in terms of an intervention. So having a means to figure that out required this new approach which was only possible because of the improved ability to do the predictions, which in turn was enabled by the use of the AI (Machine Learning based prediction) methods.

This type of pilot is a good approach for the Department of Human Services to learn about useful and responsible, well governed ways of using predictive analytics to support their social service policy, planning and delivery efforts. In this sense, it was an experiment at multiple levels.

5.1.2 The partnership between the local government agency and the university

Steven Miller: What was the nature and extent of the help of the policy experiment in figuring this out? And how was the collaboration with CMU Heinz College funded?

Ramayya Krishnan: It was philanthropically funded. Because it was an innovative experiment, and an important and interesting problem in the social services setting, there was a philanthropic foundation that funded this effort that permitted the public sector agency to work with academia, in this case, CMU's Heinz College.

The pilot could demonstrate that this type of approach could actually be valuable, and it would be feasible to carry it forward to a next stage and do it at larger scale, though of course there would be needs for ongoing operational budgets to sustain such an effort. Beyond the initial pilot effort (which is not yet fully completed) supported by the philanthropic foundation sponsor, follow on deployment for ongoing operational usage has not been done yet, so we don't have results to report on that. But that's the essence of the idea.

Steven Miller: What was their evaluation after they did the pilot, even though they have not moved forward with full production deployment. Do you have a sense of the trajectory of this effort based on the results of the pilot?

Ramayya Krishnan: The idea in the pilot that was evaluated was as follows: if we devised an intervention based along the lines of what I described above, would it actually result in a greater conversion rate of eligible (but not enrolled) to enrolled, as in, would you see a lift in terms of a reduction of the set of people who are eligible but not enrolled. That was what the pilot was about.

The pilot was small in scale because we had to invest a lot of effort in building appropriate, trustworthy and responsible prediction models in the context of this particular setting. So, it was a small scale pilot to both illustrate and test the idea.

There are several challenges to getting this type of initiative implemented. You need to have the data resources to be able to make these predictions about which households are likely households that might be eligible. Arranging for the appropriate technical, administrative and governance means of obtaining privacy protected access to the necessary data sources and getting that coordinated and arranged was a big part of this very first pilot. Another aspect of complexity is the nature of this type of policy issue. These are households that are often disadvantaged, and that can have implications for the extent to which there is complete or missing data in the Department of Human Services databases about these households. So, researchers have to devise practical strategies for working with data sets that had varying degrees of incomplete or imperfect data.

Data resources are the precursor to the use of AI. A public sector unit seeking to use AI as part of their efforts has to be able to determine: do you have the maturity in terms of data and data governance and data resources that would even allow you to do this first step?

Steven Miller: In many situations, more so than needing a highly sophisticated model, the better the data, the more you can do something useful with AI, and the better the result. In most cases, a highly sophisticated AI model cannot compensate for a lack of high quality data.

Ramayya Krishnan: Yes!

Steven Miller: This was a county agency. Earlier, you spoke about the importance of talent and you noted that cities and countries do not have the same degree of resources (talent or funding) for AI projects as state level and national level public sector units.

For this particular experiment, including the planning of the experiment, what was the role of the county officials? Obviously, they know the domain and they are experts on the policy needs and the way these social service programmes operate. What was the role with the university participants?

Ramayya Krishnan: It turns out that this county department is widely regarded in the US as being a leader in the use of technology related to data management and data analytics.

By the way, the Allegheny County Department of Human Services is headed by a CMU alum, a person who's very knowledgeable about the potential of using these kinds of approaches.[21] The leadership is open to new ideas and experimenting with approaches to improve societal outcomes. However, due to their workload demands and staffing constraints, and budget constraints, it was very helpful to them to define a funded pilot project and work with the CMU/Heinz team of faculty and students to tackle this problem. Partnering with the university (while incurring additional project effort overhead) did provide access to relevant faculty and student expertise.

5.1.3 The pilot effort is much more than getting the AI predictive model to work- you need to look at the larger socio-technical system

To run this pilot, and even more so to do any of the follow on to run this at scale on an operational basis, it's much more than getting the predictive model to work. You need a trusted case worker to actually show up at house of the person who was identified as eligible but not enrolled. This has to be the type of case worker that the individual in that household would be willing to talk to. There is a big human component to this type of effort.

This goes to a broader point I want to make. There is too much of an over indexing on AI models. What you need to think about are larger socio-technical systems that encompass the policy problem and domain and AI is just one piece of a much bigger system that one needs to address the identified issues. And this is true in both the private sector as well as the public sector. This is not just a public sector issue.

Steven Miller: What I have observed from my studies of AI applications in a number of public sector applications is that the AI model itself is often a small piece, though a critical piece, but in the scheme of the total domain related use case, workflow and process, a small piece.

Ramayya Krishnan: Exactly! I think that's important to keep in mind as we think about governance related to the use of AI in both the public and private sector. It is not just governance of the AI model. It's the governance of the entire work process and overall administration, technical and institutional systems because it's a system that either produces a benefit or causes harm.

Steven Miller: Right. And by the use of the term “system” here, you do not only mean only the information system. You mean the broader service delivery system, and all the related processes.

Ramayya Krishnan: That's right. It's a combination of human, digital, AI. Everything in a business process that's delivering the service.

Steven Miller: That was a spot-on example. Can you think of another example you're quite familiar with where as part of AI enablement, there had to be some kind of policy sandboxing.

5.2 Example #2: Deciding how to respond to allegations of child abuse

Ramayya Krishnan: So this one is a different part of that same Allegheny County agency, though I see this scenario happening in in other settings as well. This has to do with a child welfare agency. They receive allegations of child abuse that are often called in anonymously. The caseworker who receives this call has to make an assessment of what to do. Upon receipt of this call, does he or she investigate it further? Do they send out a team from the child welfare agency to this household about which an allegation has been made?

5.2.1 How do you carefully deploy AI to support a decision that is so hugely consequential

That decision has two very consequential kinds of errors. One is that I should not have sent the team out there, but I did. That's a false positive, and that effectively hurts the reputation of that household because they get a visit from the child welfare agency.

The false negative is equally consequential, which is that there really is a problem in that household, and you chose not to send somebody, and the child died or was very badly harmed later in time after the child welfare agency had received the call.

This type of decision is hugely consequential. The costs of these errors, of getting it wrong, are very large. This agency does use a recommendation system where there's a recommendation made by the AI to the case worker taking into account features about the household, features as to whether there have been other complaints about this particular household, whether somebody's lost a job there, et cetera.

5.2.2 Concerns with bias in this type of decision making

Often times these are people that might potentially be on (as in, be receiving) social welfare assistance and this is where part of the bias might come in. Households that are poorer, whether they're either poor white or poor black or poor of any background or ethnicity, may have had more child welfare visits to investigate allegations, and therefore there's a concern that the AI recommendation might actually be biassed in favour of sending out a team to those households (because there was a higher proportion of these types of incidents in the training data). If you are wealthy, white and rich, and there was an allegation made about your household, it's less likely that the AI recommendation would be to send a team to visit that household because there isn't as much support in the historical data for this situation. Potentially this is the concern.

Steven Miller: Just to clarify: you use the phrase, “people are concerned about a bias in the data.” Is that a bias in the data in the sense of social discrimination? Or is that just the way the probabilities are, and the record of incidences are?

Ramayya Krishnan: The two are connected in the way that the record is created because as a caseworker, I end up sending more child investigation teams to certain types of homes. In

other words, it (the recommendation based on the training data which is based on data of prior incidents) builds on past work. So this is where this kind of the question of how do you debias these types of recommendations where there might genuinely be the type of issue which I think you're alluding to, which might be that the incidence rate of actual problem occurrences may be higher within certain local socio-economic subsegments. Or is it the case that more teams were sent to certain types of households based on prejudicial (as in biased) assumptions? The historical data on prior visits could be a result of either or both of these types of reasons, right? In statistical methodology, we refer to this as an "identification problem."

So I think the challenge is to recognize the importance of the question of, "How do you deploy AI in something that's so hugely consequential?" Not only consequential to the lives of the kids, which is the most important thing here. But also consequential in other ways as well if you get the decision wrong.

This is an agency that's in the crosshairs of media attention because of the special nature of child welfare. As a society, we feel very strongly about the safety of children and the importance of safeguarding them. This is why these types of decisions are a very challenging situation for both the case worker and also for the agency as a whole. If this type of agency is deemed to have "screwed up," they get into a whole heap of trouble. This often leads many public sector agencies doing worked in "socially sensitive" areas to become very risk averse.

It is a different type of situation than the agency that gives you a driver's licences. They (protecting child welfare and providing drivers licences) are both very important public services but they are very different settings. Both types of public services can make use of AI for making predictions and for recommending a course of follow up action based on the predictions, but I think these are just very different contexts because of the differences in the nature of the consequentiality of the decisions and the consequentiality of errors (false positives, false negatives).

5.2.3 Using sandboxing to understand the nature and implication of false positives and false negatives

This is where I think the sandboxing can be very helpful in order to ensure the public agency in a specific situation understands the nature and implication of false positives and false negatives. In many public sector settings, there are challenges with even defining the nature and implications of a false positive or a false negative. There are many public sector settings where it is a challenge (or not even practically possible) to define what is meant by "ground truth". This is important as a baseline for ground truth is needed to assess the rate of false positives and false negatives. This gets into some of some very thorny issues and sandboxes can be an important way to address such issues given the specifics and practicalities of a particular public sector use case for AI.

Several CMU/Heinz College colleagues of mine have worked on this type of child welfare problem, including efforts to come up with better data analytic and AI tools as well as

related efforts from a human-computer interaction perspective that considers how do you provide this information in a way that a case worker develops trust in the recommendation, or knows when or when not to trust the recommendation. It's a human in-the-loop kind of problem. How do you build trust in using these kinds of systems? At the same time, how do you know when to override the AI systems recommendation, and how many times do you override this? The public agency case workers are allowed to - empowered to - override the recommendation of the AI system. This ability to override the AI system is an important issue from an organisational, process and policy standpoint. The AI system users are allowed to overrule the AI, though they are also required to document why they override it.

6. Finding the best approach in a given use case setting for creating human augmented systems, and in some cases even automated systems

Steven Miller: To ameliorate the problems associated with the risks of using AI systems for public sector decision support or decision making, one could suggest holding back on the use of AI. However, there is the other side of this that as was so well documented by Daniel Kahneman and his co-authors explained in the book called "Noise."[22] The book explains and documents how not using algorithmic decision making in consistent ways in many types of decision-making situations also leads to undesirable and often pernicious outcomes.

Ramayya Krishnan: That's absolutely correct. I think the question is what is the best approach for creating human augmented systems, and in some cases even automated systems, that might do very well.

Here is a simple example where an automated system does very well. Some years ago, when you were driving your car and entered a toll road, you stopped at a toll booth and picked up a ticket from the booth attendant. Then, when you exited the toll road, you stopped at another toll booth and manually paid the attendant. That job of toll both attendant is gone in the US and in many other places in the world. The reason why this job is gone in so many locations is that it can be done automatically with such high reliability.

This is a good example of a digital automation technology used by many public sector transportation agencies. Has this automated approach for charging for toll road usage improved traffic flow? Without a doubt, the answer is yes. That is why this "EZ pass" type of approach for toll roads is so widely used world-wide, though there are situations where some countries or regions do not make use of this type of automated solution due to their local situation.

This is an example where most people would agree that, "Yes, it makes it makes a lot of sense to use digital and information technology to automate this type of task and the related human jobs."

Let's consider another example related to driving that involves the use of technology for safety support though not for full automation, as in, not for Level 5 fully autonomous

driving. There is a lot of AI-enabled safety technology in cars that beep at you when you stray from your lane, or that indicate to you if there is another vehicle in one of your blind spots in the direction that you are starting to move. Is this type of AI-safety support information for the driver a benefit? Absolutely, yes. Both this safety support application and the automated toll both applications are examples of using digital technology, including AI technology, that I don't think anybody would argue about.

6.1 The challenge of training data bias and the example of resume screening

Where the challenge occurs is when we use AI systems in societally consequential settings where the training data used for training the AI models themselves are biased or they incorporate the biases that we as a society have.

Steven Miller: Is it the situation that the data is actually biassed in the sense of reflecting overt prejudice or social discrimination? Or is it the situation that the training data is limited due to whatever practical constraints? Or, do you see both of these situations where in some cases the data is actually biased in a social discrimination type of way and in other situations, the data is limited in that some types of information are under or over represented in the data set with respect to the overall population due to data availability and data collection practices and policies (without there being any intent of overt social discrimination).

Ramayya Krishnan: It could be a combination of these two situations. I would not leave the situation of social discrimination related bias out of it because there are classic, well known examples in the US at least related to university admission decisions and housing allocation decisions where we know that in the past there was discrimination based on ethnicity or race or other social considerations.

Many public and private institutions in the US have come a long way in terms of making sure that they are not going to use any type of decision support information system, including AI based decision support systems and recommendation systems, to try and replicate the prior historical patterns of those types of decisions again at scale.

By the way, there's a really interesting set of published papers co-authored by my CMU/Heinz colleague Alessandro Acquisti that documents the presence of prejudicial discrimination in the CV screen stage of the hiring process.[23] Other researchers, including a team at Stanford Law School, have done recent work on assessing the nature and degree of bias in responses given by Large Language models.[24]

These studies highlight the following type of situation: Suppose you submitted a resume online for a job application with the name Jamal versus the name Keith. All other things being equal, what is the probability of being called back by the company for follow-on discussion per your suitability for the job? Researchers have shown that certain types of names (for example, the name Keith, which is a common white name) have a a higher rate of call back than other types of names (for example, the name Jamal, which is a common African American name).

What is especially interesting about some of these studies is that they also show that once certain types of additional features are included in the resume screening (so not just only using the name, but also including your education background, where you graduated from, job relevant skills and experience information), almost all of that type of disparity effect goes away.

Why is there a lower callback on one versus the other if I just gave you the name? Why is that the case? Well, maybe there are fewer black people with those names that are in the training data that the company used to train the AI model it uses for screening online resume submissions. But once you add these other descriptive features about educational background and job relevant skills, the call back rate based on the initial resume screening starts balancing out.

Is this use of using an AI system to screen online resume submissions an example of discrimination? If you use it inappropriately, it could be.

This is where governance comes in, not only in terms of setting thresholds for acceptance and rejection, but also in terms of coming up with playbooks about how an AI model for this type of application should and should not be used. And for how the use of the AI model actually generates outcomes that are trustable, reliable and fair. This is an important issue.

7. Defining your playbook to use sandboxes as a way of building capability and capacity

7.1 Initial steps and questions for moving forward with a playbook and capability development

Steven Miller: You just gave me two very helpful and specific illustrative examples. In your head, based on your vast experience with public sector usage of AI and policy making, you know of many more examples like these, perhaps tens of additional examples.

Aggregating over this large set of experience you have, what are your thoughts on lessons learned for doing policy experimentation and for creating sandboxes within the public sector as a means to chart the path forward for incorporating AI capabilities, though in ways that are effective, inclusive and accountable. Given what you have been observing and seeing, how would you begin to extract and summarize some of the important lessons learned?

Ramayya Krishnan: One way to respond to your question is not to provide some general answer, but rather to provide a process-oriented answer. Suppose you are a public sector agency. What's your playbook to actually use these sandboxes as a way of building capability and capacity? Suppose you approached it in that way. Then I think the sandboxing effort could apply equally well to building not just technical capability, but also capability of your people. So think of it that way.

Steven Miller: Yes. We view the sandbox as an ecosystem, not just a technically oriented infrastructure environment such as a virtual container on a cloud.

Ramayya Krishnan: Yes. Then I would say a public sector organisation has to define and identify what are a priority set of consequential decision making settings that become good candidates for using AI. You need to start there, by saying what's the use case you want to consider first.

That balance has to be struck between i) what data do I have, where lack of perfection in terms of data availability is not going to be the enemy of the good enough data availability to do something useful, ii) the risk-reward consideration of actually doing this in a way that will determine if there is value from doing this, and iii) because this is in the setting of the public sector , the politics and the optics of a particular use case and how it serves the public interest will determine the way in which AI will be used.

Then, what would the budget be to try and do a proof of concept? If I have a budget for a proof of concept and sandboxes are the mechanism I'm going to use to actually build out a proof of concept to do the testing that needs to be done---then, how do I decide what business process I'm going to use as a use case? And what should be the criteria by which I select that? The criteria could be driven by these some of these dimensions that I just outlined above.

Coming up with those dimensions and then using them consistently is important. Who are the people who come up with those dimensions? Who decides the weights on those dimensions? These are all important factors here. This is not unlike decision making related to doing any type of pilot in any type of organisational setting. Start by picking one or two use cases and then use those few use cases to drive this decision-making process.

Early on in this decision-making process for the pilot, the public sector agency has to decide between doing this effort “in-house” or whether to partner with an external organisation that has more maturity or capability for doing AI prototypes and pilots. This is the classic make-vs-buy decision, and there are choices per the degree to which the agency can work with an external partner. These sourcing considerations for the pilot effort have to take into account the public sector unit’s internal capability for AI projects at a given point in time. If the agency considers doing this internally (choosing “make” over “buy”) their existing ecosystem would have to have the necessary staff to define and implement this pilot effort in the sandbox and to do the evaluations. And that's where the workforce and internal staffing piece of this comes in. Does the public agency have the right capabilities to programme manage and implement the sandbox effort and to know how to manage the effort as a project and a programme?

7.2 Addressing talent related capability limitations and the possibility of accessing “talent-in-the-cloud”

Steven Miller: We already know most government organisations, be they national level, state (or provincial or prefectural) level, county level, or city level are always going to be resource

limited and knowledge limited in their ability to do this. Very seldom would they have all of the necessary staff required.

Ramayya Krishnan: This is where "talent-in-the-cloud" approach for government might work. Imagine something like the following, using a context I am familiar with. Suppose the state of Pennsylvania (as Carnegie Mellon is located in the city of Pittsburgh, the country of Allegheny, and the state of Pennsylvania) creates a pool of skilled people that allows different state-level, county level and city-level agencies to use these skilled people. International consulting firms like PwC and Deloitte are now thinking of using this type of model to support their own consulting teams with their external client engagements because as we all know, AI resources and AI expertise are scarce.

In this "talent-in-the-cloud" model, I can make use of these AI "experts-in-the-cloud" and I can get them for X hours a week. That's what I mean by talent in the cloud. These remote expert AI human resources are members of your team and they support your team during specified time periods. The public agency has staff that knows the use case, the related business process, and all the business and related risk aspects of that process. The public agency staff also know the users (customers) of that process, and their needs and concerns including their risk management issues.

Still, the public agency needs in-depth help on such issues such as: i) what are all of the various important data aspects, ii) what is the AI technology capable of, and what are its limitations, especially in light of available data, and iii) given these considerations, how do I specifically target for how to make use of the AI capability to enhance or expand my existing decision making capabilities? This will lead to consideration of the next level of details pertaining to which exact step(s) in the process do I augment or replace with AI? To what extent?

The public agency has to determine for a given pilot and sandbox effort: what's the right blend of external technical AI expertise with your internal domain, business and process expertise? Because this AI expertise is scarce, including the ability to implement all of these planning and design steps, it will often be necessary for the public agency to pursue the effort using a shared resource model.

7.3 Data inventory, data governance and cloud infrastructure

Steven Miller: Are there other practical and actionable recommendations you would have for countries around the world, including the sub-units of government within those countries, who are moving in this direction are using AI capabilities? What are some of the basic things a government unit would have to have in shape before they can attempt to move forward with doing an AI pilot effort or any type of related sandboxing effort?

Ramayya Krishnan: The first thing you need is an inventory of what data is available for the government unit to work with. You have to break this down into what kind of agency and then that within the agency, what data resources are available? And the specifics of these data

resources. Are they already digitised? What kind of data, as in, is it text, audio, video, images, or numbers or records in a transaction data base? What is its quality? What is the maturity and the capacity of this particular agency or unit to actually work with that data.

So it all starts with data governance. The government unit needs an inventory of the available data that it can work with that is relevant to the use cases it is interested in. You have to start with these data issues.

The second essential piece is the availability of cloud-based Infrastructure. Either the government unit works with one of the major western (US multinational) AI cloud infrastructure players in its home country (Microsoft Azure, Google Cloud Platform, Amazon AWS), or with commercial cloud providers from other regions of the world (e.g. from China, Japan, or from other countries).

Or the public sector in other countries can do something like what India's trying to do domestically, which is a really interesting model of coming up with a distributed cloud-based infrastructure. Building upon and extending the India digital infrastructure stack (the India Stack), India has created the Open Network for Digital Commerce (ONDC) and other supporting infrastructure and application environments.[25] This type of approach could be disruptive for the India domestic market and possibly for other countries which choose to go in this direction.

The intent of recent India AI-focused Cloud Infrastructure initiatives are to create an alternative distributed approach for providing cloud-based access to Graphics Processing Units (GPU) processing cycles required for training and deploying AI models so that domestic users, including government, do not have to be tied to one of the global major providers. There are also other international open source initiatives to provide cloud infrastructure to support AI and high performance computing such as the Open Compute Project.[26]

7.4 Does your country need the capability to develop the largest scale AI models? Or only to deploy and use them?

Another important distinction for a government to consider is whether it will be developer of large-scale AI models and get involved in the infrastructure intensive "pre-training" of such large-scale models? Or whether it will be a deployer of large AI models pre-trained by others (either commercial providers or open-source providers) and the government focuses on adapting that model to its own particular context through fine-tuning training, retrieval augmented generation or other techniques for making large pre-trained models created by others work will in a specific domain and with specific data and information sources. This approach has a much smaller requirement for GPU processing cycles.

With respect to AI models, are you as a country a "deployer country" or are you a "developer country? This is an important distinction to draw. I think most countries are going to be deployer countries. I suspect that relatively few countries are going to be developer countries of very large-scale AI models (the so called "foundation models").

With the emergence of the recent generation of "smaller" language models (still large, but small in comparison to the ultra-large state-of-the-art language models) and open source

models of varying sizes, the government unit needs to have internal capability to understand all these things. When we talk about necessary AI related technical capabilities, it is not just limited to the details of how do you use the existing available tools to do implementation. In addition, you have to have people on your team who are knowledgeable about the state of art of AI.

Steven Miller: Allow me to clarify what you just said. There will only be a small number of countries that are and will continue to be AI model developers, but the context of that statement is with respect to these ultra large AI models (the scale of GPT4, Gemini, Claude, and the large versions of Llama). And you believe that going forward, most countries in the world will end up being AI-model users and appliers of these ultra large-scale models, and naturally, even those smaller number of countries that are also ultra-large scale model developers will also be model users and appliers.

As you shared earlier in this interview, each public sector setting and the specifics of a particular use case context are unique to some extent in that each use case in each setting across each country will have some of its own special considerations and its own relevant data. Everybody's use case is unique in some way, even if there are many commonalities across use cases across countries or across agencies within a country.

Won't each public sector agency have to do some amount of fine-tuning training based on local data related to its particular circumstances? In other words, won't the larger number of countries that are model users (and not ultra-large scale model developers) still have to do some amount of model training for fine tuning, or need to have the infrastructure to support retrieval augmented generation (RAG) to contextualize language model performance to its own set of documents?

It would seem that even for countries who position themselves as model users and appliers of the large scale models (which you believe will be most of the countries in the world), there is still the need for some degree of GPU processing cycles for fine tuning training, even though even though public agencies in that country are not developing and "pre-training" the large scale foundational models.

Ramayya Krishnan: Without a doubt. But your dependence is on that large scale foundation model that got developed in the US or from other sources such as Mistral from France or Falcon from Abu Dhabi. Other than those two well-known non-US international examples, all of the other very large scale pre-trained foundation models are either from the US or from the Chinese large model providers (Alibaba, Tencent, Huawei).

You are right that governments will need to have capability within their own country to do the deployment because their data and their context are your own.

7.5 AI Governance challenges when the software and infrastructure supply chain spans providers in multiple countries

Another important point is that AI governance spans multiple country jurisdictions. If you are a public sector agency in a particular country, and something is not working correctly with your deployment or something goes wrong with it, what obligation does the developer of the large-scale model you are using as a basis for your specific application have to fix your problem?

This raises the issue of the AI supply chain for executing the deployment of the models, and this supply chain can be complicated.

If you are a large public sector agency in a large country, perhaps you have some degree of "market power" in the sense that if your ministry or agency (perhaps in combination with other agencies in your national country) has a large enough market, you may be able to exert pressure on the provider of the large scale AI model and require them to locate the cloud infrastructure used to run the model in your national jurisdiction.

As a hypothetical example just for illustration: Suppose a government agency in Rwanda does its own fine tuning of its own AI model to use in the Rwandan context based on OpenAI's GPT 4 which is sitting in the US on the Microsoft Azure cloud.

In this hypothetical example, that model instance for the Rwandan fine-tuned model sits in the Azure cloud in in the US. Now, again, just for hypothetical purposes to illustrate some of the AI model trans-national supply chain issues, suppose the Rwanda government is using some commonly used commercial application for basic functions such as finance tracking, or HR tracking or procurement tracking. (For example, some local equivalent of a system that has SAP-like capability, or perhaps SAP or some enterprise resource planning system with this type of functionality). And the Rwanda government agency's fine-tuned version of GPT 4 for the Rwanda context is using HR and finance data from this locally used enterprise resource management system in order to generate predictions and recommendations. But something with this AI model in the Rwandan context is not working properly. What's the obligation of the local enterprise system vendor who installed this HR and Finance tracking systems that is being used in conjunction with my GPT 4 based AI model?

They may say the problem is not with their system but with the fine-tuned GPT 4 model that's sitting in the US. The point is you now have this AI model governance issue that spans several national jurisdictional boundaries. Once you get into using a very large language model that is only hosted in a limited number of countries (and not in your particular country), you get into this type of "AI model supply chain" issue.

Before the use of these very large foundation models, a government or commercial sector company built much smaller scale AI models using the commonly used available AI methods from prior years such as decision trees or support vector machines or deep learning systems that had a limited number of layers. This much smaller scale model was run "on-premise". It was sitting there on your own organisation's IT infrastructure, and everything was sitting within your jurisdiction. Now, given the size and nature of this new generation of very

large-scale foundation models that can be used for a wide range of tasks, and also fine-tuned with our own organisation's custom data and documents, they can only be run on very large cloud infrastructures. However, a country always has the option of using the smaller versions of open source large language models that can be run locally or domestically, though they may not have state-of-the-art performance.

Now if the government entity is very large like the EU, you would have the market power to force all of the vendors you relay on for your data management and your AI models (in this case, the enterprise systems provider with the HR and Finance application) and GPT 4 to locate the relevant applications and model instances in the EU. That way, everything related to model development to deployment and usage, and integration with other data sources and applications, all sits under the EU's jurisdiction from a governance standpoint.

However, if you are a smaller country, it is much more likely that your AI model supply chain will span multiple jurisdictions, and this can have some important implications on a number of fronts--- including your country's choice of what size model to use given the considerations of where that model can be hosted for deployment.

8. Concluding recommendations and suggestions

8.1 Recommended steps and questions for moving forward with playbook and capability development for AI

Summarising some of my comments above to help guide public sector officials considering or already pursuing AI usage in public sector settings:

- Do you have people that know your process really well? That's a starting point.
- Do you have an inventory of your data that is available, and of the properties and characteristics of this data, including whether it is digital already?
- Do you have an inventory of AI use cases? Are you capable of planning, implementing, and evaluating the pilot with your own staff? And doing all of this with computing resources within your own country? If not, what type of partnering strategy do you need?
- What are your governance policies per what must be done (or hosted) in country versus can be done (or hosted) with resources out-of-country?
- Do you want to build your own model for deployment and if you want to build your own, do you want to use open-source tools or proprietary vendor tools?

8.2 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

Here is another idea for an organisation like the UN to consider. Think of what Europe has done in terms of creating shared resources for physics research through CERN. The United States has just announced that it will provide a National AI Resource so that universities and

other entities aside from the very largest commercial providers of very large proprietary AI models can have access to the necessary compute resources and supporting infrastructure for doing AI R&D that will benefit science, research, as well as the national economy.[27]

Should there be the equivalent of a global "CERN-like" entity that can provide smaller countries and less developed countries some reasonable amount of cloud-based GPU access for civil sector (non-defence related, non-internal security related) AI model testbedding?

If you use an open source model that's already been "pre-trained" with weights, using that for further fine tuning and for model deployment is not as computationally expensive, but you still need compute. So for smaller less developed countries and for the various other categories of less developed countries, where is your compute for AI coming from? Do you have that cloud -based computation source? Where is that going to come from? Could something like this be funded by a global "CERN-like network" that less developed countries can get access to? How could this be paid for?

8.3 Key steps for getting countries at different levels of AI capability and maturity to experiment with AI technologies in beneficial ways

Steven Miller: You've been so generous with such on-target observations and sharing. Is there anything else that you would like to just say or wrap up or to synthesise?

Ramayya Krishnan: I would start by asking ourselves what is the problem we are trying to solve in order to help this UN effort on policy experimentation and sandboxing in the context of using AI for public sector applications.

If the core of the problem that we're trying to solve for is how did we get countries at a different levels of AI capability and maturity to experiment with and try out these AI technologies in ways that might benefit their citizens and their society ---- if that's the big overarching goal, then I would say:

- Make sure you prioritise the process that want to do a proof of concept with.
- Do an inventory of data to see if the available data you have enables you to work on the process and related use case that you initially prioritized. In other words, make sure the process you prioritize is a good candidate for an AI pilot given the data you have available.

8.4 Suppose international organisations could help enable access to the "digital public goods" that less developed countries need to progress with using AI?

For international organisations (including the UN), consider making efforts to identify the "digital public goods" that are needed so that less developed countries can have access to the assets they need to make progress with using AI.

- What data-related digital public goods that are needed?
- What cloud computing infrastructure public goods are needed?

- And how to address the severe talent shortage for providing these developing country public sector AI projects with the advisory and consulting expertise that they will need?

Is there a way to provide access to talent, or at least a supplement to access to talent as a digital public good? For example, through some type of international NGO focused mechanism that can serve as a platform for organising experts who want to "donate" their professional expertise in this way—in the spirit of Doctors Without Borders. This could also be a vehicle for master's students and PhD students with the appropriate types of expertise to do "service stints" on public sector projects (perhaps in their own country, or with other countries) as part of their education program training. Imagine an "AI Engineers without Borders."

Imagine various developing companies using the India Stack, or making use of parts of the India Stack to create the foundation layers for their own public infrastructure and then add what is required for supporting the deployment of AI models:

Data, infrastructure and talent are the essential elements for creating sandboxes for policy experimentation to do trials with AI models. Then, as part of the evaluation, the public sector team (with whatever external staffing support) could compare the results of delivering some specific public sector use case the AI way versus the prior way and see if this made a discernible difference. There may be the opportunity to do more formal A/B testing, or even a well-designed Randomized Control Trial (RCT) to determine if the use of an AI model as part of a government digital service actually makes a difference.

Steven Miller: On behalf of the UN's Division of Public Institutions and Digital Government and its parent organisation DESA (Department of Economic and Social Affairs), I would like to thank you.

Ramayya Krishnan: Steve, one thing that strikes me that might be really interesting is following up on the summing up that I just did. I'm wondering whether there might be an opportunity to do some partnering around this and create a white paper of some sort, and/or or a practical playbook for government agencies related to AI experimentation and sandboxing.

Endnotes

[1] For background on Prof Ramayya Krishnan, visit https://www.heinz.cmu.edu/faculty-research/profiles/krishnan-ramayya.

[2] The Heinz College is named in honour of Pennsylvania Senator H. John Heinz III who died in an aviation accident in 1991. For more information on CMU Heinz College, visit https://www.heinz.cmu.edu/. For more information on CMU/Heinz Block Center for Technology and Society, visit https://www.cmu.edu/block-center/.

[3] For more information on CMU's Metro 21 Smart Cities Institute, visit https://www.cmu.edu/metro21/.

[4] For more information on CMU's Traffic 21, visit https://www.cmu.edu/traffic21/.

[5] *CMU Dashboard Will Help Inform State Decision-Makers During Pandemic*, 22 April 2020, https://www.cmu.edu/news/stories/archives/2020/april/dashboard-will-help-inform-state-decision-makers.html.

[6] US General Accounting Office, Educators Advisory Panel Members, as of September 2023 www.gao.gov/assets/2024-01/EAP-Members_Sept2023.pdf.

[7] US AI.Gov, National AI Advisory Committee, https://ai.gov/naiac/.

[8] US AI.Gov, National Artificial Intelligence Advisory Committee (NAIAC) Year 1 report, May 2023, https://www.ai.gov/wp-content/uploads/2023/05/NAIAC-Report-Year1.pdf. See page 26. For the objective "Organise and elevate AI leadership in federal agencies, "the fifth recommendation item states, "Agencies should foster a culture of continuous piloting and experimentation, mindful of the multi-stakeholder and sociotechnical considerations addressed in this NAIAC report. An evaluation process should include testing of AI systems for safety and functionality, assessment of impact on stakeholder groups, and processes for reporting, mitigation, and redress of harms should harms occur."

[9] GOV.UK, Department for Science, Innovation & Technology and Office for Artificial Intelligence, 03 August 2023, A pro-innovation approach to AI regulation, https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper. Download page for document: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach.

[10] Think Tank, European Parliament, 11 March 2024, Artificial Intelligence act briefing, https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792. Briefing download: https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf.

[11] The Council of the EU made the final approval of the EU's AI Act on 21 May 2024. https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/. The AI Act was published in the official journal of the European Union on 24 July 2024, and went into force on 01 August 2024. https://www.artificial-intelligence-act.com/.

[12] Commonwealth of Pennsylvania Governor's Office, 20 September 2023, Executive Order– Expanding and Governing the Use of Generative Artificial Intelligence Technologies Within the Commonwealth of Pennsylvania, https://www.governor.pa.gov/wp-content/uploads/2023/09/20230920_EO-2023-19_AI_Final_Executed.pdf.

[13] Carnegie Mellon University, 27 September 2023, Gov. Shapiro Visits CMU — Birthplace of AI — To Sign Executive Order on Generative AI, https://www.cmu.edu/news/stories/archives/2023/september/gov-shapiro-visits-cmu-birthplace-of-ai-to-sign-executive-order-on-generative-ai.

[14] Commonwealth of Pennsylvania, Governor's website, 09 January 2024, Shapiro Administration and OpenAI Launch First-in-the-Nation Generative AI Pilot for Commonwealth Employees, https://www.pa.gov/governor/newsroom/2024-press-releases/shapiro-administration-and-openai-launch-first-in-the-nation-gen.html. The Pennsylvania Governor's Office released a summary report on the results of this 12 month pilot effort on 21 March 2025, https://www.pa.gov/governor/newsroom/2025-press-releases/-shapiro-administration-leads-the-way-in-ethical-use-of-ai.html#.

[15] ISO, December 2023, ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system, https://www.iso.org/standard/81230.html.

[16] FedRAMP establishes a standardized way for US federal agencies to assess the security of cloud services offered by private companies. The program aims to ensure these cloud services meet strict security requirements for protecting government data by providing a framework for evaluating cloud services against those standards. The FedRAMP program is the authoritative standardized approach used by various parts of the US federal government to do security assessment and authorization for cloud computing products and services that process unclassified federal information. US Federal Risk and Authorization Management Program (FedRAMP), https://www.fedramp.gov/program-basics/.

[17] Kristalina Georgieva, International Monetary Fund (IMF) blog, 14 January 2024, AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity, https://www.imf.org/en/Blogs/Articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity. From the IMF blog above: Singapore, the United States and Denmark posted the highest scores on the index, based on their strong results in all four categories tracked.

[18] US Consumer Protection Financial Bureau, CFPB Laws and Regulations, Equal Credit Opportunity Act (ECOA), June 2013, https://files.consumerfinance.gov/f/201306_cfpb_laws-and-regulations_ecoa-combined-june-2013.pdf.

[19] University of California Berkeley, Computer Science Department, Artificial Intelligence: A modern approach, 4th edition website, https://aima.cs.berkeley.edu. Pearson publishing, Artificial Intelligence: A Modern Approach, 4th edition, April 2020, https://www.pearson.com/en-us/subject-catalog/p/artificial-intelligence-a-modern-approach/P200000003500/9780134610993?tab=table-of-contents.

[20] Allegheny County Pennsylvania, Department of Human Services (DHS), https://www.alleghenycounty.us/Services/Human-Services-DHS.

[21] Allegheny County, DHS Data Warehouse, https://www.alleghenycounty.us/Services/Human-Services-DHS/DHS-News-and-Events/Accomplishments-and-Innovations/DHS-Data-Warehouse. Allegheny County Analytics website featuring analytics projects done by the Department of Human Services and many other county agencies: https://www.alleghenycountyanalytics.us/. Note that Allegheny County has an Office of Analytics, Technology and Planning (ATP), whose mission is to support policy development, quality improvement, planning and decision-making through research, analysis and engagement. ATP is an office within Allegheny County's Department of Human Services (DHS). https://www.alleghenycounty.us/Services/Human-Services-DHS/DHS-Offices/Analytics-Technology-and-Planning

[22] D Kahneman, O Sibony, C Sunstein, May 2021, Noise: a flaw in human judgement, https://www.hachettebookgroup.com/titles/daniel-kahneman/noise/9780316451383/?lens=little-brown.

[23]A Acquisti and CM Fong, *Management Science*, *66* (*3*), 2020, An Experiment in Hiring Discrimination via Online Social Networks, and Online Appendix link.

[24] A Haim, A Salinas, J Nyarko, ArXiv working paper, 21 February 2024, What's in a Name? Auditing Large Language Models for Race and Gender Bias, https://arxiv.org/abs/2402.14875. J Nyarko, Stanford Law School Blog, 19 March 2024, SLS's Julian Nyarko on Why Large Language Models Like ChatGPT Treat Black- and White-Sounding Names Differently, https://law.stanford.edu/2024/03/19/slss-julian-nyarko-on-why-large-language-models-like-chatgpt-treat-black-and-white-sounding-names-differently/.

[25] India Stack, https://indiastack.org/index.html. Open Network for Digital Commerce (ONDC), https://ondc.org/. S Sahasranamam & J Prabhu, Stanford Social Innovation Review website, Digital Public Infrastructure for the Developing World, 25 March 2024, https://ssir.org/articles/entry/digital-public-infrastructure-developing-world.

[26] For more information on the Open Compute Project, visit https://www.opencompute.org/.

[27] The National AI Resource Pilot in the US stated in January 2024. visit https://new.nsf.gov/focus-areas/artificial-intelligence/nairr# and https://nairrpilot.org/.

INTERVIEW 6: Prof Gianluca Misuraca, Polytechnic University of Madrid, Spain and EU AI4Gov Master Programme

Date of Interview: May 02, 2024

1. Introduction to Gianluca Misuraca's involvement in public sector AI and the AI4Gov International Master in AI for Public Service

Steven Miller: I want to thank you on behalf of the UN team at DESA and the Divion of Public Institutions and Digital Government for participating in this interview. Please introduce yourself in your own words and state your title and roles that you think are appropriate to this interview.

Gianluca Misuraca: I'm Gianluca Misuraca.[1] I'm the founder and Vice president of Inspiring Futures, a global consultancy network, headquartered in Switzerland and in Spain with operations in different countries. We are mainly working with international organisations such as the United Nations (including UNESCO and UN DESA), and with the Council of Europe and especially the European Commission, as I previously worked for EU institutions for many years.

I've been leading research on digital governance and public sector innovation, and lately artificial intelligence in the public sector, but also social innovation and specific aspects of digital inclusion at the Joint Research Centre of the European Commission which is a policy research institute with a specific focus on foresight studies.

My passion is trying to anticipate the future, and that's what I've been doing as part of my research when I was In Switzerland at Ecole Polytechnic Federal de Lausanne (EPFL) and in some of my other capacities.

Now in fact I'm also "back to school." Since I left the European Commission in 2020, I've been serving as Executive Director of the AI for Government (AI4Gov) master programme that is a project co-founded by European Union focused on AI for public services.[2]

Steven Miller: Who are the students in this AI4Gov master programme? Could they come from any walk of life and they want to understand AI for public sector in order to move into public sector job roles involving AI applications? Or, are many of your students already public sector professional people who do this part time during their regular job?

Gianluca Misuraca: You need to already have a Bachelor to access the Master, but this type of Master is what we call in Europe a first level Master so you don't have to be have already graduated from another master to enrol.

It turned out to be more of an executive master type in that many of the people who have enrolled to date are working full time. We even have people who have already completed a PhD doing our Master. We are now completing the 3rd edition (3rd round of delivery) and we will have a fourth edition next year. These initial years of the master programme are

supposed to be a pilot for a larger scale roll out through the EU's Digital Europe programme and that piloting and preparing for scaling is what we are now trying to do.[3]

When we started, we had to design a new curriculum for a student profile that did not exist yet. In fact, we found a lot of interest because the programme is trying to equip both the technical experts with the non-technical capacities and vice versa. That's where we've been quite successful. In the three editions to date, we trained 120 participants from all over the world, from about 50 countries. The master is mainly online, and we have two events in person, one in Madrid and one in Milan because the lead institutions are the Polytechnic University Madrid and Polytechnic University Milan. We also have other university partners in Estonia and Germany.

We found the conjunction of different disciplines, different profiles and different interests to be fundamental. This is both from the student side where we have participants coming from various governmental institutions, from the European Commission, from various public sector organisations, and we also have students who are consultants, researchers and civil society professionals. On the lecturer side, we have the combination of faculty experts from data science, from AI, and from the field of design (User Experience design, socio-technical systems design, organisational design, process design). The programme also includes courses on public administration and talks and visits by guests who are actually in charge of policies such as some members of the European Parliament that were drafting the AI Act.

Steven Miller: How do your professional activities give you up close visibility to public sector organisations planning for, reviewing or using new digital capabilities that include AI? In other words, outside of the classroom, could you explain some of the things that give you visibility to real-world efforts?

Gianluca Misuraca: I've been working for most of my career in the area at the crossroads of policy, research, public sector reforms and technology going way back to the time before this was even called e-government. For a long time, I have been involved with the transformational aspect of technology in public sector, going back to earlier innovation efforts before this current generation of digital technology, to efforts transforming with the more recent technologies, and now to all that is happening with AI.

AI arrived more recently. Perhaps it is better to say that AI was "rediscovered" by policymakers more recently as AI is not at all new as even 20 years ago and further back, we had been discussing it and asking policy makers to prepare for how this will eventually change our society, the way we work and we live, and even the way we think. But a few decades ago, AI capabilities were not yet as well developed to the extent that they now are, and planning for the eventually widespread usage of AI was not taken up at that time as a priority.

Now of course, suddenly our government officials and policy makers realise that we in Europe were lagging behind with respect to AI technology and usage compared to the US and China. That realization has led to all this rush across Europe with trying to equip everybody with this technology and the capacity. Still, we are just at the beginning.

When I was previously working at the EU's Joint Reserve Centre through September 2020, we were looking at what was happening with AI, going beyond the surface, beyond the claims in the media and the rhetoric. So indeed, I've been advising and working with governments about using technology, data and AI in the public sector for decades, though mainly in Europe. And more recently, also worldwide. For example, last week I was in Rwanda, and I just concluded work with Jordan.

There is an international interest in understanding how to use AI in public sector. Yet, in many countries, they first need to really understand the basics. During some of my recent public sector international visits, everyone's talking about AI and then you say, "OK, but where are your data? Do you have the data you need? Can you exchange and combine the data you need? And what about the basics of security? And the data privacy aspects?" You really need to make sure that when someone is talking about artificial intelligence, they know what they are talking about. Some of the public sector people I speak with know they don't actually know about AI but there are here to discuss and to understand.

This relates to why our AI4Gov Master has been so successful. We won the European Digital Skills award last year exactly because we are giving this very basic knowledge. And then of course, those who want to can go into more depth to better understand what AI is and what is actually involved in using in within the public sector.

I am now also working with OECD and with UNESCO. People are really hungry for this type of information on AI in public sector settings. They really want to learn more. However, you have to start with having a use case because the risk is that you embark on a journey, and you don't even know where you want to go and you spend a lot of money buying systems that maybe were not even necessary.

2. What AI related policy experimentation and sandboxing means to me

Steven Miller: This interview is to discuss the theme of AI related policy experimentation and sandboxing, specifically in public sector. When you hear the term AI related policy experimentation and AI related sandboxing, what does that mean to you?

Gianluca Misuraca: For me, this means a lot. This is really needed because otherwise we might end up just using technologies to do things that we were already doing in the past though maybe with a bit of getting them performing better but not really changing the relevant system. That bigger change is what is required especially if you want to modernise or transform the procedures or policy making processes within the public sector.

To me, it also means engaging with "the usual suspects" within government who are involved with public sector technology and the related process change initiatives, and also with citizens, and also with experts that are outside of the Parliament or the government.

It means having this capacity to extend our collective intelligence and to change the way we normally do policy making. At this moment in time, it is still a very old-fashioned way

that we go about policy making. It also means having the consultations and listening to citizens and understanding the range of the sentiments and the opinions and then really taking them into consideration in some data informed or evidence informed process. That's why experimentation in this realm is needed though, it is not yet happening very much, not to the extent that it needs to.

Last year I was involved in some hearings for the Italian Parliament and resulting from this process, they recently launched some proposal calls for experiments that would involve reaching out to citizens and experts and also involve experimenting with using generative AI to see how this can actually be done. So, there is room and opportunity for experimenting in this area. Clearly the current way we do policymaking is not adequate anymore for the current times and these limitations are showing as the tension between citizens and governments and policymakers and this is coming to the fore in many countries.

I think this topic of policy experimentation and sandboxing related to AI usage is important and is also needed as part of reinvigorating our democratic systems that are now all being put under question. Of course, there are risks that the use of AI, including AI related policy experimentation, could be used instead for reinforcing certain non-democratic governance systems.

Coming to the term "sandbox"- I am not a native English native speaker, so I naturally ask, "what is a sandbox?" And then you basically realise that a sandbox is where kids and children play, and they experiment with how they play in the sand. I had the realization, "Ah Ha, so that's what it means." There are some people who think the term "sandbox" is a very technical term. No. I do not see it as a technical term. It's about creating a setting where people can come together and "experiment" as in, play, though with supervision, and try out something new.

Now everybody's talking about developing policies for how AI is used, and I think it's important that we have these controlled "sandbox" settings for experimenting as we do this. Clearly, policy experimentation sandboxes overall, as well as more specific regulatory sandboxes, are important. I think it's a good that there is this approach, and that is very much linked in part to the experimentation. **It is especially important that there is also the possibility to experiment and fail in a controlled environment.**

Steven Miller: That's a nice expression you just said, "the possibility to fail in a controlled environment."

You mentioned the need to make changes within the public sector, the need to work with new technology, the need to modernise how certain approaches are done. Obviously, these types of changes would need trials and pilots and experimentation. When you make use of a new model, a new method, or any new kind of machinery, you have to do trials to get it to work in the real-world operational setting.

2.1 Learning to better integrate across policy making, service delivery and regulation through experimentation and sandboxing that harnesses the potential of new technology

Steven Miller: In this project, the UN team uses the word "*policy* experimentation" and "*policy* sandboxes." Does adding the term "*policy*" in front of public sector experimentation or sandboxing mean anything special? What does that mean in your mind?

Gianluca Misuraca: I think it makes a lot of sense and also has as an important meaning because it brings the discussion to a higher level. For the government and the state, the main civilian sector functions are related to service delivery, the regulatory aspects, and the policy making, and sometimes these functions are treated as separate. The members of the Parliament do the laws with the experts. Then based on these laws, some bureaucrats are trying to simplify the service delivery using sub-technologies. But often you miss the policy which is the architecture of all this. Then of course you have all the strategies that are done by the different levels.

But what is really the vision that you want to pursue or enforce? And how do you use either "carrot" or "stick", to drive movement toward that vision? How to devise some incentive to those that perform well? How do you get the various parts of public sector to cooperate with other stakeholders, and in particular the private sector and the citizens? This often requires a policy change. Of course, that is easy to say and not necessarily easy to do because it also requires a willingness to make such changes and commitments to do so, and this is something that is not always there because policymakers are in general (most of the time) trying *not* to transform too much for many reasons, maybe also for good reasons, because we cannot reinvent the government every day.

But it is true that in some cases you really can use technologies in a way to do things maybe completely new and much better that were not possible to do without technology. That is why we have to really harness this potential, especially given the fact that for the first time in history, we now have a huge availability of data and the possibility to gather and make sense of this data that we didn't have before. So harnessing this potential to improve policy making together with improving the other aspects of public functioning- that's clearly where we should be, and we therefore need to experiment with policy. So I think it's a good choice to emphasise policy experimentation.

Steven Miller: One thing I take away from what you just said is that experimentation, especially in the public sector setting, 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. How do I use the capabilities of the technology to help me learn how to bring these facets of the public sector together, and to do the execution necessary to have these facets work together?

2.2 The importance of conceptual reframing as part of policy experimentation and related sandboxing efforts that involve using AI

Gianluca Misuraca: I was recently part of a project exploring digital government transformation and trying to understand what is new in public sector innovation, along with how public sector innovation should be done in a data-driven society. We came up with a term and concept that we thought was something new. This was the idea of "reframing" which involved re-conceptualizing, re-defining as well as redesigning. We viewed reframing as even including the cognitive aspect, the mindset, and the shift in cultural behaviours that need to be embedded into a public sector innovation process.

Then I started as a visiting faculty member in the Department of Design of the Polytechnic Milano related to my role as head of the Master on AI4Gov. My colleagues from that design department had already been using that same term reframing and they explained to me that the concept has been used for a long time in the areas of system thinking and systems design theory and in other disciplines as well. The re-conceptualizing that is part of reframing is exactly what is needed when it comes to making use of artificial intelligence. Before you even apply AI or any other technology, you need to see what is it for, and what data you have at your disposal. So as part of examining this, you maybe need to redesign the services and the processes. In some cases, maybe you don't even need that technology. That's a very important concept that I think embeds what you summarised.

3. The EU's AI Act and its implications for policy experimentation and sandboxing

Steven Miller: Let's now transition into a few specific examples where you can elaborate on these themes that we've described in our introduction, but in some specific context.

3.1 The EU and member states are initiating experimentation and sandboxing efforts to test compliance with the new AI Act

Gianluca Misuraca: When I left the EU Joint Research Commission, I was asked by colleagues at UN DESA to look at how digital government can help with Sustainable Development Goals (SDGs) along with the cross-cutting issue of the impact of technology. This was one recent step towards investigating examples.

I have also been following various AI sandboxing efforts. In Europe when you say the word "sandbox" the usual assumption is that it must be a regulatory sandbox.

Steven Miller: A sandbox for public sector experimentation does not necessarily have to be specific to only experimenting with regulatory matters.

Gianluca Misuraca: That's what I was pointing out, though now in Europe, in the EU Commission, their focus with respect AI sandboxes is on regulatory, and even more specifically

on the AI Act. Now the emphasis in the EU per AI related policy experimentation and sandboxing is an attempt to test and then validate if the regulatory framework that the EU has been put in place is actually viable and makes sense, and what are the implications.

One of the most known proposals for an AI regulatory sandbox effort is in Spain because of the launch of an initiative with the Spanish Government where the Presidency of the EU Council (where Spain held the EU Presidency from June through Dec 2023 as part of the rotating Trio Presidency approach) and the Commission are trying to test the impact and explore the implications of the AI act on the ethical use of AI. This effort in Spain is one specific EU trial effort.

Now (29 Feb through 29 May 2024) there is a proposal call from the European Commission's Digital Europe Programme trying to foster and promote this AI experimentation and sandboxing approach in the context of the AI Act in other countries.[4]

Also, non-governmental industry trade groups have been involved with AI sandboxing efforts as part of gauging the impacts of the AI Act on start-ups and SMEs.[5]

I think these types of sandboxing and related exploratory efforts in both the public and private sector are needed because so far, despite all the debate about AI, we do not yet have clear guidelines on how to apply the law and on whether the impacts of the law will be positive or not. We do not yet understand the nature of the obstacles and the problems that will result from applying the law.

3.1.1 Experimentation and sandboxing for AI applications can be done for other purposes, not just for AI Act compliance testing

There may be a specific focus of how AI related sandboxes are currently being pursued in the EU, but indeed I agree with you that this is not the only way we can make use of sandboxes for AI, and there are many other ways that you can use a sandbox to experiment with innovating. There is a huge potential to use sandboxes to do experimentation in specific areas linked to the SDGs, for example for education, healthcare, and for various social and employment policies, as well as for environmental aspects. This includes experimenting with the use of AI to support the twin and interconnected transitions of digital/technology and environmental/climate change.

3.2 Avoiding getting stuck in the syndrome of never moving beyond piloting

Across Europe, I have seen a lot of attempts at piloting the use of AI in the public sector but in many of these cases, we seem to have this syndrome of "ever piloting" which means most of these efforts do not progress beyond the pilot stage and never make it to large scale operational usage. That's why I believe that bringing this to a higher level of policy experimentation probably will force or must force the entire government to then be more serious about it. The risk of this "every piloting" syndrome is that you have this pilot done in a

specific department in a specific ministry or in a specific city or region, but not scale it up or scale it out.

It's important that when a public sector AI application is done in one EU city such as Amsterdam or Helsinki that the successful aspects can be replicated, and the lessons learned regarding challenges or even failure can be applied in other European locations such as Palermo or Athens. In the EU, there are at least some attempts to do that, but I think we can definitely do more and maybe also learn more from what's happening on the other side of the oceans.

3.3 The rapid increase in public sector AI applications across the EU since 2020

Steven Miller: Are there specific case studies that you have been briefed on or you had to learn about where a government entity is taking a regulatory issue or service delivery issue or policy implementation and adding an improved ability that incorporates AI ability?

Gianluca Misuraca: In 2020 I was leading the EU's AI Watch for the Public Sector under the AI Watch Observatory. There was very little research at the time. But now research and case studies on AI in the Public Sector across the EU has exploded in the last five years.[6] The number of papers that we have on AI in the public sector has skyrocketed in recent years.

Back in 2020, one of our contributions was to show that there was not much evidence about the use of AI in the public sector across the EU. So, we gathered data by surveying the EU Member States and we found that there was a lot of interesting initiatives. Serendipitously, I found out at that time that in the US, some of our colleagues were doing something very similar in terms of data gathering and investigation. We shared our respective results and views, and we came to similar conclusions that there was a lot of interest, and that many people across the public sector were aware of the potential and even aware of the risk although to different degrees.

At that time, most of use of AI technology in the public sector (civilian sector) was still superficial. There were some machine learning application efforts and uses of more traditional types of AI and other types of digital technology. My assessment per what was happening at that time was that most EU public sector users of AI were not leveraging on the potential of combining different technologies. That is the beauty of artificial intelligence. You can gather data through sensors and in other ways, then make sense of that data, and build models with that data to anticipate, to predict and to simulate. But we did not encounter very many examples of public sector organisations doing all of these steps. Of course, you need to make sure that the data are not biased and address other aspects of the responsible usage of AI.

Back in 2020, we could see that this is an emerging national field and there were a lot of pilots and experimentation being done to produce some results. And hopefully these results could be replicated, transferred and exchanged.

3.4 The 2021 study highlighting several problematic examples of algorithmic or AI-based decision making in the public sector

After I left the EU Commission Joint Research Centre, an independent foundation asked me to do something that I thought was puzzling. They asked me to investigate examples of the "bad use" of AI technology in the public sector. There was a highly publicised case in the Netherlands where a machine learning algorithm was used as part of a government effort to prevent and combat fraud in the fields of social security and income-related schemes, including child welfare schemes, and tax and social insurance contributions.[7] However effectively the system was at detecting true cases of fraud, there were some high impact false positive examples where the government had sent notifications to mothers who had received child welfare benefits that they had to pay back the government for benefits received even though they in fact were qualified to receive these benefits. Thousands of families, often with lower incomes or belonging to ethnic minorities or with migrant backgrounds, were flagged by the system as being high risk fraud offenders and were sent notifications by the government and/or investigated. There was insufficient quality control of the system outputs for false positives and there was also a range of issues related to the privacy protection.

There was also the controversy in the UK (England) about using statistical algorithms for predicting the results of graduating high school students during the Covid pandemic when it was not possible for students to sit in person for their A-level exam. Because of the cancelled A-level exams, the government had teachers give their estimate of how they thought their students would have performed on the exams. Those predicted grades were then adjusted by England's Office of Qualifications and Examinations Regulation (Ofqual) using statistical algorithms (without any explicit mention of AI methods, so not necessarily AI methods) that weighted the scores based on the historic performance of individual secondary schools, teaching ranking of individual students, and a few other factors.[8] Because of the nature of the algorithm and the historic data used, students with high grades from less-advantaged schools were more likely to have their scores downgraded, while students from richer schools were more likely to have their scores raised, resulting in around 40% of the predicted performances being downgraded, with only 2% of marks increased.[9]

There were several other of these types of "bad examples" resulting from the use of automated, algorithmic (usually but not always AI-based algorithmic) decision making. The resulted in the report "Governing algorithms: perils and powers of AI in the public sector" published in 2021, that examined five AI public sector use case studies from several European countries that raised concern due to the considerable public backlash that emerged following their adoption.[10]

3.5 Sandboxing as a more controlled and careful way to learn about the issues of complying across a range of public sector settings

In my opinion, the area for AI applications that have the most potential, but also have high risk, is when you have to allocate social benefits and when you provide opportunities for

employment. You can design and validate an algorithm to make the best result or to optimise the result. However, the results, even if they apply decision making approaches more consistently and are carefully controlled for biases in data, may well exclude some people from benefits or from job opportunities though also help the public agency to know which other people to focus on. However, this has been challenged in some settings as being not constitutional and not fair. There are a number of such real-life experiments from real sandboxes. Maybe some of these efforts were not designed as an "experimental sandbox" and they were just real examples. Even so, this is all part of a bigger public sector experimentation process, and in that sense, I think that is good, as these examples and legal cases prove the limitations of certain of these approaches.

I think there is now a bit of reluctance, or perhaps more caution, within parts of the public sector per using AI especially because of all the debate on AI, the high risks, and the AI Act. We are beginning to see more examples of the risks and through these examples, better understand the nature of the risks.

I am not saying that is hindering the experimentation, but clearly many governments and public sector organisation are taking a positioning towards the use of AI that they initially took (and continue to take) towards the EU's General Data Protection Regulation (GDPR). This is a cautious posturing along the lines of "I don't know if taking this particular approach is OK and if this is legal per the new EU regulation, or if this is feasible. However, I do know that I don't want trouble, so maybe I better stop for a bit and then get clearer about how to go about this." **So that's why sandboxes are now becoming so popular and important because it's a more controlled and careful way to do things without having to be officially in full compliance with the regulation, and you can see how it will work out and learn more about the issues of being in compliance.**

In fact, this is what happened during the Covid pandemic because in many cases you had to work outside of some of the existing procurement and other rules. Many of these examples of new ways of doing things showed the potential to accelerate certain processes and also showed the need to modify existing rules, regulations and laws. That would have not been possible in in normal conditions.

In Europe there are the usual countries and usual cities that are doing interesting things in terms of AI related policy experimentation and sandboxing as part of their efforts to use AI in the public sector. Closely related to this, in Amsterdam and in Helsinki, they are building some repositories of algorithms and registries with the intent that these can be shared across different public sector locations and domains.[11] But I haven't seen the results of this yet.

4. Thoughtfully managing errors and risks and moving forward under uncertain conditions with AI

4.1 Even without algorithmic or AI support, there are known problems with decision making

Steven Miller: Let's take this public sector issue about determining whether people are eligible for a certain kind of social service or not. For example, are people in a given location eligible for a certain kind of employment opportunity or not? These are kinds of decisions that governments have always had to make. They have to make this type of determination at every level of government and will continue to have to make such decisions into the future.

The assumption is that using AI models to make such decisions or to support the process to make such decisions can be risky, and with the AI Act we have to be even more careful about how we use AI.

At the same time, there's also very strong documented evidence that not using any type of algorithmic decision making for these types of decisions, not using some kind of consistent, evidence-based approach, is also a problem.[12]

If you hold back from using AI and even hold back from using any kind of algorithmic approach, it also leads in some ways to undesirable outcomes in this kind of decision making. And if you use algorithmic approaches, including some of the more sophisticated ones, however far you want to take it with different applications of AI, there are a set of risk issues that have to be carefully understood, characterised and managed.

It's not like you can do nothing because even the alternative of doing nothing (not using any type of algorithmic models for decision making, or not using AI-based algorithmic models) has problematic aspects. For these higher volume types of decisions that often occur in public sector, without the support of some type of consistent, algorithmic approach, the way the public sector staff officer makes a decision on Monday could be different than the way it gets made on Thursday, or the decision made after lunch can be different than way the decision is made before lunch.

Given this reality—of the problematic aspects of insuring consistency, fairness, transparency and accountability in public sector decision making, especially for the types of decisions that are made in higher volumes, are you aware of specific efforts where some level of government is trying to figure out how to go about this?

4.2 Dealing with fears of moving forward given the many uncertainties about the impacts of using AI

Gianluca Misuraca: What I normally use in presentations or lecturing related to what you just mentioned is that the famous book from Cass Sundstein, Laws of Fear.[13]

This Laws of Fear book was republished in 2020 during the pandemic and this republishing was very successful because it dealt with this dilemma related to the fact that

policymakers had to decide one way or another. Should I do something or not? What should I do? Is there any evidence?

The premise of the book is that if you have fear, if you are afraid of doing something wrong, you prefer not doing anything but that can be problematic given the situation.

This is now the situation with the use of AI. I believe that AI is basically something very good for the public sector to use, but we don't know exactly what the consequences could be. It is like the situation with ChatGPT. In some cases, it can be a disaster.

For example, students will copy if you give them access to it, but students can copy anyway, even without ChatGPT. Then there is the positive side, that students can be more creative with the assistance of ChatGPT. But what about younger school children? Maybe kids 10 years old or younger should not be using ChatGPT for school assignments because they first have to form their own cultural and critical thinking. That is something that has to be considered really carefully.

That's the nature of the risk. For these new technology systems, there are these many concerns with the actual and potential wrongdoings, the so-called "bad things," the things that could happen without being anticipated. Of course, nobody wants to have negative consequence, but something can happen because there was a problem or because there was an issue that was not well addressed.

4.3 Can algorithmic accountability lead to "automated grace" and more compassionate use of algorithms (including AI systems)

On the other side, the American law professor Frank Pasquale, well known for his work on algorithm accountability, pointed out that there can also be "automated grace", automation designed to help beneficiaries (in contrast to automatic action adverse to beneficiaries' interests, or that is used to police or punish).[14]

For many types of public sector benefits, the resident has to know to ask for that benefit. You need to know if you are entitled to receive it, and if you think you are entitled, then you have to ask for it. Then the automated systems will check and eventually some humans will say if you are entitled or not. Based on data from the research Prof Pasquale did, he found there are many instances of denied benefits that turned out to be benefits that people were actually entitled to.

Prof Pasquale's provocative question is, "What if you actually give the benefit to everybody that asks for it, and then you use automated system to check afterward if they are actually not entitled to it? If they are not entitled, then in that case, you stop the benefits, or you can have an appropriate penalty." This approach could probably be more effective, and algorithms could also play a role in being able to do this type of screening afterwards to raise alerts of cases of potential non entitlement. I'm saying this because we have to see the possibility to use completely different approaches to our traditional complex processes and challenges. The issue of the complexities in public sector processes, and how to safely reduce those complexities, is something we need to consider.

4.4 Using a sandbox to better understand the nature and consequences of errors and algorithmic transparency when using AI tools

Steven Miller: You just mentioned an example based on these studies of Prof Frank Pasquale about inappropriate denial of benefits. That is an example of a false negative, where the automated algorithm deems that something should be "no" or "negative" when the "true" answer is supposed to be "yes" or "positive." In a wide range of government services, some person or some system needs to make predictions, and based on those predictions, also make recommendations of which action to take. This issue of understanding the meaning and measurement of true positive, true negative, false positive and false negative, which is a very classical and well understood area of study, becomes very important. Just like you said, we need to know for sure if this person is entitled to the benefit or not no.

Yet in in some situations, it is hard to get "ground truth", and without being able to determine ground truth in a given decision-making setting, then it is hard or not even possible to determine the level (and therefor risk) of getting a false positive result or a false negative result. And even when you can determine ground truth, you will still have some level of false positive judgements and false negative judgements.

Are you aware of a country or province or municipality that's doing some interesting work to better understand how to deal with these false positives and false negatives as they're trying to make more use of data and AI enabled recommendations?

Gianluca Misuraca: This issue of false positives and false negatives in public sector decision making is discussed in our 2021 study, "Governing algorithms: perils and powers of AI in the public sector" that I referred to above. In fact, one of the sections in that paper specifically addressed these types of challenges. Navigating between false positives and false negatives is exactly the problem that you find, and now even more so with generative AI where you have these famous issues hallucinations (which you can think of as a false positive as the large language model "predicts" something is true when it is not).

If we go back to the to the design aspect of these systems, there is the lack of transparency, the famous black box effects. You have all these systems that are very sophisticated, but then you don't necessarily know how the data are trained and how the systems themselves are developed.

The AI act (which comes into force as of 01 August 2024) and also Europe's Digital Service Act for online social media platforms, establish the fact that you need to have algorithmic transparency or otherwise you will not be legal and you will therefore be in trouble. **This issue of algorithmic transparency is something that really needs to be addressed. Therefore, sandboxes are something you can definitely do because through specific types of experimentation in the sandbox, you can open up this black box, or at least characterize its behaviour.**

Some parts of the private sector do not necessarily want to do this in an open space because they have their own secrets and their own proprietary things, so that is where this

policy experimentation could actually be useful. We have examples in many countries in Europe that at least are claiming they are starting or doing this type of sandboxing and policy experimentation to better understand how the AI systems they are considering or already using actually work and to demonstrate compliance with the EU requirements.

We have the usual early movers in this space like in Estonia. They are claiming that they have an AI focused government and they have started with a chatbot even before the new generative AI systems, so I suppose they have been updating their chatbot. They are attempting to get the data required to make use of AI and trying to use these AI-based capabilities to be a more proactive government and to try and provide more personalised services. This is the important part. There is big potential if a government can provide more personalised services and more proactive services, though only up to the point where ethical considerations come in because if I as a citizen get a service even before I asked it, that can sometimes lead to delicate or complex issues.

Steven Miller: That's the crux of what we mean by policy experimentation or AI related sandboxing. Sure, we already know we can do personalised service, but how personalised should it be under a given situation? How do you guide it? Sure, we already know we can do proactive services, but how proactive should it be under given situations? How do we understand the trade-offs and the related risk management of doing this?

Gianluca Misuraca: Exactly. There's a huge field of research and experimentation that is required.

5. Public sector use case example: The Italian pension system

I remember when we were doing another project focusing on ICT enable social innovation. We ended up looking at several cases of this in the context of the pension system (a domain area some people may consider not so exciting, but it is obviously very important). In Italy we have the National Institute for Social Security (Istituto Nazionale della Previdenza Sociale or INPS). They have data about all the nearly 60 million citizens in Italy. We made a case study about their innovation efforts and some of their professional staff members are students in our AI4Gov masters programme and they use their master studies assignments and projects to support their professional work to explore how they can further innovate and develop their systems. They have a lot of data and they can do a lot of interesting things to innovate and improve. Of course, these innovation efforts must be within the constraints of existing and emerging laws and regulations because there are limitations in terms of what they can and cannot do per their use of AI.

In Italy and France, there are number of interesting public sector AI application examples. Spain is also quite advanced. For instance, there's was a very interesting case we studied on predictive policing that included all the due concerns and related risk management, and clearly that's also an important area to consider.

Steven Miller: Let's further explore the possibilities for AI applications with a government's pension agency because it's such a fundamental government service.

Gianluca Misuraca: This is very much linked to the social benefits that citizens are entitled to. And it's not just the pensions. The pension is one element and it is important to see what kind of services you can actually get. The link that I don't think has already been done, but I think would be a massive innovation in Italy and other countries, would be linking information from the national pension system to the national fiscal (including tax) system because that would allow identification of potential fiscal frauds.

Italy has a reputation for a high degree of tax evasion and maybe there is some degree of truth to this. Knowing more about the pension system, together with the social service benefits, would allow the relevant parts of the government to better understand that. And on the other side, you may have people that are paying taxes and maybe they have the right to some benefits they are not receiving. This could help to better match different services from a policy perspective.

6. Additional background on the EU AI Act and the current window of opportunity to experiment with compliant AI approaches before enforcement comes into effect

Steven Miller: Earlier in our conversation, you said that you especially like to look ahead and anticipate things about the future. I want to combine that interest of yours with your interests in public administration, digital innovation related change, and usage of AI. The EU AI Act is now written and it has gone through certain stages of approval, but it's not fully enacted yet (as of 02 May 2024). Can you explain the exact status of the EU act?

Gianluca Misuraca: The European Commission first proposed the AI Act back in April 2021 based on background work that had been done prior to that time. From that time through December 2023, intensive negotiations took place between the European Commission, the European Parliament, and the Council of the European Union on the specifics of the content. There were many discussions related to provisions related to human rights, human centricity, and ethical impact assessment. ChatGPT and similar large language models were introduced during this negotiation period and many things had to be reconsidered that had previously been agreed to, including even the definition of AI. Also, other countries as well as industry, including the global providers of commercially available large language models, expressed their views.

Even with all the changes resulting from these negotiations, the basic architecture of the act remained anchored on a risk-based approach. A political agreement on the Act was reached in December 2023.[15] The European Parliament formally adopted the AI Act in March 2024. As of now, the AI Act has not yet been officially published and entered into force though

this is expected to happen soon, perhaps within the next few months.[16] Once this happens, there is a two-year period to prepare for compliance with most aspects of the act.

Everyone agrees that we need some rules, including the big tech companies, but we still don't know exactly how the rules will be applied. In fact, everybody's a bit concerned, although the period of implementation for most aspects of the act will only start two years after the AI Act law is officially published and entered into force. There are already attempts to experiment with compliance to the AI Act. For instance, there is the AI pact that was launched by EU Commissioner Thiery Breton in November at the Madrid event that we organised where there was a proposal for companies to start experimenting with their usage of AI as if the AI Act was already in force and to see what would be the implications.[17]

The EU's Digital Services Act (DSA) entered into force in November 2022, and became applicable in February 2024. The Digital Markets Act (DMA) also entered into force in November 2022 and became applicable May 2023. The AI Act will require some more time as it will only come into force within the next few months and then most of it becomes applicable two years after that date. **That's why we have a window opportunity to really experiment.** That's why it would be good to have a conversation with the Member States in Europe, because they would also be interested experimenting. Maybe that could be an extension of this project.

6.1 Healthcare and education would be high potential (and high risk) settings for experimenting with compliant AI applications and policies

Steven Miller: Given what you know about the AI Act, what would be some examples of experiments that you think would be good to do now? What are some good scenarios for AI usage examples to try and assess in the context of the AI Act's risk-based approach?

Gianluca Misuraca: Areas that are at the core like health and education. In these areas, there is high potential for benefit for both individuals and the public overall, and these are also areas where there are high risks because of the sensitiveness. These are areas where you could inevitably use this technology to the full potential because you have a lot of data and a lot of user interactions. These are areas where you could define and design policies that could anticipate and be proactive. In health for example, you could anticipate possible disease and health problems by analysing larger amounts of data and identifying patterns that we cannot even see as humans.

And in education as well because there needs to be a radical change in the way we see education systems versus the way they have traditionally been built. The new generation of children are starting to experiment by themselves with technology. The other day (30 April 2024), an expert panel commissioned by French President Macron announced their recommendations on the extent to which a child or teenager's access to mobile phones, any type of smart phone or tablet, or social media should be limited or even excluded, depending on their age.[18]

This is an example of proposed regulation. Is it based on evidence? In some cases, there is some evidence, but not in all cases. This is an area where of course it's also difficult to see how you can really experiment with the real lives of children. But this is definitely something we need to do in one way or another because it is already happening. Already, we have some schools in some countries that are using technology, and others that are not. These can be natural experiments for comparison and that type of comparative analysis could be something good.

I'm part of a group of experts on AI and education for the Council of Europe.[19] The Council of Europe's Committee on AI has recently approved the draft of the Convention on AI (officially referred to as the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law).[20]

A process has recently been promoted to explore if in addition to this recently announced Convention on AI, there should also be a specific attempt to make regulations on the use of AI in education. Member States are not so in favour of doing this, but there will be an October 2024 conference to discuss this and that conference is "bravely" called "Conference on Regulating Artificial Intelligence in Education."[21]

I think that in these two areas, healthcare and education, the AI technologies are already mature enough to useful things and to change the systems in a way that can benefit society hugely. But the risks are also quite big.

Steven Miller: Are you aware of specific country level or state/provincial level initiatives where they are starting to systematically test how do align with the EU AI Act?

Gianluca Misuraca: Systematically, I wouldn't say so, at least not yet. A lot of people are looking at Spain for this type of example because there is already an AI regulatory sandbox ongoing effort.[22] Now, how much has really been done? The jury is still out. We will see what will be reported hopefully soon.

As I mentioned earlier in this interview, there is now (29 Feb through 29 May 2024) this ongoing call for projects within the EU that would do exactly that type of testing in several countries for how do align with the EU AI Act.[23] Once proposals are eventually accepted from this ongoing call, there will be at least three or four other countries that will be doing some regulatory sandboxing efforts within different countries to explore the implications of the EU AI Act.

7. The importance of the conceptualisation phase for regulatory sandboxing as well as for designing the regulation

Steven Miller: In their prior work, the UN DESA team had talked about the following four phases of a sandbox: conceptualization, operating the sandbox, evaluating what happened, and then exiting which would lead to either sunsetting the effort, or making adaptations and continuing to run trials, or adopting and scaling up the effort.

Do you see any of these stages as being more challenging in this realm of experimenting with pilot applications that incorporate AI capability? In your view, which are the most challenging or problematic aspects of these four phases?

Gianluca Misuraca: I would say the conceptualization phase is what is really important because you need to design these in the right way. Otherwise, yes, you can do a sandbox effort that seems very good in terms of results, but then maybe it's not really answering your needs. I think that the design part is what would require the most effort. What we sometimes fail at in these sandbox efforts, including in the conceptualization phase, is putting too much emphasis on the technological part and not effort emphasis on the social aspects and the legal aspects which are also very important to consider, but cannot be the only thing to look at.

Now there is a bit of mismatch. You have this EU AI Act that seems to be the most important legislation and regulation of its type in the world, for now at least.

And then you say, "Why are you are regulating AI in this way?" Because of the risks that you want to avoid. Yet maybe there are other ways of avoiding that same risk, for example through a combination of education and training and other mechanisms. These basic questions of "why are we regulating?" and "are there other ways to address the key concerns?" are other points that are important to consider.

Steven Miller: That's an excellent point.

8. Bridging the gap between the content of the AI Act and everyday practice and organising across the EU for enforcement and oversight

8.1 Steps towards frameworks and tools for assessing compliance with the AI Act starting with the ALTI tool

Steven Miller: What practical and actionable recommendations can you give to a public sector unit that wants to pilot how to make use of AI and how to go about this experimentation? And for how to go about this sandboxing? While each AI application use case is different, is there a common framework that you can recommend? Or specific documents that practitioners can look at for guidance?

Gianluca Misuraca: There is at least an attempt to develop some guidelines. For instance, The European Commission has the High-Level Expert Group on AI.[24]

As a result of the work of this High-Level Expert Group, they have already developed a document on Ethical Guidelines for Trustworthy AI and a tool called ALTAI (Assessment List for Trustworthy AI) that is a checklist used to try to understand the issues to look at for having ethical, trustworthy AI.[25]

These prior efforts of the High-Level Expert Group also guided the AI Act to some extent. Now that we have the AI Act, how we do the checks for compliance to the Act and for alignment with other key EU guideline documents is becoming more important.

8.2 Organising across the EU for AI Act governance and enforcement oversight

Steven Miller: Consider a public sector employee doing their everyday work. Let's say this person needs to help his or her unit improve the determination of whether somebody is eligible for some particular social benefit, and they are part of a team that is using their data to create AI-based prediction and recommendation models to support this decision making. And let's say this public sector employee has the EU guideline documents for ethical and trustworthy AI and has the EU AI Act document. Wouldn't there be a big gap between the more general principles described in these guideline and regulatory documents and the everyday work setting of this person who over the next few months has to figure out with his/her team how to do a better assessment of who is or is not eligible for a specific social benefit?

Gianluca Misuraca: Well, that's exactly what we are confronted with at the moment. We can have a generic list of how we assess the trustworthiness of a proposed system or a system being piloted, but how we do this assessment in practise is a different story. That's where I guess the debate about the AI Act in Europe will become hot because there are political decisions to be taken.

There was discussion whether the EU should have a central agency or authority that takes care of every Member State per AI policy and compliance matters related to the AI Act or rather whether the Member States themselves should have their own authorities, or whether there should be a combination of both.

The decision has been finally taken in the EU to have this AI Office.[26] What was finally agreed upon is that there must be a specific EU wide authority for the AI Act that will work together with the designated authorities in each member state. As such, the administration of the AI Act for compliance and oversight will be a shared responsibility between the EU-wide authority and the designated authorities in each member state. This is to help ensure a more consistent approach to AI regulation across the EU. This approach is similar to how the GDPR is administered.

Are Member States ready for that? Absolutely not. They are trying to see how to do this and figuring out which agency within their country should be appointed as the AI Act authority. Even after the policymakers in a Member State determine which agency has authority to address issues related to the AI Act, that's just the beginning. The beauty and power and risk of these technologies is that they have an impact everywhere, so how a country makes the distinction of which agency should have oversight authority is not too clear. And then the Member States are all looking at the recently established EU-level AI Office to provide guidelines, rules and policy directions. These have been hard things to do in the past, so it will be a challenge for the new EU-level AI office to do this now and in the future.

There are additional complications. As much as Europe has prepared its own arsenal of laws, rules and guidelines regarding AI usage, this is something that needs to have global coordination mechanisms as the use of AI has global implications and there is the need for global governance per AI laws, regulations and usage. This is where UN agencies can

potentially play a role. The UN can also build upon the ongoing UNESCO AI ethics efforts and the Secretary General's AI advisory body.

Maybe a concrete and tangible effort towards global governance of AI could build on the ongoing efforts of the G7. That of course can be criticised because the G7 is "just" the few countries of the G7. But in that setting, the G7 has been informed and inspired by the work that many scholars, including myself, have done on AI governance and usage. The current G7 President (from Italy, till end December 2024) has put as a priority AI in the public sector. In a recent ministerial declaration on AI and Innovation they have issued, they said they will create a toolbox on AI for the public sector.[27]

We also have to see how the industry reacts to the AI Act. Government can say, "I give you the rules." Industry can say, "Yes, and I also have my own rules, my own guidelines, my own internal systems." So, there are issues about how industry will actually respond to the AI Act, and the relations between government and industry per the AI Act.

There is something that we haven't touched yet and that is the procurement part. This important aspect also requires an additional consideration.

Steven Miller: We know procurement is a very important aspect because in both public and private sectors, a main way that AI and other technology and change initiatives move forward is through procurement.

9. 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?

Steven Miller: Let's conclude. I am most thankful for your time and for your generous sharing of your experience.

Gianluca Misuraca: This interview was also a learning process for me. I am glad that this was useful to you and to the project. I'm very much interested in knowing the results of this work on AI policy experimentation and sandboxing in the public sector because these types of efforts are very much needed.

Through the European Commission's Joint Research Council (JRC) Public Sector AI Watch (previously mentioned above) and Public Sector Tech Watch initiatives, they continue to capture a lot of data and a lot of new cases about the use of AI and digital technology in the public sector.[28] The challenge is how we learn from that information. I think there is a need to come up with an analytical framework to better capture and assess information about the use and impacts of AI in the public sector, and to assess the implications of different types of uses of AI. I think that would be much needed as such a framework is still lacking from a theoretical and also practical perspective.

Steven Miller: As we close Gianluca, what did you take away from this interview? What did you most enjoy about this discussion?

Gianluca Misuraca: What I take away is that we are really are at the at the at the core of the problem, at the centre of this. It is not just the technological development. It is more than the ChatGPT kind of debate. The challenging issues are how policymakers and civil servants, with all the limitations they must work within, can really embrace the complexity of these epochal changes and make sure that new capabilities enabled by AI and other new digital technologies can help citizens and society to function better. Of course, it's not easy because there are many challenges and resistances, but from an Idealistic and optimistic view, this is what I believe that we should try to achieve.

I'm very happy having had this conversation with you and I hope this won't be the last one.

Endnotes

[1] For background on Prof Gianluca Misuraca, visit https://www.icegov.org/people/gianluca-misuraca/.

[2] See AI4Gov master programme website at https://www.ai4gov-hub.eu/master/.

[3] European Commission, The Digital Europe Programme, https://digital-strategy.ec.europa.eu/en/activities/digital-programme.

[4] European Commission, EU Funding & Tenders Portal, AI regulatory sandboxes: EU-level coordination and support (call for proposal), 29 February 2024, https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/digital-2024-ai-act-06-sandbox.

[5] DigitalEurope, SANDBOXING THE AI ACT: Testing the AI Act proposal with Europe's future unicorns, June 2023, https://cdn.digitaleurope.org/uploads/2023/06/DIGITAL-EUROPE-SANDBOXING-THE-AI-ACT_FINAL_WEB_SPREADS-1.pdf.

[6] European Commission, JRC Publications Repository, AI Watch - Artificial Intelligence in public services, July 2020, https://publications.jrc.ec.europa.eu/repository/handle/JRC120399.

C van Noodt and G Misuraca, "Artificial intelligence for the public sector: results of landscaping the use of AI in government across the European Union," Government Information Quarterly July 2022, https://www.sciencedirect.com/science/article/abs/pii/S0740624X22000478?via%3Dihub.

European Commission, AI Watch website, https://ai-watch.ec.europa.eu/index_en.

European Commission, Joint Research Centre Data Catalogue, EU Artificial Intelligence Observatory, https://data.jrc.ec.europa.eu/collection/id-0130.

European Commission, Joint Research Centre Data Catalogue, Artificial Intelligence Observatory, Selected AI cases in the public sector (JRC129301) http://data.europa.eu/89h/7342ea15-fd4f-4184-9603-98bd87d8239a.

European Commission, AI Watch, Public Sector: Investigating the potential use and impact of AI for the public sector https://ai-watch.ec.europa.eu/topics/public-sector_en

[7] M Heikkila, Politico, Dutch scandal serves as a warning for Europe over risks of using algorithms, 29 March 2022, https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/.

J Gesley, US Law Library of Congress Global Legal Monitor, 13 March 2020, Netherlands: Court Prohibits Government's Use of AI Software to Detect Welfare Fraud, Global Legal Monitor, https://www.loc.gov/item/global-legal-monitor/2020-03-13/netherlands-court-prohibits-governments-use-of-ai-software-to-detect-welfare-fraud/

M van Bekkum and F Z Borgesius, European Journal of Social Security, Vol 23(4), 02 August 2021, Digital welfare fraud detection and the Dutch SyRI judgment, https://journals.sagepub.com/doi/full/10.1177/13882627211031257

A Rachovitsa and N Johann, Human Rights Law Review, June 2022, The Human Rights Implications of the Use of AI in the Digital Welfare State: Lessons Learned from the Dutch SyRI Case, https://academic.oup.com/hrlr/article/22/2/ngac010/6568079.

[8] GovUK, Ofqual, 13 August 2020, Research Analysis of Awarding GCSE, AS & A levels in summer 2020: interim report, https://www.gov.uk/government/publications/awarding-gcse-as-a-levels-in-summer-2020-interim-report.

[9] B Walsh, AXIOS, 19 August 2020, How an AI grading system ignited a national controversy in the U.K., https://www.axios.com/2020/08/19/england-exams-algorithm-grading.
D Kolkman, LSE Blog, 26 August 2020, What the world can learn from the UK grading fiasco, https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/.

T Harkness, BBC Science Focus, 11 January 2021, A level results: Why algorithms aren't making the grade, https://www.sciencefocus.com/future-technology/a-level-results-why-algorithms-arent-making-the-grade.

A Kelly, British Education Research Association (BREA) blog article, 21 May 2021, The great algorithm fiasco, https://www.bera.ac.uk/blog/the-great-algorithm-fiasco.

[10] G Misuraca and T Alvarez, Digital Future Society (Barcelona Spain), 2021, Governing algorithms: perils and powers of AI in the public sector, https://digitalfuturesociety.com/report/governing-algorithms/.

[11] M Haataja, L van de Fliert and P Rautio, City of Amsterdam and City of Helsinki White Paper, September 2020, Public AI Registers: Realising AI transparency and civic participation in government use of AI, https://algoritmeregister.amsterdam.nl/wp-content/uploads/White-Paper.pdf.

UN ITU, 30 September 2020, Helsinki and Amsterdam launch AI registers to detail city systems, https://www.itu.int/hub/2020/09/helsinki-and-amsterdam-launch-ai-registers-to-detail-city-systems/.

City of Helsinki, AI Register website: https://ai.hel.fi/en/ai-register/.

City of Amsterdam, Algorithm Register: https://algoritmeregister.amsterdam.nl/en/ai-register/.

[12] D. Kahneman, O. Sibony and C. Sunstein, May 2021, Noise: A flaw in human judgement, https://www.hachettebookgroup.com/titles/daniel-kahneman/noise/9780316451383/?lens=little-brown.

[13] C Sundstein, 2005, Laws of Fear: Beyond the Precautionary Principle, https://www.cambridge.org/core/books/laws-of-fear/16124E83F371BEAA5082AB07EA892836.

[14] University of Melbourne, Centre for AI and Digital Ethics, announcement for Prof Frank Pasquale seminar on Automated Grace: Toward More Humane Benefits Administration via Artificial Intelligence, 19 July 2022, https://www.unimelb.edu.au/caide/news-media-and-events/online-seminar-with-frank-pasquale. Also see Prof Frank Pasquale web site, Cornell University School of Law: https://www.lawschool.cornell.edu/faculty-research/faculty-directory/frank-pasquale/.

[15] Topics, European Parliament, EU AI Act: first regulation on artificial intelligence, 08 June 2023 and updated 19 December 2023, https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. Also see:

- Think Tank, European Parliament, AI Act briefing, 03 November 2023, https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792.
- Future of Life Institute, EU Artificial Intelligence Act website, High Level Summary of the AI Act, 27 February 2024, https://artificialintelligenceact.eu/high-level-summary/.
- Future of Life Institute, EU Artificial Intelligence Act website, AI Act Implementation: Timelines & Next steps, 28 February 2024, https://artificialintelligenceact.eu/ai-act-implementation-next-steps/.

- Post-interview update on European Council final approval of the EU AI Act, European Council, Council of the European Union, Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI, 21 May 2024, https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/.

[16] The Council of the EU made the final approval of the EU's AI Act on 21 May 2024. https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/. The AI Act was published in the official journal of the European Union on 24 July 2024, and went into force on 01 August 2024. https://commission.europa.eu/news/ai-act-enters-force-2024-08-01_en and https://www.artificial-intelligence-act.com/.

[17] European Commission, DIGIBYTE website, Fourth European AI Alliance Assembly took place in Madrid, 15 November 2023, https://digital-strategy.ec.europa.eu/en/news/fourth-european-ai-alliance-assembly-took-place-madrid.

[18] Le Monde (English), Kids and screentime: What the committee appointed by Macron recommends, 02 May 2024, https://www.lemonde.fr/en/france/article/2024/05/02/kids-and-screentime-what-the-committee-appointed-by-macron-recommends_6670225_7.html#. Also see:

- Reuters, France must curb child, teen use of smartphones, social media, says panel, 30 April 2024, https://www.reuters.com/world/europe/france-must-curb-child-teen-use-smartphones-social-media-says-panel-2024-04-30/.
- Expert Panel commissioned by French President, Enfants et écrans, À la recherche du temps perdu, 30 April 2024 (original report in French), https://www.elysee.fr/admin/upload/default/0001/16/fbec6abe9d9cc1bff3043d87b9f7951e62779b09.pdf .

[19] Council of Europe, Education Department, AI and Education, https://www.coe.int/en/web/education/artificial-intelligence-and-education. Also see Council of Europe, Education Department, AI and Education Expert Expert Panel update, https://www.coe.int/en/web/education/-/artificial-intelligence-and-education-expert-group-meets-in-strasbourg.

[20] Council of Europe, Committee on Artificial Intelligence, 14 March 2024, https://www.coe.int/en/web/artificial-intelligence/cai.

[21] Conference announcement: https://epale.ec.europa.eu/en/content/council-europe-working-conference-regulating-use-ai-systems-education Conference concept note: https://rm.coe.int/2nd-working-conference-on-regulating-the-use-of-ai-systems-in-educatio/1680b0dfcd.

[22] European Commission, Launch event for the Spanish Regulatory Sandbox on Artificial Intelligence, 08 June 2022, https://digital-strategy.ec.europa.eu/en/events/launch-event-spanish-regulatory-sandbox-artificial-intelligence.

The Legal Wire, Spain Pioneers EU's First AI Regulatory Sandbox: A Step Towards Innovative Compliance, 20 November 2023, https://thelegalwire.ai/spain-pioneers-eus-first-ai-regulatory-sandbox-a-step-towards-innovative-compliance/.

[23] European Commission, EU Funding & Tenders Portal, AI regulatory sandboxes: EU-level coordination and support (call for proposal), 29 February 2024, https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/digital-2024-ai-act-06-sandbox.

[24] European Commission, High-level expert group on artificial intelligence, https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai.

[25] European Commission, Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, 17 July 2020, https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

[26] European Commission, European AI Office, https://digital-strategy.ec.europa.eu/en/policies/ai-office. Also see

Future of Life Institute, The AI Office: What is it, and how does it work?, 21 March 2024, https://artificialintelligenceact.eu/the-ai-office-summary/.

[27] Gov.uk, G7 nations to harness AI and innovation to drive growth and productivity (press release), 15 March 2024, https://www.gov.uk/government/news/g7-nations-to-harness-ai-and-innovation-to-drive-growth-and-productivity. Also see:

- Gov.uk, G7 Ministerial Declaration: deployment of AI and innovation (link to declaration), 15 March 2024, https://www.gov.uk/government/publications/g7-ministerial-declaration-deployment-of-ai-and-innovation.
- UNDP, G7 consensus reached on advancing AI for sustainable development, 15 March 2024, https://www.undp.org/news/g7-consensus-reached-advancing-ai-sustainable-development.

[28] European Commission, Public Sector Tech Watch web site, https://joinup.ec.europa.eu/collection/public-sector-tech-watch.

INTERVIEW 7: Dilshat Saitov, Nigmatullo Sharafutdinov and Jahongir Topildiyev, Ministry of Digital Technologies, Uzbekistan

Date of Interview: May 07, 2024

Dilshat Saitov

Head of Division for Cooperation with International Rating Organizations
Digital Government Projects Management Centre, Ministry of Digital Technologies[1]

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

1. Overview of Uzbekistan efforts with digital government and AI applications

In our view, Uzbekistan is doing well with its digital government efforts. In the 2022 UN E-government survey, we were ranked 69th on the survey's E-Government Development Index (egdi). Two years prior, in the 2020 version of the survey, we had ranked 87th, so we increased 18 positions because of the progress we had made. The UN survey assess 22 areas of online government digital services. We have all of them in place except for one (vehicle registration). Over these most recent two years (2022 – 2024) we have been focused on making further progress so that we might do even better in terms of our egdi ranking in the forthcoming UN 2024 E-government survey.

We have already applied AI in several ways to enhance the way that our citizens interact with our government digital services portal. We have already implemented the following AI-enabled functionality: 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 people who have trouble typing or who have trouble reading the screen more easily interact with our portal and its applications.

We tested our face ID system for more than one year. We also worked with an external international certification organisation to validate the performance of this system during our testing period. That international certification organisation required that we demonstrate at least a 90% accuracy level for them to provide the certification. We were able to meet and even exceed that accuracy level and receive their certification.

Our face ID system will only allow a user to access to our government digital services portal if it is at least 92% confident of the identify match. Our internal evaluations have shown that the system is performing at over 99% accuracy.

We have also implemented this face ID system for user authentication on our government digital services mobile app. We have not had complaints from our citizens about using this system, so it seems to be working very well.

The facial ID application uses a database we created called One ID that is used by the government as well as by some non-government organisations for identification. Private sector banks in our country also use this One ID database when they need to do online authentication of their customers. For example, if you want to apply for a loan online, some of our banks uses this same One ID database to validate your identity. For validating the online user's identity as part of issuing a bank loan, their threshold for the probability of a correct identity match is even higher than 92%. It is above 99%.

While our face ID system is performing very well, it is not 100% accurate. There are some identification errors. For those instances where a citizen who is trying to access our government digital services is not able to be properly authenticated by our face ID system, they can use an alternative way of authentication using the messaging capabilities of their mobile phone, or their government digital portal password.

The voice assistant is now being piloted in our internal sandbox. If a person has difficult reading what is on the screen of the computer or mobile device, the voice assistant can say the information to them so they can hear it. Similarly, if they have difficulties with using the keyboard to input information, they can ask the voice assistant to help them proceed with their request. We are also in the early stage of working with a company selling automobiles in Uzbekistan to test out the use of our Voice Assistant in their car.

We are now working on testing and deploying a new generation of chatbot that can assist users to find the government information, documents or forms they are looking for. The chatbot can determine what type of information the user is seeking, where it can be located within the government, and assuming that information is available for public access, then proceeds to make the connection to the appropriate government data base or hosting site so that the requester can access the information they are seeking. We will be deploying this new AI-based service before the end of 2024.

2. More emphasis on proactive digital services

We are also currently working on and planning for a range of other AI-enabled applications as well including applications that we refer to as proactive public services. For example, suppose an older citizen reaches retirement age, and they want to retire. Currently, they have to notify several different government agencies about their plan to retire, initiate the various request for retirement related government benefits, and fill out and submit all of the necessary application forms. This is a lot of administrative work for the person who is planning to retire. Also, if they forget to do any of these steps, it causes complications.

With the new AI enabled proactive services we are now developing and testing, this person would automatically receive notification of their upcoming retirement. If they wanted to proceed with retiring at this time, all the necessary notifications would be automatically done by our digital services system, and the necessary applications for the benefits they are entitled to would also be automatically submitted. It would make it much easier for a retiring citizen to administer the process of retiring and to receive the government benefits they are entitled to.

We are in the process of planning, creating and deploying additional proactive digital services. Examples include a more proactive approach to notifications, recommendations for available government services and support, and follow-on transaction support related to:

- Government provided insurance policies, e.g., notifications for expiration and support for renewal.
- Passport expiration and renewal.
- Payments for taxes, water, gas.
- Getting vaccinations.
- Parents with new babies and with young children, e.g., about options for kindergartens and schooling where people need to register several years in advance, and about other services specific to children.
- People with disabilities, including families with children with disabilities.

During 2024, we will be implementing an additional 10 proactive government digital services, including some of those listed above. They reason we can only accomplish implementing 10 new proactive services this calendar year is that some of these are quite complex to implement.

Another aspect of making our services more proactive is to make it easier for our citizens and residents to complete the end-to-end process related to receiving a benefit or a service. Suppose we notify a single mother with a disabled child that she is eligible for government support payments, and she would like to receive this benefit. If she already has a bank account number, then she can share that account information with us, and we can arrange for the payments to get sent automatically to her account on a monthly basis.

Sometimes we encounter situations where a person in this situation does not have a bank account. Our proactive approach is to inform the single mom of a nearby bank she can go to that will provide her with an account. She just needs to go to that bank and receive her account number, as we would already have taken care of the effort to provide the bank with the information they need to process her application for opening a new bank account (with her prior approval). Then, we can arrange to have the monthly support payment sent to that new bank account.

A supporting initiative for a more proactive approach to supporting our citizens with digital services is the recommendation system we are developing. When a citizen logs in, an AI-based recommendation system would make suggestions to them based on based on their profile, their existing transaction records in government data bases, and their prior

government services related interactions or transactions. These recommendations will also be able to be communicated using our Voice Assistant. This recommendation application is a way of combining our proactive services efforts with our AI efforts.

3. Our usual steps for sandboxing and piloting

In our view, the term “policy experimentation and sandboxing” means testing government public services in an experimental place. We think of a sandbox as a place where we are testing a new public service or experimenting with policies related to delivering or using those services. We run pilots in our internal sandbox settings where the testing is done by internal government staff from our ministry and from the other ministries we are collaborating with, and by the external vendor who may be helping us to build the service. We also run pilots outside of our internal sandbox environment when we do live trials with limited numbers of citizens who are actual users.

We typically go through the following three steps in our sandboxing and piloting efforts:

First, we build and test the application in our own internal Ministry of Digital Technology sandbox. This may involve using the services of an external technology provider. We design, build and test the application in partnership with any other government ministry involved in providing that particular government service. For example, in case of the new proactive service to simplify the process of declaring and registering for retirement, we work in partnership with the Ministry of Labour. We partner with that agency and our service provider to do extensive testing of the application and address all the problems that we can find through this internal testing.

Second, we migrate the application to our digital government portal and do a pilot where we only use the application with a limited population of citizens and residents who are external users. For example, we might initially launch the application in a smaller city. We continue teaming with any of the ministries involved in delivering the service. Together, we rapidly address any problems that surface as our citizens and residents make use of the new digital government service. We continue this pilot for as long as needed to address problems, and to establish our confidence (via evidence) that the new digital service works well and that the outputs provided to the citizen/resident user are appropriate and of sufficient quality.

The third step is to expand the access to the new digital government service to all other people across all other parts of our country. We continue the partnership with any ministry involved with the service. As needed, we continue to address any problems that arise with using the new service as more and more people across the country make use of it.

In some situations, we may include an additional step between this second and third step mentioned above where we expand the scope of the pilot, for example, to additional people in another smaller city, before we release the application to our entire country.

Our team within the Ministry of Digital Services is the central coordination and control point for getting external and internal user feedback related to any problems with a new or

existing digital service that we release. We then coordinate the process of working with whatever other parts of the government need to be involved to make the necessary corrections and changes. Any type of feedback related to a problem with a digital service comes to us. It could be a policy related problem dealing with the content or rules related to the digital service. It could be a technical error. It could be a user experience or usability problem. Whatever the problem is, we receive and review the feedback, determine the nature of the problem, decide how to address it, and work with any other part of the government that needs to be involved to address the issue to the extant that that we assess it is necessary, practical and worthwhile to do so.

4. Our approach for working with ministries to bring additional services online

Within our Ministry of Digital Technologies, the Division of Introduction of Electronic Public Services and Interdepartmental Electronic Cooperation creates and manages a master list of government services that we want to have online. This includes existing government services that are still only offered off-line as of now as well as new online services that may not exist as offline services. Most of these services on this master list are passive services in the sense that the citizen has to initiate the process of using them. As of now, only a small number are the more proactive types of services that we mentioned above. As we keep progressing with our digital government efforts, we are trying to make more of our services be more proactive.

For every government service that we already offer online or plan to offer online, we define all the ministries and units under each ministry that are involved with offering and managing that service. For some types of services, many ministries may be involved. For example, for the set of services related to families that need help with financial support, we need to work with 11 different ministries in the effort to bring this type of service online. We have a total of 26 ministries in Uzbekistan and we work with nearly all of them across our different projects to move more and more of our government services online.

To make a government service available online, we make a proposal for doing this and review it with representatives of all ministries involved with that service. If they agree to move ahead with making the service online, then we continue with making a more detailed proposal. If they do not initially agree, then we continue our discussions and negotiate until we can find some type of mutually acceptable win-win approach.

The final version of the proposal for providing a new government service online is a document that is submitted to our country's national cabinet of ministries and must be approved by the cabinet. The proposal includes 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. This proposal document is the roadmap for how to proceed with the project.

Once we have cabinet level approval, we proceed with the effort. This includes doing a detailed business process analysis to determine if there are ways to simplify the workflow required to support the process both online and offline. Often, we find that it is possible to eliminate a number of existing process steps to optimize and simplify the workflow.

We also work out all the details for the necessary data access and integration needed to implement the online service. We define what kind of data we need from each ministry to implement the proposed service online. We work out the details for how to make the connection between the relevant data bases of all participating ministries with the centralized government digital services portal, and how to handle the necessary data integration. We encounter situations where the ministry may have the necessary information, but it is not in the needed format, or not even in electronic form. We figure out how to address and resolve those types of issues. We teach our ministries how to move forward with improving the accessibility and quality of their data, and we work with them to take care of all the necessary data connectivity and integration. Working out these details related to data access and integration, and related to data formats and quality, is a big part of our effort as each of our 26 ministries has their own information system.

Once an agreement is made to proceed with making a service available online, we must work with every one of the ministries in parallel involved with providing that service. Even if 15 ministries are involved in a service we are proceeding with, we must take care of the business process simplification, data connectivity, data integration and data format efforts for all of the involved ministries in parallel, or else it would not be possible to fully implement the service. When we are implementing a new online government service that requires participation from a larger number of ministries, it is a more complex and time consuming effort.

Naturally, our Ministry of Digital Technologies has our own capacity constraints in terms of how many of these projects of putting services online we can execute in any given time period, especially given that many of these efforts involve working with multiple ministries in parallel in order to implement one particular service. We manage this by prioritizing the items on our master list and deciding which new projects to take in a given time period given consideration of ongoing projects that must be brought to completion.

5. Plans for increasing AI usage in our online government digital services and implications for sandboxing and piloting

We think that AI is not just important. It's very important. We are studying more about AI and are already planning to implement and use AI as much as possible as we proceed with our online digital service efforts.

Currently, our call centres are still playing a major role in the way we provide services for our citizens. We need AI to increase the capacity and productivity of our call centres and to reduce the load on those call centres. Our call centre staff receive many calls on the same topics, and they have to repeat the same answers over and over again. We could offload this

type of call to an AI system so our call centre staff can handle the non-routine and more complex types of calls. This is an example where we want to use AI to automate the more routine tasks of providing government services to our citizens and residents, so our government service employees can spend their time on more complex things and on new things.

Will our plans for using AI in a growing number of projects lead to new types of risks and require new or expanded approaches to our risk management efforts compared to what we are already familiar with? We don't know yet. We are still in the very early phase of using AI and our deployment experience is still quite limited to the few examples we discussed earlier. At this early phase of our journey with AI, we do not have enough specific experience to know if we need to take new measures to deal with new types of risk that may result from increasing use of AI in our online government digital services. We will have to see how this evolves and respond accordingly.

Even with the few AI applications we have worked on and implemented so far, we have seen that we need to do a lot of internal sandbox testing to confirm and improve the accuracy level and the need to do the next stage pilot testing with external users and use their feedback to make additional iterative refinements to improve accuracy and reduce inappropriate responses. As we proceed with implementing more AI in our online government digital services, we will keep doing this type of careful output testing and piloting in order to improve the accuracy of the AI outputs to a satisfactory level.

As a result of this interview, we better appreciate that as we continue expanding our efforts to make use of AI, and as we use increasingly sophisticated types of AI, our testing, evaluation, and iterative refinement efforts to control the quality of the AI outputs will have to expand. We may have to expand our internal sandboxing efforts. We may also have to expand the scope and duration of our pilot efforts with external citizens and residents prior to releasing the new AI-enabled service to the entire country. We may also have to allow for the possibility that this sandbox and pilot testing may take longer as we work with more sophisticated types of AI capabilities.

This interview makes us realize that as we proceed with our AI efforts, we should not underestimate the time it may take us to more clearly understand the quality control issues, and related risks and risk management needs, of using AI outputs for new types of online government services.

Endnotes

[1] For background on Uzbekistan Ministry of Digital Technologies, visit the official website at https://mitc.uz/en/pages/about_ministry. Also see the Single Portal for Interactive State Services which provides interactive public services in the "one window" mode to the public and business entities, https://my.gov.uz/.

INTERVIEW 8: Esther Kunda, Director General, Innovation & Emerging Technologies, Ministry of ICT and Innovation, Rwanda

Date of Interview: May 31, 2024

1. Introduction to Esther Kunda and her portfolio

My official title is Director General in charge of Innovation and Emerging Technologies within the Minister of ICT and Innovation (MinICT).[1] The three main departments under MinICT oversee:

- Infocomm Technology (ICT) in general including connectivity, digital literacy, and access to devices.
- The innovation mandate related to encouraging and supporting innovation within our government and across our economy related to the use of digital and ICT (including AI).
- Future planning, which is trying to understand where ICT, Digital and AI technology is going, and how can we better position Rhonda to be a front runner in using ICT as a cross cutting enabler for economic development to help us better adapt to the future.

Under my department for Innovation and Emerging Technologies, our work can be viewed as coming under the umbrella of innovation ecosystems. This includes supporting the growth of startups in Rhonda and trying to understand what kind of programmes we can build for them. It includes finding ways of unlocking capital for startups. We also look at digital financial services and e-commerce as these areas, including capabilities for cashless transactions, are closed linked to our innovation ecosystem. My department also covers emerging technologies with a focus on how we can implement them both in the public and private sector.

My department defines and implements appropriate interventions and policies. We build advocacy and partnerships across our stakeholders and participants. Much of the work of my department involves building and implementing policies, interventions and supporting partnerships.

2. Prior Rwanda AI efforts with local language-based chatbots

We started working with chatbots as our first use public sector usage of AI several years ago (pre-COVID) to see how we could use them as part of dealing with the public and delivering government digital services. During COVID, we created a chatbot for our healthcare sector to support government efforts to educate our population about our COVID measures and related government policies. The chatbot provided another means of communicating with our citizens about COVID related restrictions and curfews, and about the type of

vaccinations that were or would be available in the country using our native language of Kinyarwanda so that our local people could use it as an information source.

The ability to create this native language chatbot during the Covid years grew out of a prior partnership project we had started with the Mozilla Foundation in 2018-2019 to create voice datasets in our native Kinyarwanda language.[2] We organised community volunteer efforts to come and record Kinyarwanda spoken sentences and to review how they were converted into text. Using this large data set, we were able to build speech to text models in our native language, and to link this with text-based translation models between Kinyarwanda and English. This enabled us to use our native Kinyarwanda language within any type of AI solution that uses language input or output in the forms of either voice or text.

So fast forward during COVID. Building on this prior local language data set effort, we were able to quickly build a chatbot for our citizens that they could use in their native language to understand COVID related restrictions, policies and support information. Since that time, we have been steadily expanded upon this language localisation ability and have been creating other types of AI solutions based upon this.

2.1 Example #1: Ongoing pilot of an AI chatbot system to support community health workers in villages

We are now piloting the use of a local language-based AI support tool for our community health workers. These people already have basic digital support tools. We are trying to give them better AI-enabled digital support tools.

Rwanda's population is approximately 14 million people, and just under 10% of these people live in the capital city of Kigali and its major suburbs. The remaining 90 percent of the population live in the smaller towns, villages and remote rural areas.

As part of our healthcare system, for every village and smaller town, there is at least one person who has been trained in first-aid support. While these community health workers do not have the formal education and training at the level of a nurse, they are trained to provide first-aid. In practice, they provide first-aid and the first line of medical advice support in situations where nurses or doctors are not accessible or available. They often support pregnant women during their pregnancy, helping with many of their symptoms. They help address many symptoms displayed by young children. Through practical experience, they get good at assessing if someone is sick enough that they should immediately go see the nearest doctor or if they need to go to the nearest hospital.

These community health workers already have a digital environment in the form of digital access to a manual they use with first-aid and related medical information, and to information on the processes and procedures for how to do their job. To take their digital support to the next level, we are now building and piloting a Kinyarwanda (local language) voice-based AI solution they will eventually use in the field during their patient visits and consultations where the AI system responds to their questions and their descriptions of patient symptoms. The AI tool responds with a listing of more likely diagnostic possibilities

given the community health worker's inputs about the patient and provides information on treatment recommendations related to the identified diagnostic possibilities. In essence, this AI tool with the local language chatbot front end interface is designed to support the community health worker in making a diagnosis and in determining follow up action.

Our ongoing pilot phase for this AI solution involves intensive testing by a controlled number of our community health workers. Over the past 15 years, units under our Ministry of Health have worked with our community health workers to put together and regularly update a manual that is specifically designed to support their needs and circumstances of doing their first-responder and first-aid fieldwork in the village and smaller town settings. The protocols and procedures in the manual are designed to reduce diagnostic and treatment risk given these field workers have not received the more comprehensive education and training of nurses or other certified healthcare professionals, and also acknowledges the constraints that the next level of support - qualified nurses, doctors and hospital facilities - may not be so easily or quickly accessible.

This manual, which has been widely used by our community health workers in the field, serves as the "ground truth" in our pilot testing. The community health workers serving as testers are providing input queries to our new AI chatbot and comparing the responses to the content that is already in the manual. By asking the same questions and following the same protocols that would normally do when using the manual, we can test the AI system by comparing its responses to content already in the "trusted" support manual. We can see when and to what extent the new AI system hallucinates. We can also see how prompting the system in different ways influences the quality and reliability of the output and the degree of hallucination. This pilot testing effort is helping us to assess and characterise the capabilities and limitations of our new AI chatbot. This gives us guidance on where we need to improve the AI system and also where we need to manage and guide the way that the community health workers provide inputs and prompts to the system.

Most of our pilot testers doing these comparisons are sitting somewhere at a table or desk and doing these exercises where they compare the AI system's output to the manual's content. They are not yet actually using the new AI chatbot in the field for real-time support. Because they can take their time to carefully compare the AI system's outputs to what is given in the manual, we have the flexibility to use community health workers of a wider range of experience and skill levels to do testing under these types of circumstances. Even community health workers with intermediate or lesser amounts of experience can do testing in this mode, giving us a larger set of community health workers we can draw upon for this type of testing.

In parallel, as an additional part of our testing effort, we also have some of our more experienced and skilled community health workers using the new AI system in the field while they are doing the actual work of their patient visits. Because they are so knowledgeable, and so familiar with the contents of the manual, they can assess in real time if the AI system's outputs are correct or not, and they know how to quickly cross-check the AI system's output against the manual. They would defer to the content in the manual as the AI system is still under evaluation. Only our most capable community health workers are suitable and skilful

enough for doing this type of pilot testing in the real-time field setting at this phase of our effort. This restricts the number of people who can do the testing in this mode.

Both testing approaches, the off-line desk-based approach and the real-time field-based approach, enable us to stress test the new AI system with realistic case scenarios derived from field experience and with actual field cases. Because we are controlling the size of the pilot through controlling the number of participating community health workers serving as testers in both off-line desk-mode and real-time field-mode, we are able to carefully train them on how to do the testing, and to carefully monitor the testing results. The need to properly train and manage our participating testers and the need to be able to handle and act on the feedback puts a constraint on the size of the pilot effort that we are comfortable managing at this stage of the effort.

We are closely partnering with the Rwanda Biomedical Centre (RBC) in this AI system development and pilot testing effort. They are the medical experts. The national community health programme for villages is part of their Maternal, Child and Community Health Division.[3] Our development and piloting team includes members from the RBC department in charge of the Community Health workers. We interact with experienced doctors and staff from this department on a regular basis, showing them examples of system outputs, discussing with them how the system is performing, and getting their inputs and help.

Through our back-and-forth interactions with RBC doctors and medical staff, we jointly decide where and how to make changes and improvements. This includes identifying ways to improve how the community health workers serving as testers should provide their inputs and prompts to the system, ways to improve input data (the manual and other supporting data we are using), ways to change the AI system itself, as well as ways to change supporting processes and procedures. It is important that we pursue all of these pathways for how to improve the overall performance of using this new AI system.

2.1.1 Stages of piloting across the language localisation loop

Our first stage in our development and pilot effort was to get to the point where we could achieve close to 95% accuracy in translating Kinyarwanda voice inputs into Kinyarwanda text. Our language technology team, including companies we were working with, worked on this internally before we reached out to the community health workers for their involvement in testing. In parallel, we also needed internal efforts (including working with external vendors) to put together the other modules of the AI support system that could take Kinyarwanda text and translate it to English, and then use the English text to do the necessary medical diagnostic assessments.

Once we reached the point where we were achieving 95% accuracy in translating Kinyarwanda voice input into Kinyarwanda text, and we had the other necessary parts of the system ready that were necessary for medical diagnostic support, we started our engagement with the community health workers for the next stage of user-based pilot testing. Even though we knew that a 95% accuracy level in converting local language voice inputs into local language text would introduce errors, our thinking was that this was “good enough” as a

starting point for them to start testing the system. Using their human language understanding capabilities and their background knowledge of the subject matter, they could compensate for or correct the errors resulting from the imperfect conversion (the approximately 5% error) in going back and forth between Kinyarwanda speech and text.

With the local language voice interface ready (or ready-enough) to use, community health workers could start pilot testing the overall outputs of the end-to-end AI support system that includes all six steps illustrated in Figure 1 below.

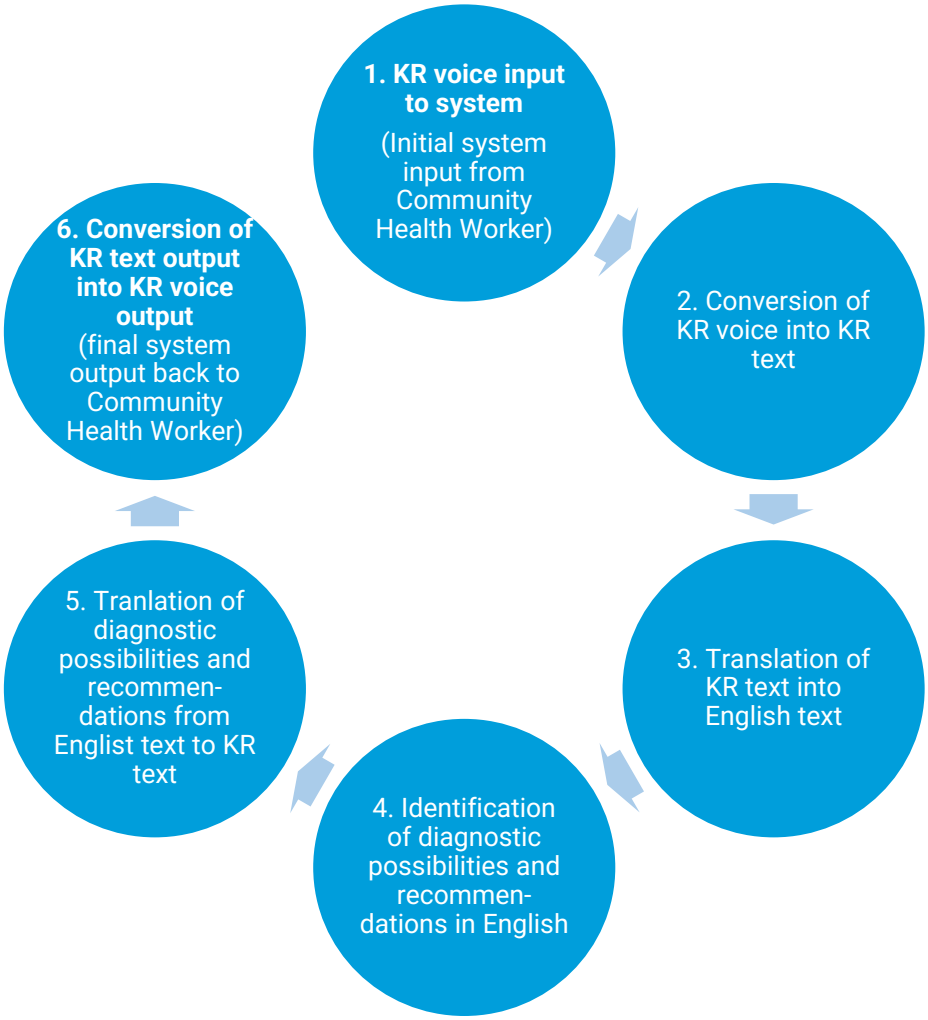

Figure 10: End-to-end language localisation loop for the AI chatbot system supporting Community Health workers in Rwanda's villages

For the community health workers to do their fieldwork in the village and small town settings, they often need to communicate with the AI support system via voice in the local Kinyarwanda (KR) language to provide inputs (Steps 1). For the AI system to perform its support tasks of using the inputs about the patient's symptoms and condition to identify the more likely diagnostic possibilities and to provide treatment recommendations for responding to each possibility, it needs to work with text (and sometimes with images),

requiring the conversion between KR voice input into KR text input (Steps 2). Internationally, as medical information about disease diagnosis and related treatments is most widely published and distributed in English, the KR text often needs to be converted to English text (Steps 3) so that the AI system can do the search and matching with the likely diagnostic possibilities and the treatment recommendations for each possibility (Step 4). This English based information then needs to be translated into KR text (Step 5) which then needs to be converted into KR voice so that the Community Health worker (and sometimes the patients they are supporting) can hear the response (Step 6) so that the output is accessible in "hands-free" mode and also accessible to those who are not able to read the text output.

We are now in the process of assessing the performance of the end-to-end AI support system (Figure 1, Steps 1 through 6). Though we are not yet finished this current phase of pilot testing, we are already anticipating additional types of questions we will need to address as we proceed with this effort that will require subsequent phases of pilot testing:

i. To what extent do we scale the use of this tool across the much larger base of community health workers supporting the large number of villages across the country? And how would we do this?
ii. To what extent should we integrate this AI chatbot system for supporting community health workers in village settings with other systems and processes that are part of country's larger national healthcare system? And how would we do this?

We have not yet reached the point where we are ready to make these decisions. We will continue with the current pilot testing effort, evaluate the results, and based on that experience, "build our plans as we go", and iteratively figure out how to address these two questions stated above.

2.1.2 Risks to manage as we proceed with further piloting and scaling of this AI chatbot solution for supporting community health workers in villages

As this is a healthcare application, we need to be concerned about the precision of the chatbot's diagnostic suggestions. Rwanda's national and public healthcare efforts have done really well in reducing both maternal mortality and infant mortality over the past two decades. Our ability to properly diagnose disease, especially for malaria, is of great importance because if we are able to do this diagnosis early and encourage people with malaria to get to the health centre early, then they get the medicine, and this reduces malaria related death. This also applies for any other treatable disease that is prevalent in our country.

With this in mind, we need to be careful in our considerations of scaling the use of this AI chatbot solution. A key risk we need to be mindful of and avoid is a scenario where scaling the use of this diagnostic chatbot ends up reducing the effectiveness of our community health efforts and reversing progress to date in addressing health challenges in the villages. How could this happen? There could be scenarios where our community health workers come to overly rely on the AI chatbot's diagnostic outputs and recommendations and are not careful

enough in verifying the responses, or do not even realize there may be errors where they need to verify the response.

As noted above, there are multiple steps of language format conversion and translation involved in the end-to-end AI support system process (as shown in Figure 1). This introduces risk as each language format conversion or translation leads to more possibilities for the AI system to make errors. There is also the source materials we are relying on. This includes our local diagnostic manual that we have developed over the years and that we are now using to test the AI system. While we believe this is high quality, accurate information that is contextualized for our local Rwanda setting, it is not perfect. There are also the other sources of medically based diagnostic and treatment information from various English-based international sources that are used to train the AI chatbot. Most of this international information is not specifically contextualized to our local Rwanda setting.

These factors contribute to error rates in the diagnostic responses, which is why we need to carefully navigate the path of finding ways to harness the useful capabilities and advantages of using the AI chatbot while appropriately managing the risks.

This current stage of our piloting is purposely done at smaller scale with a relatively small team of community health workers serving as testers. It is also a relatively small team of others involved in the details of supervising, overseeing, and evaluating the pilot and using the interim pilot feedback to guide improvements. Because of this carefully controlled smaller scale, we can have very close collaboration across all members of the pilot team and we are also able to carefully prepare and continuously support the community health workers serving as our testers.

Suppose that based on the results of our current pilot, we eventually decide to proceed to a next phase where we attempt to take the next step towards expanding the scale of our efforts. We would need to rethink many aspects of how piloting would be done in a larger scale next phase. As we incrementally scale, we would have to work out the appropriate ways of providing the necessary support to a larger pilot effort, including determining the kind of training and support required for a larger pool of community health workers who would be using and testing the system.

Each incremental step toward scaling will introduce or reveal new risks in the overall process. We would have to think critically and carefully about how we would change our piloting process with each successive iteration of scaling to address and manage these new types of risks. One implication is that piloting efforts for this type of AI application cannot be thought of as a one-time effort. With each step towards scaling to higher numbers of community health worker users and broader usage across more village settings, we will need to do additional stages of piloting designed to carefully monitor the new areas of risk associated with larger scale usage.

2.2 Example #2: Piloting and deployment of an AI chatbot to support customer service complaint response across the banking sector

In our financial sector, our central bank has implemented a sector wide customer complaint chat portal as part of an initiative to improve customer service across the banking sector.[4] They are now using this chat portal to centralise the reporting and tracking of complaints that banking customers have with their banks. The central bank created, tested and deployed this customer complaint chat portal in consultation with the banks and then asked all the different banks they regulate to integrate this complaint portal into the bank's own website in a visible and prominent way. While this chatbot for handling complaints was initially created for the banking sector, its usage has subsequently been expanded to include other parts of the financial services industry that are also regulated by the central bank.

If someone with a bank account in Rwanda lodges a customer complaint with their bank, the complaint ticket created by the customer's interaction with this chatbot is also tracked at the central bank. This means the central bank sees how long it takes a certain bank to close a particular customer complaint ticket. They see this for all of that bank's customer complaints, and they also see this across all of the banks they regulate.

Prior to the deployment of this new approach enabling the central bank to track of each bank's customer complaints, banks were often perceived as "taking their sweet time" to get back to a customer about a complaint. That has changed as a result of the introduction of this new complaint tracking system. Our banks are now much more responsive to complaints. Shortly after a complaint ticket is opened, banks are following up with the customer by phone call, messaging or email to ask about the complaint and the problem. The banks are now responding rapidly to make sure they are closing that ticket as soon as possible, often in less than 24 hours.

Previously, the central bank had specified service level agreement (SLA) requirements for responding to customer service complaints as part of their process of issuing and periodically renewing bank licenses. Even though the central bank had such SLA requirements in place, they did not have the visibility to observe what was actually happening. When each bank would provide reports on there adherence to agreed upon service level standards for responding to complaints, they could always assert that they in compliance. There was no direct way for the central bank to measure if that was really so. The central bank could occasionally sample how banks responded to complaints, for example by hiring "mystery shoppers" to make complaints and follow happens, but that was not viewed as an adequate or scalable monitoring approach.

This new customer complaint system is automated and is the same across all banks. When the customer makes a complaint, they can see that the central bank as well as their own bank is aware of this newly submitted complaint ticket so both the customer and the bank know that they central bank is aware of the problem and monitoring the response time. This gives the bank a strong incentive to respond quickly. The bank knows that when they have their next discussion with central bank regulator about their license renewal, or about

any other topic, the central bank now has clear and reliable statistics on how that bank is (or is not) adhering to their own service level agreements for customer service.

2.2.1 The bigger strategic importance of more rapidly resolving the smaller customer complaints

Most of the customer complaints received by the bank are not that complex. Most complaints just need someone to answer them, and they can be solved immediately or very quickly. A much smaller number of complaints involve more complex issues and need to be dealt with over a longer time period. Suppose you have 500 people who issue complaints, and 495 of them are very small matters that are very easy and fast to resolve, but it takes the bank a week or even substantially longer to get around to responding to them. This gives the impression that the bank does not have a good internal operational process for handling the complaints. It hints there may be an even bigger issue happening with the bank's internal systems and with the broader banking system.

This is why the central bank thought it was important to implement this centralised chatbot for customer complaints as part of trying to create a new image of an effective financial sector where services are always up 99% of the time. If most of the routine queries and related complaints can be answered immediately or very rapidly, then only the much smaller number of more complex complaints remain (e.g., the 5 of the 495 complaints in the example above), and the bank can take the necessary longer time period to solve them.

2.2.2 The central bank's process of collaborating with the banks to design and implement the centralized AI chatbot for customer complaints

The central bank created the initial versions of this AI chatbot for receiving and processing customer complaints with the support of an external vendor. Internally, they worked with the system to understand how it responded and what it looked like, and they created the initial chatbot prompts and dialogues.

After this initial round of internal pilot testing, they engaged the banks they regulated to get involved. The banks provided many inputs as part of helping to review and revise the system. They providing guidance on the flow of prompts for different dialogue scenarios, on how to answer questions, and on the types of information needed to give appropriate answers. They also helped to review and evaluate chatbot responses. This ended up being a lengthy and detail process of consultation with the banking sector.

When the central bank had a product that they thought was good enough for external customer usage, they put it on their own website so members of the public could try using it. In parallel, they gave all the banks a transition period deadline, giving each bank the time to integrate the centralized customer complaint chatbot portal into their own IT system platforms and into their own website. Once the banks completed this period of integration and testing within their own environments, the central bank joined together with all the banks to launch a common, sector wide outreach and awareness campaign to inform all customers

of all banks across the country about this new customer complaint chatbot portal so they would know that this new tool exists, what is for, and how it is used.

2.2.3 Ongoing refinements to the customer complaint chatbot

Even after its public release, the customer complaints chatbot continues to be refined and improved by the central bank team responsible for this application. In the earlier phases of the project, an external vendor played a key role in the design, development and implementation of the chatbot system. Post-deployment, as the usage of the chatbot has increased, the central bank has been gradually building up its internal capacity to maintain and support the sys and to incrementally improve it.

3. The evolution of how Ministry of ICT and Innovation works with the other ministries on digital transformation and AI efforts

Within MinICT, our view of digital transformation has evolved. When we started doing our digital transformation efforts across various parts of the government years ago, there was a need for our ministry to do many efforts to demonstrate to the other ministries the usefulness of having digital and digitised services. In that earlier phase, most of the development and/or project management was directly done within MinICT or within its affiliated agencies.

This situation has evolved. As of several years ago, we have been putting more of our emphasis on supporting the other ministries, helping them in their efforts to frame, plan and concretise their digital and AI initiatives. While each ministry is at a different capability level with respect to their digital and AI capabilities and experience base, to the extent possible, they have been taking on more of the leadership for identifying and driving projects.

The vision and rationale for why a new digital or AI initiative should be undertaken needs to come from the respective ministries. Staff and officials within each ministry need to be able to explain how an envisioned application of digital or AI answers to or supports a key pain point within that ministry, or within the sectors of the economy under the jurisdiction of that ministry. For example, within the Healthcare ministry, if a key goal is to build a healthier Rwandan population and to keep improving the condition of health across the population, each digital initiative put forth by the ministry should align with and “answer to” this type of strategic goal statement.

Within MinICT, we continue to take the leadership role on important horizontal (cross-ministry and cross-economic-sector) digital and AI initiatives that benefit multiple ministry efforts and national initiatives. We also take the lead on emerging technology demonstration projects that are strategic for the country.

Across the various MinICT portfolios and initiatives, we therefore work hand-in-hand with the other ministries helping them with 1) planning and implementing the digital and AI goals that are specific to own ministry, 2) their involvement with the horizontal cross-ministry

Digital/AI initiatives, and 3) their involvement with the emerging technology demonstration projects. Through these various types of interactions with our other ministries, we help them to accelerate their rate of progress towards achievement of both their ministry level as well as national level digital and AI goals.

The digital and AI journey is completely different for each of the ministries and their associated economic sectors. On one end of the spectrum, we have a few ministries that have already fully embraced digital transformation and related technology efforts. They are already quite experienced with such projects and in some cases, they have built up their own internal teams and external ecosystems to support their efforts. For these types of ministries, we serve as a resource and consultant to them in a 50:50 type of equal partnership. They bring on board ideas. We also share ideas and give them feedback.

On the other end of the spectrum, there are some ministries that are just getting started on this journey. For these types of situations, we go to them, make pitches, help them to plan how to get started with solutions that are important to their ministry's mandate, and provide them with strong support for building and demonstrating digital and AI efforts. Then there are other ministries that fall between these two ends of the spectrum. The key point is that the nature of our engagement with each ministry is quite different depending on where they are in their journey with digital, AI, and other emerging technologies.

Our ministry has put in place a framework for having a Chief Digital Officer resident within each of the ministries across the government.[5] This is helping all the ministries, whatever their digital capability level, to move forward with their digital and AI efforts. Having a Chief Digital Officer within each ministry is a good mechanism for facilitating and accelerating cross-ministry experience sharing, and for improving MinICT's ability to support and align with the digital and AI initiatives of the various ministries.

3.1 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

It is good that the ministries across the government are increasingly stepping up (though at different speeds) to take a larger leadership role in identifying their respective digital, AI and other innovation and emerging technology needs and to more actively participate in leading these types of projects. At the same time, across the government, we need coherence and consistency in our approach to piloting, risk management, policy experimentation, and adaptively designing and implementing appropriate regulations across our entire national portfolio of digital, AI and related innovation efforts.

This raises an important question for MinICT. How should Rwanda combine a centralized and coordinated approach for guiding how piloting, policy experimentation and regulation is done related to digital and AI efforts, while at the same time encouraging each of our ministries and the vertical sectors they represent to increasingly take ownership for moving forward with their respective sector specific efforts?

While we are encouraging each of the ministries to take ownership of their respective digital transformation and AI efforts, we still require that each ministry needs to fully involve MinICT at every stage of the process. Through various formal and informal mechanisms, MinICT - including our affiliated agencies - has what amounts to a "green light capability" to review and approve digital and AI solutions proposed by other parts of the government. When we give that green light approval, and a particular ministry proceeds with develop and implementing the effort (as per the two examples discussed above from our Ministry of Health and our national bank/Ministry of Finance), MinICT also closely follows the effort. We have teams within MinICT and our affiliated agencies tasked to follow, and as needed, to support these efforts.

Another mechanism that helps us with following, supporting and coordinating the many digital and AI efforts across the government is the Chief Digital Officers (CDOs) scheme that we started putting in place as of the beginning of 2021 (mentioned above). These CDOs are sitting within each of the other ministries as part of their senior management teams, and also reporting back to MinICT via a joint reporting line to our digital technology implementation agency.

What we learn from these CDOs, and through our other mechanisms of collaborating with the ministries, helps MinICT to work towards achieving a greater degree of coherence and consistency in our approaches to piloting, risk management, policy experimentation and regulation.

It is important to keep the following in perspective. For most ministries, for many of their projects, they are implementing already proven digital or AI solutions that we are already quite familiar with in terms of technology, risk management issues and regulatory needs. It is a much smaller proportion of projects where the situation involves a new and unfamiliar situation and where we need to invent the policy approach and regulatory needs as we go through learning-by-doing experimentation.

4. Rwanda's national AI policy and related efforts

We published our National AI Policy document in 2022 and this policy document was approved by our Prime Minister and Cabinet in April 2023.[6]

This policy is designed to support the adoption of AI and related digital services within both the public sector and the private sector. When we think of AI adoption in the public sector, we are trying to understand the value addition of using AI for how we do business on a daily basis within government and for delivering government services.

We set a goal of having all Rwanda government services online by end of June 2024 and we are almost there (as of this interview on 31 May). We are almost at the finish line and trying to finalise getting the last few government services online.

With 100% of our government services available soon to be available online, that enables us to look across our entire portfolio of government services and to understand and assess those services where we would most need AI.

Two years ago, we did an economic impact study on AI for Rwanda. We looked at Rwanda's current and forecasted economic activity as of that time and at some of the existing AI solutions on the market. We assessed if there were existing AI solutions that could be adopted for our needs, especially for the public sector. We have a long list, sometime we refer to it as a "laundry list", of all these various solutions that we would like to eventually test and see if they are suitable for our needs.

Once we complete the current pilot testing phase of our chatbot for supporting our community health workers (described earlier), then we will start ramping up our efforts to revisit that prior AI economic impact study. We will do the necessary updates related to our national priorities by economic sector and to our list of AI solutions available in the market corresponding to those economic priorities. Based on that, we will further prioritize which projects to focus on and initiate additional piloting and implementation projects that use AI-based capabilities. Part of this effort will be to get both the public sector and private sector to increase their adoption of AI solutions over the next several years, and to enhance public sector-private sector collaboration to make this happen.

4.1 Commercial Cloud

We are actively trying to understand the issues related to using the commercial cloud. We are figuring out where and how we can access larger commercial cloud data centres for both public and private sector needs. We believe that cloud will play a big role in helping us reduce our cost of compute for both our public and private sectors, and that access to cloud resources for compute, as well as for software components, applications and other support resources, will have an important enabling impact on our ability to move forward with our AI plans.

Our National AI Policy document approved by our cabinet in April 2023 states "We need to provide mechanisms to enable access to international and world-class cloud computing services which offer competitive performance and cost for Rwandan companies and the research community, in order to facilitate and fast track AI adoption." That same document also says that we will create the necessary critical infrastructure through both partnering with global players and also through building local infrastructure. We are looking within the country as well as across the region to assess the situation as part of figuring out how to move forward with providing both our public sector and private sector with access to cloud-based infrastructure and ecosystems. We don't know yet what is possible. As we proceed with this, we will also have to work out the policies and processes for the governance of using the commercial cloud.

5. "Big Picture" challenges as we continue moving ahead with digital transformation, AI and other emerging technologies

Given my portfolio and responsibilities within MinICT, I summarise my three "big picture" challenges in the following way.

For government policymakers and regulators, our first big challenge is just keeping up with the technology advances so that our own internal teams have sufficient understanding to determine how to adapt policy and regulate as we go forward. This requires that we upskill our own staff members and senior officials so that we have the understanding we need to do this.

The second challenge relates to navigating the complex geopolitical landscape (regionally and globally) for determining our sourcing and partnership strategies for cloud access and for the other enabling aspects of AI capabilities and applications. Rwanda is a small country in terms of population and financial resources. We have to have close collaboration with other countries to continue making progress with our digital transformation, AI and overall innovation efforts. Our continued ability to create meaningful partnerships and appropriate synergies for our private sector as well as for our government is going to be very important. Also, as the conversation on the governance of AI usage becomes increasingly globally, it becomes very important for us to make sure that we have the collaborations, partnerships and ecosystem participation to voice our considerations for what small countries like ours might face as the technology and its applications evolve.

The third big challenge is how fast can we develop our own within-country ecosystem? We need to further develop our country's ability to do relevant research, to make use of innovations, and to link these capabilities with our AI and digital transformation efforts. We need the market, the private sector and academia working together with the government to achieve this and to determine the appropriate government, private sector and market roles for how to make this happen. We need to determine how we invest in creating and accelerating this ecosystem. It is crucial to make sure that we are making progress towards this type of ecosystem.

6. Making Rwanda a proof-of-concept hub for national scale piloting and "learning-by-doing" policy design

We know the technology will continue to evolve much faster than our government can adapt its policies and regulations. This is especially true with everything related to AI. The gap between the speed at which the technology develops versus the speed at which policies and regulations can be adapted or newly created will only continue to increase.

After observing and thinking about this for several years, one way we are positioning for this is a strategy of making Rwanda a proof-of-concept hub. That means we actively work towards creating an environment where we can quickly and safely test emerging technology solutions. A good example of this is what we did together with Zipline, the US start-up

company for drone-based delivery and logistics.[7] In 2016, they started their first field-based company operations by demonstrating their ability to deliver blood packs in Rwanda. This was the world's first example of operating a regular commercial drone delivery service.

We have done other significant demonstration projects with various companies to test the usage of other types of emerging technologies. We now understand what it means to test out something that is not yet operationally tested in any other market or where there is not yet a global consensus on what are the best deployment and operational practises, or what are the important risks that need to be managed. Because we have been through this cycle several times, we understand how to manage and operate these types of efforts, and we understand how to approach the unknown risks.

In areas related to innovation, emerging technology and new types of use cases, we are practicing a "learning-by-doing" approach to policy making and regulation. We know that when we are pilot testing under new circumstances, we will not initially have appropriate policies or regulations in place that would be appropriate for governing something new.

We use the piloting and testing efforts to build up our understanding of the initial approach to regulation that seems appropriate. We are open enough to adapt and improve the regulations as we gain field experience from larger-scale pilot trials and the various stages of operational usage as we deploy at scale. Our goal is to make sure that our citizens are well served by whatever new digitally enabled services that we put in place and that we eliminate or reduce the risks of harmful occurrences.

We are practicing and refining this learning-by-doing approach to our digital and innovation policy making and regulation, including for AI, in a way that meets the needs of our situation in Rwanda. In my view, this strategy of combining a systematic approach to operational piloting with an adaptive (learning-by-doing) approach to policy and regulation is a good way to move forward because the speed of innovation is always faster than the speed of doing policy making and regulation.

For every emerging technology, including current AI developments, it takes time for our policy makers to develop the necessary understanding of the new technology and of how it is used across industry. This learning-by-doing approach is a good means of bridging this gap. It is a good way to develop the policy making talent we need within the public sector to manage these new developments, applications and use cases. Move forward in this way helps us to develop our public sector talent base that we need to manage these efforts.

It is important to keep the following in perspective. For most ministries as well as the private sector, for most of their projects, they are implementing already proven digital or AI solutions that we are already quite familiar with in terms of technology, risk management issues and regulatory needs. It is a much smaller proportion of projects across the ministries and the economy where the situation involves a new and unfamiliar situation and where we need to invent the policy approach and regulatory needs as we go through learning-by-doing experimentation.

7. The role of Carnegie Mellon University Africa in building up Rwanda's manpower and ecosystem for digital and AI innovation

CMU Africa was established in 2011 in Kigali.[8] The CMU-Africa website states that this is the only U.S. research university offering its master's degrees with a full-time faculty, staff and operations in Africa. In 2019, CMU Africa moved into a new location in Kigali Innovation City.[9]

It has been a very interesting journey with CMU. The size of the very first cohort that graduated in 2014 was 22 students and they were all in one programme, the MS in Information Technology (MSIT).[10] By the time of the 5th cohort in 2018, there were 45 graduates in total from two masters programmes, the MSIT and the MS in Electrical and Computer Engineering (MSECE), and they came from six different nationalities.[11] By the time of the 10th cohort, there were 158 graduating students in total across three programmes (MSIT, MSECE and the first graduates of the new MS in Engineering AI Systems (MSEAI), coming from 19 different nationalities.[12] In May 2024, they recently held the graduation ceremony for the 11th cohort and the size of the graduating cohort continues to increase.[13] Currently, according to the CMU Africa website, they have over 300 students currently enrolled across their three masters programmes, and they have produced over 550 alumni to-date.

This is Pan-African talent. Considering both the students who are currently enrolled and as well those who have already graduated, this talent come from more than 30 countries across Africa. The quality of education of CMU Africa has helped us in a couple of important ways. First, it has strengthened our ability to attract a talented African workforce to Rwanda because a growing number of people are coming here to do these masters programmes.

Second, this in turn helps us in our investment attraction capabilities. In our efforts to attract companies in the tech sector as well as companies in other sectors that require the type of technology talent produced by the CMU programmes, we can assure them that they are going to have a reliable supply of talent coming straight out of CMU Africa. As a growing number of the CMU Africa graduates come from other African countries, as a government, we promise that we will work hard to make sure that these graduates have the right work permits to work in or from Rwanda after graduation.

Third, CMU Africa is contributing to the Rwanda and the Pan-Africa R&D ecosystem. CMU Africa has very good faculty who are living here in Rwanda. We can already see the steady ramp up in research activity, capability and output. This steady progression of CMU Africa R&D efforts will play an increasingly important role in building up R&D capability in Rwanda and more broadly across the region.

This will be very helpful to our future efforts to create and apply emerging technologies. It will also strongly support our goal of nurturing Pan-African faculty members. A small proportion of the CMU Africa students are getting seriously interested in going into research. This would contribute to our national and regional pool of well trained researchers and also to the pool of faculty candidates. A few years from now, we will see more of the

effects of good research coming out of CMU-Africa and more examples where that research is done in collaboration with our other local universities, as well as with other universities across the African continent.

The growth in the size and activity levels of CMU Africa will give us a larger and stronger manpower pool, more people well equipped to become tech entrepreneurs, and much richer research capability. All these dimensions will contribute towards our ability to realise our National AI Policy aspirations and our ability to contribute to questions and challenges that are important to the African continent as a whole. As the size of the CMU Africa student body increases towards a total of 400 enrolled at any one time across the various programmes (whereas now it is currently around 300 in total), it will be helpful to our national and Pan African efforts.

7.1 Engaging the innovation ecosystem through CMU Africa student internships and practicum projects

Some of our Chief Digital Officers deployed across the various ministries take on CMU Africa students for summer internship and for supervising practicum projects that are part of the curriculum.

What is especially interesting for us in MinICT to observe is that Rwanda's private sector is increasingly "waking up" to the possibilities and benefits of engaging with CMU Africa through taking on the students as summer interns. A small but growing number of our private sector companies are structuring this internship opportunity as part of their proper recruiting process. When these students come into the private sector company or into a public sector department for the summer internship, it has turned out to be a good way for that organisation to quickly pilot an innovative idea. This helps our national effort to build innovation ecosystems and to increase the number of people who can be part of those ecosystems.

8. Suggestions for other small countries moving ahead with digital transformation and AI

8.1 Suggestion #1: The importance of contextualisation

A key lesson I have learned from my current and prior professional roles (all based in Rwanda) it that it is always very important and worthwhile to take the time to contextualise any type of solution that comes into the market. Especially for small countries, take that extra time to think about and experiment with a new solution in order to understand what it means, and to see how it compares to needs and circumstances for your country. This will help you to determine how to best contextualise the technology and related application systems.

8.2 Suggestion #2: Utilizing the advantages small countries have with piloting emerging technologies

The second suggestion is to be aware of the very exciting possibilities for doing pilot experimentation and sandboxing with AI and other emerging technologies in small countries. Even when you pilot something at a smaller scale relative to the need for scale in a larger country, the local impact in the small country can be very big.

A good example is the pilot testing that we started with Zipline back in 2016 that I mentioned earlier. This drone delivery service for emergency medical logistics has expanded beyond the delivery of blood packs to now also included special drone deliveries of essential medicines, diagnostic test kits and vaccines. Because of Rwanda's smaller geographical size, Zipline delivery drones can reach nearly 80% of our health centres and hospitals across the country.[14] This range enables deliveries to nearly all of our healthcare facilities outside of our capital city Kigali. Another practical advantage related to our small size is that all drone operations can be supported with only two ground-based centralised distribution centres, the first one established in 2016 in Muhanga near the centre of the country, and the second one established in 2018 near Kayonza, in the eastern part of the country.

Because the drone service can reach healthcare facilities across the entire country (outside of the capital city), any healthcare centre in Rwanda that needs an emergency delivery of blood can get it within one hour. This ability to provide full national coverage across all of our rural and town areas outside the capital has had important and lifesaving impacts. Larger countries in Africa and in other parts of the world have piloted drone delivery services for emergency medical supplies. However, they were not able to achieve the high degree of national coverage that we are able to achieve and do so with so few supporting ground-based distribution centres.

Because of our small scale in terms of both geographic and population size, we can set up pilot testing efforts that can more quickly achieve a high degree of national coverage, of course with attention to proper precautionary measures. This allows us to learn lessons very fast, including how to manage risks. This in turn enables us to make the necessary changes and improvements very fast. When you are a much larger country and require much larger scale pilot testing efforts to do trials with significant coverage or participation levels, it is much more difficult.

This is why as a small country we have an advantage for the pilot testing of emerging technologies that are contextually appropriate for our needs in Rwanda. This also relates to my point made earlier as to why Rwanda can be a proof-of-concept hub for national scale piloting of innovation efforts, and why we need to be quick in our ability to learn how to do the relevant policy making and regulation.

Endnotes

[1] Rwanda Ministry of ICT and Innovation (MinICT) website: https://www.minict.gov.rw/.

[2] R Mohire, Mozilla Foundation, How Rwanda is Making Voice Tech More Open, 16 September 2020, https://foundation.mozilla.org/en/blog/how-rwanda-making-voice-tech-more-open/. Also see:

- Mozilla Foundation, In Rwanda, Voice Technology Innovation Helps Fight COVID, 02 September 2021, https://foundation.mozilla.org/en/blog/in-rwanda-voice-technology-innovation-helps-fight-covid/.
- Kathleen Siminyu, Mozilla Foundation, Lessons from Building for Kinyarwanda on Common Voice, 05 April 2022, https://foundation.mozilla.org/en/blog/lessons-from-building-for-kinyarwanda-on-common-voice/.

[3] Rwanda Biomedical Center (RBC), Maternal, Child and Community Health Division overview webpage, https://rbc.gov.rw/index.php?id=672

[4] National Bank of Rwanda, Introducing INTUMWA Chatbot, October 2022, https://www.youtube.com/watch?v=xjKxqVNyLLY

[5] MinICT, THE TASK THAT AWAITS GOVT'S NEW 'DIGITAL LEADERS', December 2020, https://www.minict.gov.rw/news-detail/the-task-that-awaits-govts-new-digital-leaders

Rwanda Information Systems Authority, Chief Digital Officers (CDOs) celebrate their impact in Rwanda's Digital transformation journey, 24 June 2024, https://www.risa.gov.rw/news-detail/chief-digital-officers-cdos-celebrate-their-impact-in-rwandas-digital-transformation-journey

[6] Republic of Rwanda, Ministry of ICT and Innovation, "The National AI Policy," 2022, https://www.minict.gov.rw/index.php?eID=dumpFile&t=f&f=67550&token=6195a53203e197efa47592f40ff4aaf24579640e

[7] Ministry of ICT and Innovation, RWANDA SIGNS AGREEMENT WITH ZIPLINE TO USE DRONES FOR DELIVERY OF ESSENTIAL MEDICAL PRODUCTS, February 2016, https://www.minict.gov.rw/news-detail/rwanda-signs-agreement-with-zipline-to-use-drones-for-delivery-of-essential-medical-products
Gavi (the vaccine alliance). Also see:

- Rwanda launches world's first national drone delivery service powered by Zipline, 14 October 2016, https://www.gavi.org/news/media-room/rwanda-launches-worlds-first-national-drone-delivery-service-powered-zipline
- ITU, How medical delivery drones are improving lives in Rwanda, 24 April 2020, https://www.itu.int/hub/2020/04/how-medical-delivery-drones-are-improving-lives-in-rwanda/
- K Korosec, Tech Crunch, Zipline is now the national drone service provider for Rwanda, 15 December 2022, https://techcrunch.com/2022/12/15/zipline-is-now-the-national-drone-service-provider-for-rwanda/

[8] Carnegie Mellon University Africa, https://www.africa.engineering.cmu.edu/

[9] Carnegie Mellon University, CMU-Africa Celebrates New Location, 21 November, 2019, https://www.cmu.edu/news/stories/archives/2019/november/cmu-africa-celebrates-new-location.html

[10] Carnegie Mellon University, Carnegie Mellon University in East Africa Graduates First Class, 24 July 2014, https://www.cmu.edu/news/stories/archives/2014/july/july24_cmurwandafirstgraduates.html

[11] T Moore, Carnegie Mellon University, CMU-Africa Graduates Poised To Engineer the Future (5th graduating cohort) 05 June 2018, https://www.cmu.edu/news/stories/archives/2018/june/cmu-africa-graduation.html

[12] M Sumbi, Carnegie Mellon University Africa, CMU-Africa celebrates 10th graduation, 13 June 2023, https://www.africa.engineering.cmu.edu/news/2023/06/13-cmu-africa-graduation.html

[13] M Sumbi, Carnegie Mellon University Africa, 11th graduation ceremony celebrates innovation and excellence, 30 May 2024, https://www.africa.engineering.cmu.edu/news/2024/05/30-graduation.html

[14] M Mhlanga, T Cimini, M Amaechi, C Nwaogwugwu, A McGahan, From A to O-Positive: Blood Delivery Via Drones in Rwanda, report published by Reach Alliance, Munk School of Global & Public Policy, University of Toronto, April 2021, https://reachalliance.org/wp-content/uploads/2021/03/Zipline-Rwanda-Final-April19.pdf. Also see:

- T Amukele, Using drones to deliver blood products in Rwanda, Comment in The Lancet Global Health, Vol 10 (4), April 2022, https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(22)00095-X/fulltext.
- HH Jeon, C Lucarelli, JB Mazarati, D Ngabo, H Song, Last-mile Delivery in Health Care: Drone Delivery for Blood Products in Rwanda, working paper on SSRN,12 October 2022 (revised 03 May 2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4214918.

INTERVIEW 9: Dominic Chan, GovTech Assistant Chief Executive and CIO, Singapore

Date of Interview: May 10, 2024

1. Dominic Chan's role and background

Dominic Chan works for the Government Technology Agency of Singapore known as GovTech.[1] He has two roles. One role is Assistant Chief Executive for government digital products. He oversees the product management side of GovTech which includes overseeing all the product management teams that build and deliver products for the various parts of government.[2]

Concurrently, he serves as the GovTech Chief Information Officer (CIO) where he oversees planning and execution efforts for GovTech's internal digital transformation initiatives.

In the last five years, he was the senior product manager for three major GovTech products families: National Digital Identify, Moments of Life and the Trace Together Token developed for Covid-19 contract tracking as an alternative to using the Trace Together mobile phone app.

2. My understanding of the purpose and meaning of AI policy

AI is not something new, although there has been an increased emphasis on in recently. AI has been in the background and part of how digital technology has been evolving for a long time. The recent advancements in AI have brought a lot more prominence to the term and to the related capabilities, and more concerns regarding how AI usage may impact our lives and impact our online security and privacy.

When people talk about policy for AI, they often focus mostly on precautionary measures to protect citizens, to protect data and to protect people and ecosystems at large.

My personal opinion is that focusing AI policy mostly on precautionary measures and restraints is not a very fertile approach because whether you like it or not, the technology is going to be moving forward and be used more widely. Therefore, the question we need to address is how to have appropriate policies to govern how it is used that include reasonable ethics around how it's used.

As we move forward with greater usage of AI in our government services, we need to be extremely aware of the direct and indirect downsides of using AI as well as the upsides. We need to be very proactive in developing an understanding of risks and of the things that can be potential threats, and work on ways to manage and counteract these types of issues. Simultaneously, we need to be proactive in enhancing our understanding of beneficial uses of AI. It is always a trade-off. This is how I think about AI policy.

3. Sandboxing, piloting and experimentation - including policy experimentation - are a regular part of our product management efforts

In my product management roles for various government digital service products, sandboxing and experimentation are important steps of what we usually do as part of product management. Simply put, there are a lot of uncertainties and unknowns with anything that we set out to build (whether or not it includes the use of AI). As such, as part of our product management process, we experiment all the time.

When we do the software and related digital technology development for creating a new government digital service, we initially create a Minimally Viable Product (MVP) as part of our early-stage experimentation and testing efforts. We are systematic about explicitly testing what we believe are the riskiest assumptions and key hypothesis underlying our assumptions. Given our systematic approach to product management, sandboxing and experimental testing happens all the time as part of every product and service we create. Our regular efforts for experimentation and testing as part of product management sometimes also requires that we explore how the usage of a new digital service (which may or may not include AI enablement) aligns with or pushes the boundaries of existing government policies.

Given what we regularly do as part of product management, I do not look at experimenting (including policy experimentation) and related sandboxing with new digital products and services that make use of AI differently. I view it as part of our ongoing and always evolving approach to product management that is based on experimenting with MVPs, piloting within both internal and external sandbox settings, continuously evaluating feedback, and iterating to make improvements related to risk management, policy compliance, operational considerations, usability and experience, and technical improvements. We will sandbox both internally and externally as we deem needed to test for problems and negative implications and for how to counteract or address such issues.

Whether or not the new digital service makes use of AI, there may be risky assumptions and underlying hypotheses that relate to policy compliance or relate to determining whether the digital service has aspects that may not even be covered by an existing policy. When we identify such situations, then we must do that type of policy related experimentation and testing as well.

While policy-focused experimentation and sandboxing is a regular part of our overall product management effort, this needs to be put in perspective. Of all the different aspects of experimentation, testing and validation that we do as part of piloting and rolling out a new government digital service, specific efforts for policy related experimentation and testing do not happen as often or as regularly as other aspects of our experimentation and piloting efforts. However, when we need to also include policy related experimentation and piloting, we do so as part of our product management efforts and we go about it in a way that is no

different from all the other hypothesis that we test. This summarizes our product management perspective on AI experimentation, piloting and testing.

4. Identifying the risky assumptions and underlying hypotheses that need to be tested and evaluated

If there is a new government digital service that makes use of AI, we should go through the same exercise of experimentally testing and evaluating what we deem to be the riskiest assumptions and the key hypotheses underlying our assumptions.

AI does not change this process that we already have as part of our product management. What AI might do is to add new scenarios for what may be the riskiest assumptions that should be tested and similarly for new hypotheses that are underlying our assumptions. This means that our challenge related to AI experimentation and sandboxing is to understand how the AI capabilities – including supporting processes for enabling AI - incorporated into the new digital service we are developing may lead to new types of uncertainties and related risks that we need to understand and test.

Also, there may be some specific governance, risk management and compliance requirements that need to be followed when using AI based systems that are based on our own national frameworks or on international ones that we are using. We would need to incorporate the need to adhere to these requirements as well as part of our product management efforts.

An important aspect for us to test, especially when AI is used as part of producing the output, is how people will use and respond to the outputs. When piloting and testing a new digital service from a policy perspective, we need to see how people will react to using the service and its outputs, and how people will be affected by it. In my experience, the most uncertain part usually relates to people's perception and their subjective assessments of whether the digital service and its output is something that they find palatable and acceptable, and useful. If it is a hypothesis relating to policy that we want to pilot and test, the key considerations of the policy from our perspective as the providers of the government digital service are usually revolving around these types of concerns.

The technological part of what we need to test and validate usually has more certainty and well-defined structure. The engineers and others involved in designing and building the digital service or product have a better understanding of the technological issues and their implications, and these aspects are usually more straightforward to test and evaluate in an empirical manner.

5. Navigating through the process of policy experimentation

Within GovTech Singapore, we have well defined descriptions of the product manager’s role and of the product management process.[3]

When I was serving in the role as a senior product manager prior to my promotion to my current position, what I usually did was to approach the policy owner and the other people in government who had most interest in guarding the relevant policies and have an open conversation with them to discuss if they would be willing to go on this journey together with me. I would ask them if we could collaborate to assess and determine the boundaries and interplay between their policy and the digital service product that I was working on as product manager. Through this collaboration and joint assessment, we could work together to achieve the best practical outcome in terms of integrating across the dimensions of policy, operations and technology.

5.1 Knowing when you need to do policy investigation or experimentation and how to frame it

It is important to recognize the overall context of the tech environment evolving rapidly, so it is also important for us to adjust our policy positions from time to time to remain relevant.

When we are working on the product, there will always be a product vision that we have already defined and a definition of a roadmap of where we want to go with that product and its related digital services. Those defining product vision and road map documents would also include a statement of risk considerations that summarise our initial assessment of risky items or areas where we need to be especially watchful of.

We would use that initial identification of risk considerations and areas requiring special watchfulness as a starting point for determining what we need to initially look into. We would look to see if there were relevant precedents or examples of how this had been done before. We would identify the various relevant existing policies that we need to be aware of and refer to.

Many government policies are initiated in response to some previously identified risk or previous problem occurrence, and those previously created policies are designed to protect us against those risk. As such, it becomes natural for us to look at our initial list of the things that we suspect are risky and to follow up by looking for existing government policies related to those same types of risks to see if there are any pre-existing policies or precedents that we can refer to.

Then, the product manager for the digital service will start talking to the relevant people both inside and external to government to see if people have previously tried doing something similar to what we are now trying to do with this new digital service and we collect whatever background information we can.

Then we get to the point where we need to make judgement calls on our assessments of the nature of these risky and watch-out-for items. We do some follow up to determine:

- Do some of these items seem too trivial in the scheme of things and therefore there's no need to worry about them in a special way, especially if there are already relevant policies in place?

- Are there items that are important risks to manage where there are already existing policies (legislation, regulation, rules) that guide how to govern these issues?
- Are there items that seem risky enough for special concern though they are not clearly or adequately governed by existing policies? For these types of situations, we would do further follow up and have deeper conversations with relevant government people who are responsible for that particular domain and discuss how to work together to proceed. We would work out an appropriate type of approach for managing and mitigating the risks. That might involve us creating a joint sandbox with another government unit for testing assumptions about the type of governance and oversight needed. Or it might be an initiative we do on our own within GovTech in consultation with that other unit of government though in an extremely cautious way.

In summary, from my point of view, policies are usually established to de-risk certain decisions or products or operations that we do. When we manage a new product for providing digital services, the risk dimension and related risk management Is usually already quite well aligned with policies.

For most of our efforts to create new digital services, we are operating within domains that already have reasonably well-defined laws and regulations that provide the structure for the governance needed to roll out the new digitally enabled capability. The existing policies may not have been designed or written to cover some specific aspects of the new digital services use case I'm working on, but they usually would have certain important overlaps. Then we figure out how to work within those predefined rules and through our pilot efforts, we come to understand the interplay between the new digital service and those rules.

Much less frequently but occasionally, we're rolling out some new capability in a space where there are not any or hardly any definition of laws and regulations.

In my prior product management work for our national digital identity product SingPass, we were trying to tackle online and digital scams and there was not so much existing policy that was directly relevant initially. To the extent there was existing policy, it did not have enough teeth to provide the necessary protection against fraudulent scamming with this new SingPass service for national digital identity. This caused a lot of issues for the SingPass product team at that earlier point in time in terms of control related to national identify protection and in terms of the ability to enforce laws to act on certain breaches of usage. Before we were able to get the right policies in place, there were bad actors who were abusing the use of SingPass and selling the identity credentials for criminal use.

In this SingPass example, the pre-existing legislation and policy turned out to be insufficient or inadequate to protect the product and its users against new modes of fraud and crime. Legislation was eventually passed that addressed these issues by criminalising the act of selling digital credentials for usage in criminal activities. This was an example where the lack of rule was constraining us from using the new digital service – in this case the national digital identify service- more pervasively and safely, and where a new rule needed to be created to address this need.

5.2 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

To my earlier point, AI has been around for a while, and we have been incorporating AI capabilities into our digital services for a number of years so we already have this familiarity with issues associated with using AI. At the same time, the recent capabilities of AI have rapidly broadened and expanded, progressing into spaces that we are still trying to understand. On one hand, even with these recent new developments with Large Language Models, I do not feel there are tremendous changes or fundamental shifts in the perspective of how we consider and manage risks related to using AI in our digital services. On the other hand, we are in fact updating and enhancing some of our risk management approaches because of the need to manage outputs of services based on new AI capabilities.

One of the first products I worked on as a product manager when I joined GovTech in 2018 was a customer service chatbot for responding to questions from our citizens and residents that combined natural language processing with a rules engine. The concerns we had back then was whether the response given by the chatbot was accurate, appropriate, and relevant for handling the customer service inquiry. We needed to make sure the AI did not misrepresent the intent of the customer's request, and that the response given back to the customer did not misrepresent the government in any way in terms of the information we provided back to the citizen or resident.

Our current generation of customer service chatbots based on Generative AI and Large Language Models (LLMs) are much more capable than the earlier chatbots we built that were based on the prior generation of natural language models. Even so, we know that GenAI based LLM's can hallucinate and fabricate inaccurate information. From my product management perspective, I feel the basic concerns of making sure we properly understand the customer's input and that we do not make any misrepresentations in the output that we provide back to the customer are unchanged. In this sense, I feel that the basic risk considerations are the same, though we must adjust for the new and expanded language capabilities of the GenAI based LLMs and also the unpredictable ways in which they can hallucinate and provide inaccurate information. As such, the need for enhanced approaches to the quality control of the chatbot output increases, even though the capability of the large language model is getting better.

For our recent and new chatbot efforts making use of LLMs, we are continuing to proceed with a "man-in-the-middle" (or "human-in-the-loop") approach to enhance our quality control efforts. We are using AI in new ways to support this human oversight of quality control so we can do this in a scalable and economically practical way. It is an approach of using "man-in-the-middle" in an overview and supervisory way.

Our chatbot responses need to be fully automated and occur in near real time. However, these automated responses draw from curated information. That curated information is the result of an ongoing effort that takes place in the background where a small

team of our staff members use AI tools (including LLMs) to continuously monitor and analyse the different types of chatbot inputs that stream in.

They review and assess the nature of the newer generation of LLM chatbot responses to different types of input requests. In this background mode, these people provide the human oversight and feedback to improve the quality and accuracy of the LLM's response. For example, they will provide guidance that certain categories of inquiries should (or should not) make use of certain types of internal government information sources if the LLM did not initially use the correct sources. Or once the LLM generates responses based on the correct government sources, they will refine the way the responses are communicated back to the customer.

Using the LLM together with Retrieval Augmented Generation (RAG), the AI system can quickly generate many possible question and answer sets given a set of data that is being fed to it. During this content curation process taking place in background mode, the human administrators read through what has been generated and make corrections in areas where the chatbot's response is inadequate or inappropriate. Then, this curated content is made available to the live chatbot system to make its automated, real-time responses to the public.

In this way, the automated response given back to the public in real time is drawing from curated information, and the ongoing curation process taking place in the background (off-line) involves humans whose efforts are highly augmented with LLMs. We tremendously improve the efficiency and quality of the human curators who are administering the ongoing operation and performance-related fine tuning of the chatbot through their use of the LLMs as support tools for the curation process.

Using this type of off-line supervisory oversight in the background, we still always have someone curating the content that is used to provide the chatbot's responses to the public. Even with the introduction of AI and the continuing improvement of LLMs and other AI tools, it doesn't reduce our responsibility to provide good services and accurate information to the public. In fact, we need to strengthen our ability to do this. That is why it is so important to increase the productivity of our public officers, especially those who are ensuring the quality of AI outputs used by the public, so they can do their supervisory curation effort more productively and sustainably.

5.3 Understanding the technology (including AI) well enough to establish the boundaries of trust and the appropriate risk management measures

As we continue onward with using technology, including AI, for government digital services, we must be very pragmatic and take a risk-based approach towards it.

Technology continues to abstract more types of work away from the human. Up to recently, digital automation and AI was able to handle the more mundane and repetitive types of work cognitive done by human staff. With recent advances, AI is starting to infringe on more parts of the less routine and more complex types of cognitive work done by humans.

Each time we use technology to elevate to the next level, it is important to understand the risks and to take the necessary risk management measures. We can only minimise the risk of using technology including AI as it is not possible to eliminate all risk. We need to keep calibrated on the fact that even doing the work with technology that does not use AI, or doing the work manually with humans, there is risk as any technology-based process, and any work involving human beings, is subject to mistakes and errors and to failing in various ways. So however we use people and technology, including AI, to deliver our government services, we have to strive to better understand the risk associated with whatever level of technology we are using and minimise that risk to an acceptable level.

As the capabilities of AI keep increasing, it becomes even more important for our government digital services staff members and decision makers to remember that they as humans are still accountable for the AI outputs and the related outcomes. Our responsibility for the outputs and outcomes of the service doesn't go away just because it has been automated with AI. With that human accountability comes the obligation to understand what the technology is doing, how it is working, how to validate it, and what risk management measures are required.

From my prior work experience in the private sector as well as my more recent government work experience, I have found it helpful to encourage some members of our staff (especially business and domain users of the AI outputs) to be cautiously sceptical of the AI system's outputs and to openly encourage their efforts to doubt, test and cross validate the results. I have found that for the few "business user" members of our staff willing to take on this type of extra effort, it positively reinforces their sense of personal accountability, and this type of effort enables them to do their own learning on the extent to which they can trust the AI system's outputs. These types of employees make very useful contributions to our risk management thinking.

Systematic efforts to pilot, experiment and sandbox are key ways that we develop the understanding and experience needed to establish the boundaries of our trust in the AI system, to identify the risks that need to be managed, and to devise appropriate risk management measures. As I mentioned earlier, this is all part of the regular effort of doing good product management. However, as we make greater use of AI systems and in parallel also use more sophisticated types of AI capabilities and tools, we need to be doing even more piloting, experimenting and sandboxing.

In the past, managers needed to be very good in human management. Now we are deploying AI technology to augment the cognitive capabilities of the team, including some of the less routine and more sophisticated cognitive capabilities. That means our government leaders and managers have to expand their view of human management to understand how this new augmented employee and augmented team works, and how this relates to how the supporting AI technology works. This new type of augmented human/augmented team understanding is extremely important for both the government staff involved in using the outputs of the system and also for the policymakers who need to make relevant policies around the use of AI both within the government and across the rest of the economy.

6. Implications of more AI usage for Product Management within government digital service units – don't lose sight of the basics

I go back to adhering to the basic principles of product management and I use that to set my expectation for how I want our product managers to approach and treat the usage of more AI in our digital services. The basic principles of product management deal with balancing and prioritising needs relating to the technology, the user, and the desirability, feasibility and viability of the product.

Product managers have to be obsessed with understanding how the end user actually makes use of the digital service, how they feel about using it, and why. The product manager must spend sufficient time with the experts in the technology to gain enough knowledge and understanding of how the product works and issues related to technical feasibility. However, the product manager's assessment of "product feasibility" must be broader than just the technical aspects of whether the product can be implemented from a technical perspective. They must also assess the feasibility of the output meeting the actual needs. It's not simply a matter of whether this new digital service using AI works or not. The more important consideration is whether it can eventually meet the outcome of delivering that particular service or information in ways that achieve the goal of the product.

The assessment of product viability in terms of cost effectiveness is also critical. While it may be possible to use AI for the product's functionality, the product manager must also consider the availability of lower cost alternatives for achieving the necessary functionality for a given situation. We have had examples in recent times where product managers have proposed incorporating machine learning based prediction models into a product to enhance it with proactive anticipatory abilities to predict what the users may need as a basis for making recommendations to the user. In some cases, that can be an essential requirement and worth the cost. However, in other cases, it can be a "nice to have" but not essential ability that substantially increases the cost of developing, validating and sustaining the product.

It is important that product managers do not get carried away with chasing after the latest technology – especially the newest types of AI technology - just because of the hype about AI. Currently, there is a mindset in many private and public sector organisations worldwide (and we are not immune from this) where senior decision makers are tilting towards approving a new product that claims to be making an innovate use of AI, even if the benefit-cost justification for the usage of the AI is not well understood. Per product management fundamentals, when we consider whether to add AI capability to a product or consider the scope and extent of AI capability to add, we assess whether the incremental value that the product would gain from this AI investment is commensurate with the amount of money that would need to be spent on the AI across the entire product lifecycle.

My expectation for how to manage the way forward with what will inevitably be more AI usage across our government digital services is no different than my expectations for going forward with good product management fundamentals. There will be layering or embedding of AI in most aspects of our product management, and we have to understand the key product

management considerations and trade-offs from the perspective of this always evolving and improving AI technology. Of course, this view reflects my GovTech role as the senior executive for Product Management so I emphasise that perspective.

7. Staying focused on doing things for citizens and not to citizens

When we build or enhance our products, I always tell my product managers that we must remember that we are doing things for citizen, not to citizens. That is an important principle. While this is something that may seem quite easy to understand, it is also something that is easy to lose sight of, and there can be profoundly undesirable impacts if the product team neglects this principle.

When we introduce new digital services or enhance existing ones, it is important for us to be very transparent in our thinking and in our communications about how this new product effort actually benefits our citizens and our other residents. Within the product teams, after we decide on our priorities for the initial release or a subsequent release iteration, it is also important for us to have the necessary discipline to reign in our attention and focus so we do not get too distracted by the many possible options for "interesting" additional features and functionality that were assessed as being outside the scope of our targeted set of priorities.

When we introduced our national digital identify product called SingPass, we focused on designing it so its use would simplify the citizen or resident's task of identifying (authenticating) themselves when they make use of our various government services. We did not emphasise the more abstract concept of "national digital identity" for its own sake because many people do not understand what that means and could not relate that concept to a practical benefit in their everyday life. For those who do understand the concept, some of them might have doubts about what the government is trying to do with the product because they are only thinking of all the negative things that are hypothetical possibilities rather than the practical positive things like easier access to essential government services which were our reasons for creating the product.

By keeping our product design and public communications focused on doing things for our citizens and residents that have tangible and practical benefits, we were able to deploy SingPass and achieve widespread public usage and acceptance. Of course, this was only possible because of our overall context in Singapore where there is a relatively high degree of inherent trust with the government.

In the same vein, as we continue introducing new technology-based digital service capabilities and products to our citizens, included those that are AI-enabled, the heart and mind of our government team members doing this work must always be on designing the service to make lives better for our citizens and residents. We need to stay grounded in this way and avoid getting swept up in the presumptuous mindset of, "Because we are using new technology or AI, we know this is really going to be great for the public." We also need to avoid the related mindset of "We must use this new technology or new type

of AI because it will make us look good in some international index or make us seem more 'advanced' to the rest of the world." Fundamentally, when we use any technology for our digital services, including new types of AI, we must always stay grounded and be guided by how this would make lives better for the people who would be using it.

If we always put that consideration in the centre of everything we do with our government digital services and related AI usage, and be very transparent about what we are trying to achieve with our products and services, then our experience has been that our people will start to appreciate it and understand it. Our citizens and residents, in their capacity as individual users or as business users, are making their own trade-off decisions every day. In this sense, they are like a product manager regarding their own everyday choices. They will ask themselves, "If I accept the use of this new government digital service (including those that are AI-enabled), what is the trade-off for me?" They do their own intuitive benefit-risk assessment. If they can easily see there is a huge benefit that they receive (e.g., in terms of ease of usage, overall convenience and time savings, or in reliability and assurance), they are willing to make some degree of trade-off, which may include forgoing some amount of privacy as long as they have an overall high degree of trust in the government.

This type of thinking guides our approach to how we use analytics and AI to determine the extent to which we personalise our digital service offerings. There is no simple rule that determines our judgement on where to draw the line on personalisation. I don't think there is a fixed or static line to say that this degree of personalisation is possible and that degree is not possible. It's a trade-off decision that we carefully make considering practical benefits to our users, a wide range of other user considerations, and other product viability related considerations. It is also a trade-off decision that our users are making every day regarding whether they choose to use of a more personalised digital government service that benefits them versus what they may be giving away.

As product managers, we must always go back to the roots of what is the problem that we're trying to solve and be very focused on meeting the needs of solving that problem, never forgetting that our products are meant to make the lives of our users better. As we design and deploy our government digital services, it is very important that we do not get carried away with using AI or other technologies for their own sake, and that we never lose focus on addressing real user needs and providing practical user benefits.

If we keep these fundamentals in mind, then everything related to the nature of how we need to experiment and sandbox for new AI and other technology initiatives will fall into line because we will have a clear sense of the problem we are trying to solve and what we need to test. Rather than having a pre-specified checklist or playbook for specifying how we should go about sandboxing, piloting and testing any new product (including those making new uses of AI), I feel it is more robust to make sure we get more of our internal government people focused on the fundamental culture and practise of product management as a means of providing an overall process for guiding the trade-offs and managing the risks we need to be identifying, evaluating and making choices about.

8. Comments on phases of the sandbox lifecycle

Note from interviewer: These phases are taken from the description of the sandbox lifecycle as given in this UN DESA Policy Brief (December 2021) on Sandboxing and experimenting digital technologies for sustainable development:

- Conceptualisation
- Operation
- Evaluation
- Exit

I think each of these phases has equal importance but is solving for different types of things. If you don't do the first stage or any subsequent stage properly, the next ones will get weaker because one stage is built upon the next and they are all stacked one upon each other.

8.1 Conceptualisation phase

As we're conceptualising, we are exploring options. This is where we need to be divergent in our thinking. Having the ability to consider as many options as possible is extremely important.

Very often, people, including our government staff members, tend to have a biases for solving a problem with new technology. In the conceptualization phase, it is important not to be too narrow minded about certain things, because this can sometimes put the whole problem-solving effort off tangent. We have to be open to assessing options for addressing the problem with new technology, including new AI capabilities. We also have to be open to assessing options for addressing the problem in the simplest, least complex way that may provide a highly effective and low-resource solution without the use of AI or other new technologies.

8.2 Operations phase

During the operating building phase where you are doing the experimentation and testing, it is important to be very targeted on what your hypotheses and related assumptions are that you want to test. Clearly defining this is really important. Otherwise, if we are not testing the most important or the riskiest assumptions that we are making, we may draw wrong conclusions, and then subsequently make inappropriate decisions. Being very aware of what the key risks are, and being very targeted in how to test them, is really important. Related to this, designing our experiments so that what we are testing relates accurately to what we want to evaluate is also very important. Otherwise, the test results will lead us into making wrong decisions. As stated earlier, we must work on understanding how the use of AI capabilities, especially newer types of AI methods, influence our choices on what we need to test and how we go about testing it.

8.3 Evaluation phase

A big challenge during the evaluation phase (as well as in the earlier phases) is dealing with confirmation bias. If staff members already have a strong opinion in mind about the "right" solution approach, when they look at the test data, they will selectively look for confirming evidence and ignore or downplay the rest of the test results. Such selective filtering of the test results could lead to reinforcing a particular preconceived idea that could be wrong. The reality and common occurrence of confirmation basis is a risk that permeates all the phases of sandboxing, especially the evaluation phase.

Overall, aside from the background issue of guarding against confirmation basis, the actual evaluation work during the evaluation phase comes down to the quality of the thinking used to evaluate and interpret the test results. That is usually not so risky because in most cases, my product managers are reasonably logical enough to make good decisions based on the data being collected.

As I mentioned earlier, these sandboxing phases build upon one another and influence one another. My comments on the evaluation phase are predicated on the ability to identify the appropriate range of options during the conceptualisation phase, and on the ability to define and execute the appropriate testing during the operations phase. If there are issues with the work done in these prior phases, it complicates the evaluation and assessment efforts during the evaluation phase.

9. Advice for learning from the digital services and related AI efforts in other countries

Whenever I present to government digital service teams from other countries about what we have done in Singapore, I always emphasise that whatever we have done in our local setting was done in a way that was useful and workable for our particular situation. Their local priorities, needs and circumstances may be very different. As such, I caution them about jumping to the conclusion of assuming the same problem definition and related solution approach we used would be appropriate for the situation in their country.

This works in the other direction as well. When I visit other countries and learn from some of the successes of their locally contextualised government digital service efforts, I remind myself that I cannot simply copy and paste their successful solutions into my local Singapore setting.

What I find most useful from sharing sessions with my counterparts in other countries is the opportunity to understand more deeply about the problem they were trying to solve, and how they went about solving it given their local context, priorities and constraints. I draw inspiration from this type of learning. It stimulates me to consider how I might apply aspects of what they successfully accomplished within my own setting and circumstances in Singapore.

It is important not to assume that the context and nuances of the problems are the same across national borders. That I why I suggest to my counterparts in other countries to not get too distracted and carried away by reports or presentations on what other countries have done with their government digital service efforts as the context, priorities and circumstances will always be different than those in your own setting. Learn from what's going on in other countries, but don't start from the presumption that what you want to do is to directly copy.

10. Concluding thoughts: balancing the ability to try with the discipline to manage risk

As I continue to learn more about AI and how we might make use of it in our government digital services, I am often asking myself whether we are missing out or not anticipating certain issues. I am always cautiously considering what we may have overlooked as we proceed with using more AI to deliver our digital services. I am always scanning and questioning to probe if there dimensions or aspects that my teams and I need to better understand so as to avoid being blindsided when we deploy our solutions.

Since I joined the public service nearly six years ago from the private sector, I have seen a few examples where the outcome of the technology was what we wanted in the sense that the application worked as intended in terms of features, functions and technical aspects, but the outcome of people's acceptance and interpretation was not what we had expected or hoped for. With increasing usage of AI in our digital services, including the use of more capable types of AI, we need to be even more thorough in increasing our understanding to avoid being blindsided by these types of surprises, and in our testing and evaluating user acceptance.

The Singapore government wants to encourage responsible usage of AI both within the government sector as well as across the entire economy. To encourage adoption, we need to encourage and enable AI related understanding through education and training efforts, and through practical experimentation, testing and evaluation.

While we have existing and evolving polices related to AI usage, we know ongoing trends and the trajectory of our ongoing efforts will lead to us into new territory that we may not fully understand and where we may not yet have the appropriate written policies for new situations and use cases. However, if we were to start writing too many policies around our evolving and future AI usage before more fully understanding the situation, we may overly hamper our ability to explore and apply the technology for public sector applications that benefit the public and for helping us to improve our overall future economy.

One bridging mechanism that we have put in place to deal this rapidly evolving situation was to add a few staff with deep AI technical and application expertise, and with expertise in Singapore policy related to AI matters, to support our senior policy makers in our Smart Nation and Digital Government Group. This way, when my team members and I have questions about policy related to AI applications and usage that may still be undefined or in a

"grey zone," we can consult with this senior expert who can give us guidance on the spot on how to manage and navigate these undefined policy-related issues for the time being. This way, we can at least get a certain degree of intermediary guidance on what we can or cannot do in certain new types of situations. Our queries would also help this expert and the senior policy makers inside the government to understand the existing policy gaps.

In terms of how I guide my product management teams, we don't want to assume that everything involving AI cannot be done unless there is already explicit policy or special permission deeming that it can be done. This would be highly restrictive for us. At the other end of the risk management spectrum, we cannot simply assume that anything involving AI can be done, especially for new application scenarios where it is not clear whether or to what extent existing policies apply, as this would be to risky. We want a mindset that enables us to explore how a new type of government service using AI in a new way can be done, though with parallel pursuit of all the doubts, risks and relevant precedents as part of our early-stage and subsequent investigations to guide us in determining whether and how to proceed.

I feel this is a balanced and responsible approach. It does not prevent us from responsibly exploring how to innovate in ways that better serve public needs. At the same time, it requires us to be very disciplined about our early-stage and follow-on identification, testing and evaluation of risks, potentially sensitive user acceptance issues, and potentially ambiguous or undefined policy issues. It gives us the space to at least do initial sandbox trials to determine whether and how to keep proceeding forward. In essence, they key is to create a culture and supporting practices within organisational units responsible for government digital services that balances being innovative while simultaneously being disciplined, restrained, evidence-based and responsible. In this type of setting, we do not feel so restricted that we cannot even make a first step to investigate the situation. We can take that first step and learn whether and how to proceed forward through various the various stages of sandboxing and iterations of piloting.

Endnotes

[1] Visit Dominic Chan's LinkedIn profile at https://www.linkedin.com/in/dominicchanjh/?originalSubdomain=sg.

[2] Information on GovTech products (government digital services), including some of the products mentioned above where Dominic was the senior product manager. Digital products for citizens: https://www.tech.gov.sg/products-and-services/for-citizens/digital-services/. Digital products for business: https://www.tech.gov.sg/products-and-services/for-businesses/corporate-transactions/.

[3] The Digital Academy of GovTech, Product Management course description website: https://www.thedigitalacademy.tech.gov.sg/category-product-management. When GovTech has openings for product management roles, they are posted on this website: https://sggovterp.wd102.myworkdayjobs.com/PublicServiceCareers?Agency=27bc56da9e6a01dcff9491800407da09&Job_Family_Group=27bc56da9e6a01598012e66f50087e59. Product Manager job listings provide a description of the product manager's role.

INTERVIEW 10: Prof Rhema Vaithianathan, Auckland University of Technology, New Zealand

Date of Interview: September 27, 2024

1. Introduction to Prof Rhema Vaithianathan and her work

I'm a professor of health economics at Aukland University of Technology in New Zealand and serve as director of the Centre for Social Data Analytics, a centre at the university I co-founded in 2016.[1]

1.1 Early career realization of the gap between building models that make "good" predictions and providing useful tools for real-world decision support

Earlier in my career, When I was a Harkness Fellow at Harvard University Medical School in 2007-2008, I continued my work on building predictive analytics tools to predict hospital readmissions. The purpose was to predict which of the inpatients admitted into the hospital were highly likely to return to the hospital for another inpatient stay within the near term. My assumption was that if this type of prediction for readmission could be made, the hospital staff would be able to take action through some type of intervention during the current inpatient stay in order to prevent or reduce the likelihood of subsequent readmission.

After I returned to New Zealand in 2008, I built a localized version of my hospital readmission tool and deployed it in three hospitals in Auckland region. I was able to do a good job on predicting who would soon return to the hospital for readmission. However, I was surprised that doctors did not find this prediction tool to be useful. What the doctors said to me is they "get" (as in understand) the likelihood of readmissions score, and they totally agreed with the prediction that this person is going to soon return to hospital. However, they did not know what decisions to make or what actions to take because of seeing this prediction.

While the model identified those with a high likelihood of readmission, it did not provide an explanation of why and correspondingly, did not provide guidance to the doctor on what to do to reduce the predicted risk of that patient soon returning to the hospital. Also, the hospital had not yet set up processes to support follow on action to reduce readmissions.

This experience had a lasting influence on my professional career from that time onwards. I realized that prediction tools must be coupled very closely with the real-world business process and be embedded within the workflows related to the everyday decision making and follow up execution. This experience made me realize that there is no point in giving professionals the results of a prediction model if they are not able to use it at the right point in time to make a decision that leads to practical choices or follow up actions.

1.2 The start of applying predictive risk modelling to child welfare

In 2012, I started working with a collaborator to see if we could use available social, economic, demographic and criminal justice data to build analytic models to predict those households and situations where children had a high likelihood of being victims of child abuse. We used various demographic data sets available in New Zealand. Initially, it was a theoretical and proof-of-concept exercise. We showed that available historical data sets could be used to predict quite well those households where a child would subsequently end up being abused. When we studied these predictions, we also noticed, unfortunately, that when we went back and looked at the social services records, these children who we predicted would be abused - and subsequently were eventually abused - received very little social services support prior to the abuse occurring.

This led to my initial reports and publications with colleagues on using analytic tools (including AI methods for prediction) to build models to predict the risk of child abuse in New Zealand.[2]

1.3 The predictive risk modelling partnership with the Department of Human Services in Allegheny County, Pennsylvania, USA (Greater Pittsburgh metro area)

Based on this initial work done in New Zealand, my colleagues and I won a competitive tender from the Allegheny County (Pennsylvania, USA) Department of Human Services in 2014 to use their data to build, validate and operationalize a model for predicting child abuse that could be used by social workers and their managers as a decision support aid. I was the principal investigator for that grant award.

Allegheny County's Department of Human Services is quite famous in the social services professional community. They are internationally recognised for the quality and scope of their integrated data platforms, and for their ability to use their data for doing analytics and predictions to support their everyday work and their strategic planning.

Given my experiences interacting with multiple government social services agencies globally, my assessment is that Allegheny County has the most integrated, comprehensive and operationally useful social and human services data platform in the world. Their case management system genuinely gives the case workers a 360° view of the client's history. For a researcher like me focused on using data and analytic (including AI) models for making predictions to support government social services agencies, it was a great opportunity and privilege to be able to work with Allegheny County.[3]

Our first contract with them to build, pilot and evaluate a risk prediction model for child abuse led to the development of the Allegheny Family Screening Tool (AFST), a first-of-its-kind social services predictive risk screening tool which was put into deployment in September 2016.[4]

Over the years, we have published numerous methodological and evaluation reports and peer reviewed articles based on our child abuse risk prediction work with Allegheny County. A sample of these write-ups are listed in this endnote.[5]

1.4 Expanding the predictive risk modelling partnership with Allegheny County beyond child abuse to include homeless housing and other areas

In 2017, we expanded our collaboration with the Allegheny County Department of Human Services to build prediction tools to support the agency's decisions related to homelessness and the allocation of the limited available capacity of temporary shelter and longer-term housing. There's a very big challenge in dealing with homeless people at the local (township, city, county) level in the US and in many other countries as well. In many localities, there are many more people who are homeless than there are shelter and housing slots available for the local government agency to provide for support. This gap exists both for temporary shelter housing as well as for permanent supportive housing (which is far more expensive).

Examining their data, we were able to identify major issues with the ways these housing allocations for homeless people were being made. We observed that in a high proportion of cases (up to 50% in some settings), the homeless people who were receiving the housing support from the social service agency had low risk of adverse outcomes. For example, we found that people who were at very low risk of future mental health crises or involvement with the criminal justice systems were being prioritised.

At the same time, many people who were in various stages of ending up being perpetually stuck in the trap of homelessness and were at high risk of the sorts of adverse outcomes that housing is best able to mitigate were being denied long-term housing support due to capacity limitations. To us, as experts in social services analytics and in creating useful tools for predicting risks and needs, this indicated there was a high degree of randomness (noise) in the existing decision-making process.

The predictive risk models we create as decision support tools for the front-line case workers making these supported housing allocation decisions reduces the degree of randomness in these types of decisions and leads to more effective allocations. Across the US, the local level data for most homeless systems is not so well integrated with other types of social-economic, social service and demographic data. As such, we must be pragmatic about building simpler yet still very useful models that can be used with the limited types of data they already have available. We first deployed this type of prediction model to support homelessness related housing allocation in Allegheny County about 2 ½ years ago and have subsequently built and deployed this type of predictive model in several other municipal locations.

We have also published numerous methodological and evaluation reports and peer reviewed articles based on our predictive risk modelling efforts to support housing allocations for homeless people. A sample of these write-ups are listed in this endnote.[6]

In 2018, we further expanded our collaboration with Allegheny County Human Services to predict if there were families with new children who were not making use of available social services though their circumstances were such that they likely had a need for such services.

Examples of reports and publications based on this type of prediction are given in this endnote.[7]

We are now in the early stage of working on predictive risk models for new types of use cases. One new tool in early-stage development is for supporting case worker decisions related to people with mental health issues who do not have housing. This builds on our work related to supporting housing allocation decisions for homeless people but focuses on a more specialized sub-population of those with severe mental illness as this involves special circumstances and risks.

A second new tool in early-stage development is for supporting case workers who counsel people being released from jail terms. Statistically speaking, this sub-population of people have a higher probability for ending up as victims of violent crime incidents perpetrated by others. A challenge in making use of a predictive risk model in this special context is to find an approach for the model and for how the case worker communicates with and counsels those being released from jail that turn out to be helpful (by reducing the risk of undesirable outcomes) and that do not inadvertently turn out to be harmful (by further increasing the likelihood of the undesirable outcome occurring.)

1.5 An example where the use of an analytics-AI based predictive model helped to reduce bias in decision making

One of our findings is that the use of the predictive model for homeless related housing allocation helped to reduce racial disparities. Why might this be so? While this is a very complicated and nuanced situation, we think there are two main reasons why the use of this model as a decision support aid helped to reduce racial disparities in these allocation decisions. The first reason is that some segments of the population, especially impoverished minority subsegments, tend to systematically understate (as in omit or only partial reveal) relevant aspects of their history and situation when asked to self-report as part of determining their need for homeless housing. We suspect this is to avoid the stigma that might be associated with more fully sharing this information when talking to the social worker who is evaluating the case.

The second reason relates to a well-documented phenomenon in the social science literature that a person of a given demographic group (or ethnicity) has a stronger emotional or affinity response when they experience or consider something that happens with a member of their own affinity group. This is referred to as “own-group empathy” or “in-group bias.” As such, it is not surprising that case workers of a given demographic (e.g., race or ethnicity) may subconsciously see more need in communities where the people in need of housing mirror their own background.

The key point is that the well-managed use of well-designed and properly tested predictive models can help to counteract biases in decision making in ways that help make decision outcomes fairer and more equitable. Granted, there are many predictive models that are not so well designed and where the usage is not so well managed, and this has led to further propagation of biases in decision making. But it does not have to be that way. It is

important for public officials to know these tools can be designed and used in ways that help to reduce bias and achieve fairer outcomes.

1.6 Applying and evaluating our predictive risk modelling work in other geographic locations in the US and internationally

The work we did on predicting the risk of child abuse in Allegheny Country Pennsylvania led to conducting two randomized control trials withing two counties in Colorado, USA - Douglas County and Larimer County. Both randomized controlled field trials demonstrated positive effects from using the predictive risk modelling support tool.

We are also applying these tools in other locations in the US and starting to apply them in some international use cases that are built for different approaches to child welfare.

2. Our "guard-rail" guidelines for the ethical development and adoption of predictive decision support tools for high stakes social services

Based on our experiences to-date, and also drawing on other published efforts proposing guidelines for usage of AI tools in the public and private sectors, my collaborators and I have articulated a set of "guard-rail" guidelines that we follow in our work with social services agencies. We have found these guidelines lead to an ethical and responsible approach for developing, deploying and using these predictive tools. These guidelines are as follows:

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.

Guideline #1 (Agency ownership): The human services agency needs to be in full control of the end-to-end effort including defining the rationale, policies and processes for how the analytics and AI-based predictive risk modelling tool will be used. We strongly recommend that the agency should own or co-own all parts of the tool including the data that is used to train the model, the algorithm that is deployed and any ancillary documentation. This can sometimes lead to complicated intellectual property (IP) ownership negotiations and/or pragmatic but still acceptable accommodations when the tool is being supplied by a commercial product vendor. Another aspect of ownership is that the agency, including the leadership level, must be able to explain the rationale for using these tools as well as the underlying conceptual principles and logic of how they work to members of the community as well as to lawmakers.

Guideline #2 (Fairness, ethics and transparency): The agency must be transparent about the design and use of the predictive tool, especially about the implications on fairness of

decision making. It helps if the agency openly provides extensive documentation, including technical reports and easy-for-the-general-public-to-understand responses to 'Frequently Asked Questions' on their public website. It substantially helps if the internal or external experts designing and evaluating the model can openly publish their work as either technical reports, conference papers or journal papers. Publications that are subject to external peer review further help with vetting and validating key assumptions and results, and arrangements should be made for peer reviewed publications to be openly accessible via a link from the agency's or supporting research partner's website. In support of fairness and equity considerations, there should be specific analysis and discussion of 1) the alignment between the characteristics of the data sets used to train the machine learning prediction model versus the characteristics of the population the model will be applied to in post-deployment actual usage, and 2) how the model performs on minority sub-populations and on other already disadvantaged sub-populations. The external audit of model performance called out in guideline #4 should eventually be openly published on the agency's website.

Guideline #3 (Community voice and engagement): The agency must seek and incorporate community concerns related to the use of these predictive decision support tools. There are often broader, non-technical concerns related to more general social and governmental system-wide issues that emerge from the community (e.g., related to the ethics and/or morality of using any type of predictive tool, pre-dispositions that use of data and analytics in any way will lead to increased bias). These broader issues can be more complex and challenging to discuss and pragmatically respond to than more targeted issues and technical inquiries that are specific to the proposed adoption of the given type of predictive risk model being considered. It helps if the agency can make use of already existing strong community feed-back loops. If these don't exist already, they need to be put in place. Key members of the agency's internal and external (vendor, consultant) staff involved in the design and evaluation of the system should sit in on some of these community discussions to make sure they are aware of and sensitised to the spectrum of views and concerns across the community.

Guideline #4 (Audit and evaluation): We strongly encourage the agency to support an external evaluation of the predictive tool's usage and performance as a supplement to internal evaluation. This can be done during the field pilot period or after the conclusion of the field pilot. After the field pilot, if the predictive tool is used for ongoing operational usage, there is also a need for ongoing periodic audit and evaluation to ensure the predictive power of the model that was verified and claimed at the time of initial deployment is still being maintained at later points in time as internal and external aspects of the setting evolve. Another important dimension of audit and evaluation, one that goes beyond evaluating the quality of the predictions being made, is whether the use of the tool as a decision support aid results in the agency being able to achieve the broader beneficial outcomes and impacts they earlier put forth when they initially presented the rationale for implementing and using the tool. The agency also needs to have the management commitment and supporting resources to follow up on the results of audit and evaluation results in order to address issues raised and any performance gaps that may be identified.

Guideline #5 (support and augmentation, not automation): The one additional guideline that we follow is to counsel and train the agency not to use the prediction tool as the final arbiter or maker of a decision. We strongly believe the prediction model should be used as a decision support tool that augments the thinking and judgement of the human case worker and the agency's management and not as a fully automatic decision-making tool. This guideline strongly complements the efforts described above for guideline #2 (fairness and ethical review) and #3 (community voice).

We provide a discussion of guidelines #1 through #4 in a research book chapter that we published in 2024.[8]

2.1 How the illusion of validity influences the initial way case workers respond to the availability of our tool

We want the front-line social worker using one of our predictive risk tools to use it as an input and aid for supporting their decision making and not as a fully automated replacement for making the decision. As each of these tools moved into deployment, we initially wondered if we would have issues with social workers quickly becoming overly reliant on using it.

To our surprise, it was the other way around. When the tool was initially introduced to a new group of social workers, it was more often the case that some of them would ignore it and not make use of it versus our supposition that some would take too much account of the tool.

The "illusion of validity" is a term first used by acclaimed decision researchers Daniel Kahneman and Amos Tversky back in 1973 to describe the fact that people are often overconfident when making decisions; that it is often the case that assessments and predictions people make when analysing a situation and related data are less accurate than they believe them to be. They showed that most people exhibit overconfidence when they are dealing with uncertain situations, and that judgements are highly fallible.

As this illusion of validity applies to all people making uncertain decisions, it also applies to our context of social services case workers. When they are required to make a high stakes decision, they often tend to overestimate their own ability to make that decision in a rational way. They often overly consider the importance of an individual situation and the most recent situations, and under consider (or totally neglect considering) the available date on longer term base rates across the larger relevant sample. It is not easy to admit to ourselves that we do not know as much about that decision as we believe we do. We don't really want to believe that when we follow our subjective judgement and intuition, that it is often the case that we are doing no better than tossing a coin every time we decide on something important.

From what I have observed, when social services case workers are in the early stage of having our tool made available to them, they are more inclined to believe that the decision-making recommendation of the algorithm is wrong more so than their own intuition about the decision is wrong. I believe this is why we have not encountered the issue of social workers overly relying on the tool after we initially introduce it.

We also need to keep in mind the context of the people who are the front-line users of our predictive risk tools. We are talking about social workers who spend their time interacting with families in difficult situations or with homeless people at shelters. While these social workers are very capable and smart people, for the most part, they are not particularly quantitatively- or data-oriented people. Many of them went into social work because they want to talk to people much more so than wanting to interact with computers or numerical reports or do formal analysis of risk probabilities.

For our predictive risk modelling tool used for homelessness-related housing allocations, when the social worker enters the client's identification information into the supporting information system, we show text that indicates whether this client is eligible for a permanent supportive housing, or not eligible for such housing. This "recommendation" by the system is based on the outputs of our risk prediction tool. The social worker makes the final decision as to whether or not the client is eligible for this type of housing. When the social worker chooses to override the recommendation of the information system, they can do so by clicking on the override button and providing comments on why they want to override.

I've done some preliminary work to study the override decisions for our users of the homelessness risk prediction tool. The data hints at some types of systematic bias in the overrides. This supports my point above of how these front-line social works are sometimes influenced by the illusion of validity, and how consistent use of our tool can help move decision making towards fairer and less biased outcomes, even with the fact that some of the override decisions made by social workers may sometimes be unwarranted, influenced by either bias or randomness.

2.2 The practical challenges of social workers learning how to make better decisions given the context of their work

These social services case workers have heavy caseloads. They are often making 5 to 10 high stakes decisions a day. Yet, they have no way of receiving fast feedback to know how good their decisions were because the impacts of these types of decisions can often take a long time to materialize.

Because of the heavy caseloads, along with the fact that many of the cases have similar features, and the reality that these social workers do not have research assistants or other underlings to help them keep track of all the details across all of the cases they are juggling, they often cannot remember at later points in time (days or weeks or months afterwards, and sometimes even years later) why they made a particular decision on a prior case in the way that they did. These conditions make it very difficult for individual case workers to learn from the results of their prior decisions unless special information systems support is provided for analysing the outcomes of prior decisions.

Contrast this to the situation when a doctor sees a patient in the hospital. The doctor can read the card or the report printout at the bottom of the bed that was prepared by support staff. The doctor has a whole team of people, including junior clinicians, who tell them what's going on, helping them to remember the important aspects of the patient's situation.

These social workers usually have to make an ongoing stream of very high stakes decisions on their own with minimal or no help for support staff. Given these circumstances, it is completely understandable and to be expected that they often cannot remember the specific context and circumstances in which they made a prior decision that ended up with a bad outcome.

This highlights that one of the challenges in social services settings is building the environment for the individual case workers as well as for the organisational unit to more systematically learn. There needs to be support for the individual social worker and for the organisation overall to learn from the accumulation of feedback on outcomes over extended time periods, match that to prior decisions, and use this knowledge to guide getting better at their everyday decision making. In many social service organisations, these types of feedback loops that can practically help with improving the case worker's decision making are minimal or non-existent.

The types of predictive risk tools we have been building and deploying are a stepping stone in this direction of supporting more systematic and cumulative learning for both the individual case worker and for the agency overall. Especially in light of the enormous caseloads of these social services agencies, it is a big challenge for management to have accurate and operationally useful knowledge at a summary level of what types of decisions their front-line workers are actually making. Our models, together with the way they are embedded into the agency's information systems and workflows, now make it possible to give management a quarterly report that provides a summarized comparison of the actual case worker decisions made (including overrides) versus the decisions recommended by the predictive risk model.

This type of comparisons makes visible what the decision patterns look like in aggregate across the entire population of clients served as well as within specific sub-groups of the client population. Appropriately summarized and privacy protected versions of this information can be shared with case workers to make them aware of these patterns. These summaries provide management as well as case workers with better visibility of where social worker decisions about follow up investigations or resource allocation decisions seem out of proportion in one way or the other (too many, too few). This visibility provides guidance for where management and supervisors should hold discussions with case worker staff to better understand these situations from the perspective of the case workers.

Without these types of comparisons of actual decisions made versus recommended decisions based on the predictive risk tool, management can only rely on the front-line case workers saying "we're doing a good job. We're making good decisions." Also, the case workers would have no way of viewing their actual decision making against the "reference standard" recommendations of the model. It is important to emphasize that there must be a high degree of trust between case workers and management within the social services agency for this type of feedback to be used in ways that are viewed across the organisation as being constructively supportive of learning and improvement and for the benefit of employees as well as for management.

In this sense, the predictive risk tool is like an objective observer in the organisation that can help all levels of staff by providing comparisons and summaries that can be used for various types of feedback, which in turn support cumulative learning. This is another reason why it is so important to put the necessary rigor, validation and ongoing quality assurance effort into the way these predictive models are created and used. We need to make sure that the feedback provided by using these tools is as accurate as possible as well as ethical, fair and un-biased.

2.3 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

As they use the predictive tool over time, case workers see more instances of how outcomes unfold over time and gradually see more evidence that the probability-based risk assessments of the tool have some useful degree of validity. Many case workers consider the recommended output of the tool (e.g., a risk sore, a housing recommendation) more seriously as they make their decisions. For example, case workers who use our predictive risk tool for child protection have seen children come back with severe abuse or even dead and have gone back to the information support system and saw that at prior points in time, those kids were repeatedly being flagged by our tool as high-risk cases and these recommendations were overridden for whatever reasons.

The more extreme realizations of adverse outcomes (such as a death or severe occurrence of one type of another) happen less frequently so it may take several years for the social worker to build up trust in using the recommendations of the tool. When they go back to the case history in the information system and see those examples where children who ended up later in time being severely abused had been flagged earlier in time as being at high risk, it really changes the case workers mindset towards the usefulness and value of the predictive tool.

These are very complex situations and there are indeed well characterized limitations to the accuracy with which we can predict risk. The sensitivity rates of our tool (the ability to predict a “true positive”) and specificity rates (the ability to predict a “true negative”) are much lower in these types of domains than in other settings (e.g., predictive models used in healthcare to screen for the presence of eye disease). Correspondingly, even with our best performing models, there are still relatively high rates for false positives and false negatives. Even with these limitations, case workers eventually notice and take heed of the fact that our predictions with the higher risk scores usually end up eventually having a higher occurrence of adverse outcomes.

The predictive risk model has limitations which we try and characterize as carefully as possible and make the case workers aware of. As they see more examples of results over time, the case workers gradually update their view of the tool’s usefulness as they realize that the predictions are “not as terrible” as they initially assumed they might be. They also come to

appreciate that their own judgements (unaided by the tool) are not always as good as they had been assuming.

The key point is as follows: In these very complex, high stakes social services settings, the ability of our models to predict risk are far from perfect. 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.

Even though most case workers do not understand the technical details behind how the agency's data is analysed and how machine learning and statistical methods are used to create and validate the tool, they get more comfortable over time with using the tool's recommended output and using it as a supplement and cross-check to their own intuition and judgement for a particular case. Here is a typical illustrative example. A case worker using our child protection predictive risk tool has become familiar with seeing how the tool's predictions unfold over multi-year time periods. The case worker is alerted to investigate a case of a child potentially at risk and based on his or her initial assessment of this family and the overall situation, their judgement is that there is no need for an immediate intervention. If the predictive risk tool did not exist and if they did not have experience with observing its predictions over time, based on their intuition and subjective judgement, they would have just closed the case and moved on to the next one.

However, if they saw that the output of the predictive risk tool was a high-risk score (e.g. a risk score close to the max score of 20), the case worker in this illustrative example would think of some way to take extra care and caution and perhaps offer some types of supportive services to keep an ongoing engagement with the family. They are using their own professional judgement of the situation not to immediately intervene. Yet, because they have come to learn that the predictive tool often provides useful, objective information, especially for cases flagged at the highest levels of predicted risk, they take some type of additional risk management measure that keeps the situation under view to allow for the possibility that the predictive risk tool is picking up something they are missing.

2.4 Why our team focuses on predictions for high-stakes social service decisions where extreme adverse events only occur infrequently

When you predict things that occur frequently, the eventual outcomes of the predictions are very adjacent (in terms of the passage of time) to the decision maker. If you make predictions where you can see the results of outcomes in the near term, for example, within a month (or week or day), a decision maker who is systematic about comparing decisions to outcomes can get quite good at making those decisions anyway, even without AI-based prediction support tools, because they have many repeated, near term opportunities to build and refine their judgement based on many observations. If they are using an AI-based prediction support tool in this type of situation, they can repeatedly observe the results of the tool's predictive performance within these same short time frames, more quickly learn to

calibrate to the tool's predictive performance, and more quickly build trust in these predictions if there is operationally useful performance.

The types of predictive risk management tools that my colleagues and I specialize in building for public sector social service agencies are for managing the risk of very high impact but less frequently occurring events, most notably risks related to child safety, as well as risks related to homelessness, and a few other high stakes areas. Our tools are designed and validated to predict the likelihood of occurrence of less frequently occurring but high impact, adverse outcomes. This is why it can take a case worker a while, even a few years or more, to build up trust in using our tool because of the longer time cycles required to observe the ultimate outcomes and assess the performance of the tool's predictions over time.

Through the work we do with public sector agencies, we have come to appreciate that AI-based prediction tools are especially helpful when decisions are about less frequently occurring events where the eventual outcome can only be observed a longer while after the case workers made their decisions about the case. As explained above, in this type of setting, the human case worker usually does not have the opportunity to learn in ways that lead to systematically improvements in their decision making because of the delay in receiving feedback; there are typically long gaps between when the decision was made for a particular case and when the adverse outcomes were experienced by that particular client. In addition, the case working is handling many cases in parallel, and always taking on new cases, so it is nearly impossible to for the case worker to retain the details needed for understanding causal attribution.

We choose to concentrate our efforts to support public sector social services agencies around decision support related to these rarer, longer term, high impact events because we believe AI-based prediction tools are especially valuable in these types of situations. They provide help that makes a really useful and important difference to the social worker, and to the clients they are serving as well.

3. The end-to-end process for developing and piloting our predictive risk models in high-stakes social services settings

We go through four steps related to model development, testing and validation prior to the social service agency deploying the model and initially using it in one or more field pilots. Then there are at least two major steps within each post-deployment field pilot. It can take one or two years, and sometimes even longer, to carefully complete all of the preparatory work prior to field piloting. It takes at least several years for the subsequent field piloting. **The key point is that the end-to-end process for carefully and responsibly developing, testing, validating and evaluating our risk prediction models in the real-world contexts of these type of high-stakes social services decisions is an intensive, multi-year process.**

3.1 The four steps related to model development, testing and validation prior to deployment for field piloting

The first step is to develop and test a model that is as good as possible at predicting the target outcome, given the available data. We use well established machine learning and statistical methods for doing this. We take appropriate precautions in managing the portion of the data set used for model fitting versus the "hold-out" portion used for testing to prevent data leakage or to minimize its impacts to the greatest extent practically possible.

The second step is to do external validation. This step is especially important when we are using a proxy variable as an indicator of the actual ground truth outcome we want to predict. There are often circumstances where the only practical way to create and test the predictive risk model needed to support the decision making for the specific use case we are considering is to use a proxy indicator of the outcome measure we ultimately want to improve (or reduce if it is an adverse type of outcome). For reasons related to data availability, or for other types of reasons related to social or political sensitivities, it may be necessary to use a proxy variable for our prediction model.

For example, in the domain of child welfare protection, a predictive risk modelling approach often used is to predict the likelihood of the child ending up being removed from the family within the next two years due to a very high threat level to the child. While this has proven to be a practical and operationally useful variable to use for the prediction model, it is a proxy for the more direct, ultimate outcomes of concern to the agency such as whether the child will be hospitalized for injury, or get severely abused, or die of maltreatment. There are often a variety of reasons for why these more direct outcome measures cannot be used in the operational version of our child protection predictive risk models. Similarly, in our predictive risk modelling efforts for homelessness and related allocations of housing, there are hospitalization outcomes, related physical and mental health outcomes, and substance abuse outcomes, that are important direct ultimate outcome concerns that may not be usable as the target of the predictive risk model used by the social services team making the decisions.

When using a proxy outcome variable is the only way possible forward for the predictive risk model, we must very carefully test the degree to which our proxy variable is a good indicator of the ground truth indicator of ultimate concern. We do extensive testing with all possible external (as in, outside of the social services agency we are working with) data sources to characterize and calibrate the relationship between the prediction of the proxy variable and the prediction of the ground truth variable. We go through elaborate efforts to work with deidentified local government data from the appropriate external agencies to do this external validation of our proxy outcome variable.

This external validation step, especially import for using proxy outcome variables, takes a lot of time, but is essential to do as part of assuring the quality of the predictive model.

The third step is to carefully test the predictive performance of the model for different subgroups (e.g., gender subgroups, race/ethnicity subgroups, income subgroups, nondisabled versus disabled subgroups) to test and assess predictive ability for each subgroup and for the

"cross-products", the interactions across some or all of the subgroup categories. We determine if the accuracy of prediction is similar or not for different subgroups. For example, is the model as good a predicting our outcome variable (whether it is a proxy variable or the direct measure of interest) for woman as for men? For a disabled person as compared to nondisabled person? If these types of model testing results are not satisfactory, we work with the social services agency to expand data availability or improve data quality, or to refine model design, to improve the predictive performance across the various subgroups relevant to the issue being supported by this model.

The fourth step is to do a check if there are special cases where the model's prediction of a high degree of risk can lead to a follow up intervention pathway by the social service agency that is likely to result in the exacerbation of undesirable or adverse effects versus the reduction of such effects. While these types of situations are not common, we know from experience that this sometimes can happen. For example, an intervention may require placing the "subject" (be it a child requiring protection or a homeless person) in a special type of facility, and there may be multiple providers of that special type of facility. Some providers may actually be very bad at working with the highest risk families and actually end up exacerbating adverse outcomes for these families. We need to know that before we use a predictive risk model to choose the best cases for this intervention.

In essence, if we are using a predictive risk model to select cases for specific intervention, we need to know that the intervention is actually effective for these cases. Evidence from historical and recent data may show that under certain types of situations and conditions applicable to a particular case, a person sent to a subset (which may be just one or two) of the available facility providers usually ends up later in time in a situation that is worse off rather than better off.

This is an illustration of how an intervention motivated by reducing the likelihood of undesirable things happening in the future can end up with the opposite effect. Because we have been working on building these types of models for over 10 years, we have learned that it is important to add this fourth step to our process of testing and evaluation models before they are ready for being deployed by the social services agency for use in a field trial.

3.2 The two major steps of post-deployment field piloting

There are many aspects and sub-steps to getting our predictive risk model deployed within the agency's supporting information systems and operational workflows and using the outputs of the model within the context of the agency's field pilot (or multiple series of pilots). Many aspects of these steps are highly contextualized and therefore specific to the particular social services agency, the particular type of social services decisions being supported by the predictive model, and the community setting.

For purposes of this interview summary, I highlight two higher level steps that are always important parts of the social service agency's field pilot efforts.

The first step of the field pilot, which typical lasts for much or for all of the first year, is to very closely monitor how the case workers use the predictive model within the context

of the operational and real-world problem setting, and to verify if there are any unintended "adverse effects" directly or indirectly arising from the use of the model. If adverse effects are detected, the pilot would be cancelled, the problems with using the model would be carefully analysed, and the agency would decide whether or not they will address the source of the problems and eventually restart another field pilot Within this first year, there is very careful examination of whether the use of the model is leading to decisions that are biased in one way or another (e.g., by race, by gender), or decisions that are leading to any other type of undesirable concern.

In essence, during this first step of the pilot, the agency is making sure that no one is being harmed as a result of using this new predictive risk modelling tool. If there are no unintended adverse effects detected during this initial step, the pilot continues onward into the second step.

The key aspect of the second step of the pilot is to evaluate the impacts of the model's usage on actual outcomes. This is different and more encompassing than a technical evaluation of the model's predictive capabilities. A technical evaluation of predictive capabilities was already the focus of the four steps that occur *prior to* deployment for field piloting described above. Even though we continue to evaluate the model for these technical aspects during the field pilot, this is not the key concern for this broader evaluation of impacts.

The key concern for this broader evaluation of impacts is whether the use of the model is improving the agency's decision making and actions is ways that lead to the improvements in the relevant social outcomes. Are there fewer occurrences of child harm? Are there reductions in occurrences of social disturbances and harms resulting from better informed decisions related to the allocation of scare supply of housing resources across the large pool of homeless people? Is the social service agency doing what they said and delivering on what they promised to the community when they previously put forth the rationale for developing and using this specific type of predictive risk model as a decision support tool?

3.3 Engaging and informing the community

At the very outset of the project, either prior to the model building effort, or during the model building effort, the agency would have previously held engagement and feedback sessions with representatives from the community to discuss moving forward with this effort. These interactions during these earlier community outreach sessions form the basis of a shared understanding between the agency and the community for the objectives and scope of this predictive risk modelling tool effort and its incorporation into the agency's decision-making process. The feedback from community representatives per their concerns with this effort and per the conditions of their support are the basis of a "social license" to proceed with the effort to develop and use of this new type of decision support tool. The community acceptance and related social license assumes that the agency eventually delivers on realising the positive aspects and objectives for improvement they put forth during those initial discussions.

As part of the second phase of the field pilot, the agency, supported by my team and me as the model developers and experts, goes back to community representatives and discusses the results to-date of this ongoing impact evaluation. We discuss whether and to what extent these results show if there is progress towards the social outcomes and performance objectives promised earlier on. We compare what the agency previously said it wanted to do and achieve related to using this predictive tool to the evidence to-date on what is actually happening.

As the agency is eventually able to demonstrate progress towards the type of benefits and improvements they previously promised, the leadership of the community representatives usually provides their affirmation. In successful situations, community leaders end up saying something along the lines of, “Yes, this is what you said you wanted to do. This is what it seems to be doing. And now we are supportive of the agency’s plan for rolling it out more widely, beyond the scale of the deployment needed for the field pilot.”

3.4 The recent field pilot of one of our predictive risk models at the Los Angeles Department of Children and Family Services

A recent September 2024 press release from the Los Angeles County Department of Children and Family Services provides a good example of the field piloting process we go through with a social services agency with one of our predictive risk models.[9]

The press release states, “The risk stratification model – a data-informed tool that helps social workers serve and support families based on their level of need – continues to show promise.” They note that this field pilot effort has involved the usage of the tool at only a subset (three) of their regional offices and that this pilot evaluation is currently in its fourth year. Based on the data and experience from pilot usage to date, they note the use of the tool has helped managers to prioritize their time consulting on specific cases and improve communication between emergency response and continuing social service workers.

They mention they had provided an update on this field pilot effort at a recent community forum (July 2024), and that the agency has been comparing actual pilot results to concerns previously raised by the community in earlier conversations held prior to undertaking pilot, especially about whether the use of this tool would result in race-based increases in out-of-home placements. The agency found that the use of this tool has not resulted in such increases.

They also noted that based on the findings from these nearly four years of piloting the tool, they are planning to permanently adopt it, expand its use to additional reginal offices, and that as part of ongoing permanent usage, they will continue to “closely evaluate the technology and adapt the pace of the work in order to make informed decisions with ongoing community input.”

3.5 Accepting the reality of multi-year periods for the end-to-end AI project effort

This is a good illustration of the reality that it indeed takes multiple years to do a careful field piloting evaluation of the impacts of using a predictive risk tool to support decision making for these special type of social service situations we are addressing. It takes the first year of the field pilot effort to get the social service agency and the social workers who are the direct tool users familiar with using it in the context of their overall decision making and workflow, and to monitor for the occurrence of adverse effects that might be associated with the transition to making use of the tool. It takes at least several additional years of field pilot effort to have a long enough observation window to accumulate the data needed to assess the extent to which the use the tool to support decision making translates into the border impacts and outcomes that are of ultimate importance (e.g. reduction in child fatalities).

You cannot overly rush this type of field pilot and evaluation effort, especially given the high stakes and highly sensitive nature of the specific types of social service decisions we are supporting. Time is needed to inform the community and bring them along. Sometimes there are changes in the agency's key personnel during the field pilot period and time is needed to bring new agency leaders on board and up to speed on this effort. Time is needed to accumulate the data needed to do the impact analysis to assess if the outcomes of ultimate importance are being realized.

Based on my many experiences of working with public sector agencies over my multi-decade professional career, my assessment is that in the setting of a large and highly complex public sector context like Los Angeles County which has a population of nearly 10 million people, a four-year time cycle for this type of field piloting is actually not a long time, relatively speaking. In fact, compared to many other municipalities, this is brisk progress.

Keep in mind that these nearly four years of effort were only for the field pilot effort that occurred after initial deployment. There was also the additional project time required prior to the field pilot effort for the proceeding stage of model building and technical testing as per the four steps previously described above. So, the entire end-to-end time period for the project is even longer than the field pilot effort.

Public sector officials at all levels (local, state, federal) need to understand the need for these multi-year project time cycles and find ways of supporting the continuity of these efforts over this extended time duration. This is a critical factor for being successful with using of AI-based predictive analytics to support high stakes social services decision making in responsible ways.

Even in other types of public sector domain settings and use cases, end-to-end AI project efforts are typically multi-year efforts because of the care and caution required across all phases of the project lifecycle: the pre-deployment model building and testing, the post-deployment field trials, and follow on iterations for making improvements to operational deployments. AI technology and methods are obviously evolving very rapidly. However, 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.

3.6 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

The terms "testing" and "evaluation" are appropriately used to describe important activities that occur in both the technical stage of building the AI prediction model that precedes deployment and also in the field pilot stage of using the prediction model that follows after deployment. The term "validation" is also often used in the context of both these pre and post deployment stages as well. However, these terms refer to different things in the pre-deployment model building stage than in the post-deployment field piloting stage.

I have gone to many professional conferences and meetings that discuss the use of AI models in public sector and other settings. There have always been many more presentations on the technical aspects of evaluating the predictive performance of a new AI model using only historical test data sets for training and testing, without an accompanying field pilot effort. There are far fewer presentations on evaluating post-deployment field pilot (or ongoing operational usage) impacts to evaluate if the use of the model in the real-world domain setting made a difference on the outcomes of ultimate interest.

This situation is especially the case with technology focused conferences and professional meetings. So many examples are presented where R&D teams develop a new AI model and spend a lot of money and time to demonstrate its predictive accuracy only by testing it against a pre-existing data set. Far fewer examples are publicly presented where teams are studying what actually happens as a result of usage of the model in the hands of the actual human users doing their work in the real-world domain setting. This is why I have been attending more professional meetings in the policy analysis and economics communities lately as they put more emphasis on rigorous analysis of broader impacts and relevant outcomes in post-deployment settings of using AI.

It is obviously important to do rigorous and careful testing of a new AI-based predictive model trained on a data set prior to deployment. This is an essential pre-requisite for moving to the next stage of a deployment for a field pilot. My caution to public sector decision makers involved with shaping AI strategies within their organisations is to not restrict your considerations of AI capabilities, limitations and risks to the many technically oriented assessments based on the use of training and test data sets. Make sure to additionally seek out evidence from post-deployment evaluations that examine the use of the AI tool in the real-world domain setting.

3.7 Views on the role of using Generative AI with unstructured data vis-à-vis AI-based predictive modelling with structured data

The need for building public sector decision support tools using the kind of machine learning models, including deep learning models, for making the types of risk-related predictions that we have been discussing in this interview is not going away, even with all the recent attention given to GenAI and the related large language models.

Just because we now have GenAI and the resulting large language models, the need for predictive models built from structured data sets still persists and remains important. If anything, the availability of structured data and the need for using it as part of making better decisions with the support of machine learning-based predictive models is only going to increase as the sources of structured digital data continue to increase. GenAI-based large language models have very impressive capabilities, but their core strength is not for using structured data to carefully create and test predictive models, though they can play an important role in supporting this overall workflow.

I think GenAI does have an important role to play in the types of settings we have been discussing in this interview. Over time, I expect that certain types of GenAI applications will increasingly be incorporated into supporting aspects of social service agency workflows.

One preliminary idea we are now discussing with some of the social services agencies we work with is to use a large language model to summarize cases—initially not to make recommendations or decisions, but just to summarise the case history in ways that would be useful to the social worker. A large language model could summarise a large case file, the equivalent of tens or even hundreds of pages of case notes.

This is a practical and important need. Often a social service case worker is transferred to a new case they are unfamiliar with, and they need a quick way to get a briefing on the case history. Even if a case worker is familiar with a case they have been working on over a few year period, they have frequent occasions where they must refresh their memory of key points: on details on what the agency has done with regard to this person over time, on all of the services this person has received to-date, and on the overall chronology of the case history. It would be useful for the social worker to be able to use a large language model to quickly generate this type of summary.

Earlier in this interview I mentioned that social services case workers are juggling a huge case load, do not have research assistants or junior underlings helping them manage all of the details, and cannot retain in their memory all of these details across multiple cases they are supporting in parallel. The ability to use GenAI based large language models to create these types of case summaries could help in this regard.

I have considered the possibility of whether LLMs could process a large number of case files and come up with new types of features that would be useful for us to use in our predictive models, feature that we have not yet identified as being important based on our experience to date with designing and training our models with structured data sets. While we have not attempted to do this type of experiment, I do not think this would be very helpful.

We already use many features across the portfolio of the different types of predictive models we have developed to date (approximately 1,500) that have been derived from the many structured data sets that we work with. In the types of social services decisions we support, as well as in many other public and private sector high stakes decisions, the key sources of data are structured data sets that have been carefully curated and managed over time. In the organisational settings we work in, these structured data sets are regarded as the most reliable and coherent sources of data. In most situations involving high stakes operational decisions, front-line decision-making staff usually do not rely on free text or other forms of unstructured data to make their case assessments and situation specific decisions. Rather, they mostly use and rely on the information within structured data sets because they know this is usually the most trustworthy and reliable source of information about the situation.

From what we have seen in social service agency settings, the quality of free text information added to the case files is too variable. For example, there is a lot of pro forma language and cutting and pasting. As an illustration, I recall shadowing a senior worker who showed me how she helps junior workers by providing them all with “cheat sheets” of pre-written verbiage that workers can cut and paste where free text is required. As such, I feel that using LLMs to derive prediction model features from these archives of free text comments that may be part of the case record is not going to lead to new and useful types of features that will help us to build and validate new and better versions of our predictive models.

LLMs are very good for summarizing data from the web. They are very good for summarizing spoken language either in the form of audio files or written transcriptions. However, these bodies of content are not required to be as precise as the content of the fields of a structured data base used as the official “system of record” and source of truth for supporting an agency’s high stakes processes and decisions.

Following this line of thinking, I think LLMs can also be used very effectively for helping a case worker to review their own or their co-worker’s free text notes or recorded spoken comments that were created as a supplement to the structured data fields in the case files. Case workers or their supervisors sometimes add these types of free text or spoken input annotations to remind themselves as to why they made a decision in a particular way, and to capture special aspects of context or circumstances they thought were worth noting. An LLM can quickly provide a summary of these types of unstructured data annotations to help the case worker remember their own prior thinking process and decision rationale when they made their prior assessments and choices.

In these types of high stakes settings we are working in, my intuition as of now is rather than using a LLM for designing model features or other aspects of a predictive model, it will be more useful as well as more reliable to deploy the LLM for summarizing cases. This includes summaries based on structured data records as well as summaries based on the unstructured text and voice data used for supplemental annotations. Case workers would find it very useful to have such summaries.

It is so important for leaders and managers in all types of public sector agencies, including the social service type agencies I specialize in working with, to realize that GenAI is not the only type of AI they should be paying attention to and considering for use. And related to this, to also realize that the types of machine-learning AI that they may have been using over the past 7 to 10 years to build better and better types of prediction models are not "obsolete". They will continue to be used in many ways. I suspect that there will be many new ways in which GenAI capabilities (drawing from unstructured data sources) will be combined with machine learning-based predictive AI capabilities (drawing from structured data sources).

3.8 A transitional Catch 22 situation with getting municipal level social service agencies familiar with using GenAI

One thing to keep in mind is that as of right now, it is still the case that most local government agencies keep their data related to child protection, homelessness, and other areas of high sensitivity on-premise and not on a public cloud. They may not want to place this type of data on a public cloud infrastructure even with all the steadily improving security and privacy protections that the public cloud infrastructure can provide for a government user.

At this moment in time, many local-level agencies do not have the appropriate types of processors (graphics processing units) within their own on-premise IT infrastructure to run large language models, even for running the smaller size LLMs that have been released open source and can be run on-premise.

As such, we are working through the Catch-22 situation where the social service agency cannot move ahead with more specific planning on how to use GenAI- especially large language models – until they see a demonstration of what it can do for them. Yet, they cannot see this demonstration unless they can run these models within their own on-premise infrastructure and they don't yet have a way to do this.

Creating the necessary on-premise infrastructure will cost them some additional money but they don't want to spend this money - even if they are able to work it into their budget – until they can see what LLMs can do for them, and there are data restrictions constraining the extent to which they can use public cloud infrastructure off-premise to run these demonstrations.

Sooner or later, there will be a way out of this type of Catch-22 situation for most municipal level social service agencies. At the moment, this type of situation is an obstacle for many agencies.

4. Concluding thoughts and suggestions for public sector agencies

4.1 Understand the reality of how noisy decision making actually is unless front-line workers making the decisions have good support

Front-line workers in social services agencies face daunting challenges, and routinely must make difficult, high stakes decisions about these situations. Under these circumstances,

there is bound to be a lot of noise (randomness) in how these decisions are made. There are still many social services agencies where senior management does not understand just how random this decision making can be.

That's why I often suggest to senior management of an agency to spend time on the ground with their front-line workers to see the realities of the conditions under which they must operate, to see why it is unavoidable for there to be so much randomness in the everyday front-line decision making. Once the senior management of an agency understands this, they can decide what actions they can take to better support their front-line workers in making these high stakes decisions.

It is all too common for an agency's senior management to sit in an office setting and work with staff to create and study diagrams of how their "system" (including people, processes and supporting information systems) is supposed to function. What is missing from this type of idealized view is that the most important parts of the "system" are the humans on the front-line making sequences of decisions under time pressure and many other constraints. While they are "getting the job done" as best they can under the circumstances, there is a lot of randomness in how these decisions are being made. It is no wonder that families or individuals being served by these case workers sometimes feel like they are living in a pinball machine.

I highly recommend that public sector agency managers read or look through the book Noise: A Flaw in Human Judgement by Daniel Kahneman, Oliver Sibony and Cass Sunstein that was initially published in 2021 and then republished in following years.[10]

The book explains how errors in judgement, including making risk predictions, are the result of both bias and noise. Because there is already strong general awareness of the problems associated with bias in decision making and risk prediction, organisations end up putting effort into reducing the effects of bias. However, decision making noise (unintentional randomness) often goes unnoticed and unattended to. These authors show the detrimental effects of noise in many fields, including medicine, law, economic forecasting, forensic science, bail, child protection, strategy, performance reviews, and personnel selection. They explain why wherever there is human judgment (decisions, risk predictions), there is noise, and usually a much higher degree of noise than the front-line decision makers or their senior management is aware of. They highlight how reducing decision-making related noise in and of itself can substantially improve the quality of an organisation's decision making.

The tools we build help in reducing the influences of both bias and noise on decision making.

4.2 Use AI-based decision support for helping human decision makers, not for criticizing past or current performance

When I start a new engagement with a social services agency, I am careful to clarify that I am not there to criticize the prior or current approach to decision making or to "correct" it. I communicate to front-line staff and to their supervisors that the purpose of our work is

not to “check” their decisions, but rather to help them to improve their decision making as they go forward.

As we work together with front-line staff and their management to build, test and then field pilot the tool, they come to see for themselves that the support they receive from using the tool aids their judgement in identifying the families or individuals at higher risk, and also aids in reducing the influences of bias as well as noise. The case workers realize that if they can put more of their time into the higher risk cases and do this in a fairer and less random way, there is more value to their work efforts and better outcomes for the agency.

4.3 Emphasizing the functionality and use of the support tool more so than emphasizing the use of AI for its own sake

We refer to our tools as “predictive risk models” because that is the essential nature of what these tools are being used to do. Rather than refer to them more generically as “AI algorithms”, we think it is more useful to explicitly refer to the functionality that these models are being used for. In our setting, that functionality is to predict risk in ways that support specific types of decision making for targeted social services high stakes use cases.

When we explain how we create the model to internal agency staff as well as to the external community, we of course describe the specific methods we use, including the use of AI in the form of a particular machine learning method for building the appropriate type of predictive model.

While we are transparent about our objectives and methods, and our use of AI, we minimize the use of the word AI in our conversations with management and staff, and to the community, to keep the focus on the essence of the problem, on the decision making, and on the best tooling they can use in practical ways to support their assessments and decisions. We don’t want agency personnel or the community to get pre-occupied with “using AI” for its own sake, or with viewing AI as some abstract concept, or some mysterious or magical thing. We want our users, their management and broader stakeholders to be focused on the functionality provided by the support tool, on the tool’s capabilities and limitations, and on how to use the tool for the relevant type of risk prediction as part of supporting the case worker’s overall decision making.

This helps to demystify the tool and the process of using it. It brings the concept of using “AI” down to a practical and understandable level. Also, by emphasizing that this is a tool with a targeted functionality and limited scope, it makes it more natural for the user to keep in mind that like any tool, this tool also has its limitations, and therefore the user must always be alert and careful when making use of the tool.

4.4 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

An important aspect of being a good professional in my chosen area of work is to acknowledge that there are a wide range of views related to the use of these types of

predictive risk models in high stakes social services decisions. In terms of understanding the impacts of using these tools in real-world field settings, there is strong evidence of benefit and promise, as well as evidence and information related to controversy, confusion, and even misinformation. A summary of this situation is given in a 2023 commentary written by Child Welfare Monitor blog publisher Marie Cohen.[11]

The social services agencies we work with are addressing very complex human problems. There are no easy solutions or perfect solutions for these types of high stakes situations and related decisions. While the AI-based predictive risk tools we are building and very carefully testing, piloting, validating and evaluating with the agencies have their limitations, the growing body of evidence is showing they are helpful and beneficial, and when designed and used properly, not harmful.

There are voices who argue that predictive AI should not be used in these types of situations. However, what are the alternatives? Without any type of algorithmic decision support, the situation for these agencies, for the front-line social workers and for the clients they serve will only get worse as caseloads continue to increase, case worker staffing challenges persist, and agency budgets and support resources become further constrained.

Without some type of algorithmic decision support for the case worker, it is unavoidable that both bias and noise would further increase in the everyday case worker decisions, and this would be to the further detriment of the clients who need the help.

Given the practicalities of the real-world situation faced by these social services agencies, I believe the most sustainable pathway forward for helping them to handle these high stakes decisions, especially in larger urban metropolitan areas with high social services demand, is to continue working with them in the way that we are to very carefully develop and use predictive risk tools that are used for decision support. And the best way to do this at this point in time is to make use of AI in the form of machine learning methods and to also use whatever are the best analytic methods and sources of data available.

Endnotes

[1] For Prof Rhema Vaithianathan's background, visit https://academics.aut.ac.nz/rhema.vaithianathan. For information on Centre for Social Data Analytics at Aukland University of Technology, visit https://csda.aut.ac.nz/.

[2] R Vaithianathan, T Maloney, N Jiang, I De Haan, C Dale, E Putnam-Hornstein, T Dare & D Thompson. Report Prepared for the Ministry of Social Development, Auckland NZ by the Centre for Applied Research in Economics (CARE), Department of Economics, University of Auckland. "Vulnerable children: Can administrative data be used to identify children at risk of adverse outcomes." September 2012. Also see:

- R Vaithianathan, T Maloney, E Putnam-Hornstein & N Jiang. American journal of preventive medicine. 45(3). "Children in the public benefit system at risk of maltreatment: Identification via predictive modelling." September 2013.
- A Blank, F Cram, T Dare, IA de Haan, B Smith, R Vaithianathan. Report Prepared for the Ministry of Social Development, Auckland NZ. "Ethical issues for Māori in predictive risk modelling to identify new-born children who are at high risk of future maltreatment." January 2015.
- R Vaithianathan, B Rouland, E Putnam-Hornstein. American Academy of Pediatrics 141(2). "Injury and Mortality Among Children Identified as at High Risk of Maltreatment." February 2018.

[3] See the additional discussion on the Allegheny County Department of Human Services and their integrated data platform in the interview summary with Prof. Ramayya Krishnan from Carnegie Mellon University.

[4] Allegheny County Department of Human Services. Allegheny Family Screening Tool. https://www.alleghenycounty.us/Services/Human-Services-DHS/DHS-News-and-Events/Accomplishments-and-Innovations/Allegheny-Family-Screening-Tool.

[5] R Vaithianathan, E Putnam-Hornstein, N Jiang, P Nand & T Maloney. Report Prepared for the Allegheny County Department of Human Services by the Centre for Social Data Analytics, Auckland University of Technology, New Zealand. "Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation." April 2017. Also see:

- A Chouldechova, D Benavides-Prado, O Fialko & R Vaithianathan. Proceedings of Machine Learning Research Vol 81: Conference on Fairness, Accountability and Transparency. "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions." February 2018.
- A Brown, A Chouldechova, E Putnam-Hornstein, A Tobin & R Vaithianathan. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. "Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services." May 2019.
- R Vaithianathan, E Putnam-Hornstein, A Chouldechova, D Benavides-Prado & R Berger. JAMA PEDIATRICS Vol 174(11). "Hospital Injury Encounters of Children Identified by a Predictive Risk Model for Screening Child Maltreatment Referrals: Evidence From the Allegheny Family Screening Tool." August 2020.
- R Vaithianathan, D Benavides-Prado, E Dalton, A Chouldechova & E Putnam-Hornstein. AI Magazine. Vol 42(1). " Using a machine learning tool to support high-stakes decisions in child protection." April 2021.
- K Rittenhouse, E Putnam-Hornstein & R Vaithianathan. Unpublished working paper. "Algorithms, Humans and Racial Disparities in Child Protective Systems: Evidence from the Allegheny Family Screening Tool." June 2023.

[6] R Vaithianathan & CI Kithulgoda. Report Prepared for the Allegheny County Department of Human Services by the Centre for Social Data Analytics, Auckland University of Technology, New Zealand. "Using Predictive Risk Modeling To Prioritize Services for People Experiencing Homelessness in Allegheny County: Methodology Paper for the Allegheny Housing Assessment." September 2020. Also see:

- CI Kithulgoda, R Vaithianathan & DP Culhane. Journal of Technology in Human Services Vol 40(2). "Predictive risk modeling to identify homeless clients at risk for prioritizing services using routinely collected data." April 2022.
- L Cheng, C Drayton, A Chouldechova & R Vaithianathan. ArXiv preprint working paper. "Algorithm-Assisted Decision Making and Racial Disparities in Housing: A Study of the Allegheny Housing Assessment Tool." August 2024.
- CI Kithulgoda, R Vaithianathan & C Parsell. Housing Studies Vol 39(8). "Racial and gender bias in self-reported needs when using a homelessness triaging tool." September 2024.
- R Vaithianathan & CI Kithulgoda. Research Handbook on Homelessness, Chapter 6. "Predictive risk modelling and homelessness." August 2024.

[7] R Vaithianathan, D Benavides-Prado & E Putnam-Hornstein. Report Prepared for the Allegheny County Department of Human Services by the Centre for Social Data Analytics, Auckland University of Technology, New Zealand. "Implementing the Hello Baby Prevention Program in Allegheny County Methodology Report." September 2020. Also see J Reuben, R Vaithianathan & R Berger. Child Abuse & Neglect Vol 151. "Identifying infants at risk of sudden unexpected death with an automated predictive risk model." May 2024.

[8] R Vaithianathan & CI Kithulgoda. Chapter 6 in Research Handbook on Homelessness, edited by G Johnson, et. al. "Predictive risk modelling and homelessness." August 2024.

[9] Los Angeles Country Department of Children and Family Services (DCFS) press release. "New Analysis of DCFS Data Shows Improvement in Safety with Use of Data-Informed Technology," September 05, 2024.

[10] D Kahneman, O Sibony and C Sunstein. Noise: A flaw in human judgement. https://mitpressbookstore.mit.edu/book/9780316451390 May 2022.

[11] Child Welfare Monitor blog. Using algorithms in child welfare: promise, confusion and controversy. https://childwelfaremonitor.org/2023/04/05/using-algorithms-in-child-welfare-promise-confusion-and-controversy/ April 05, 2023.