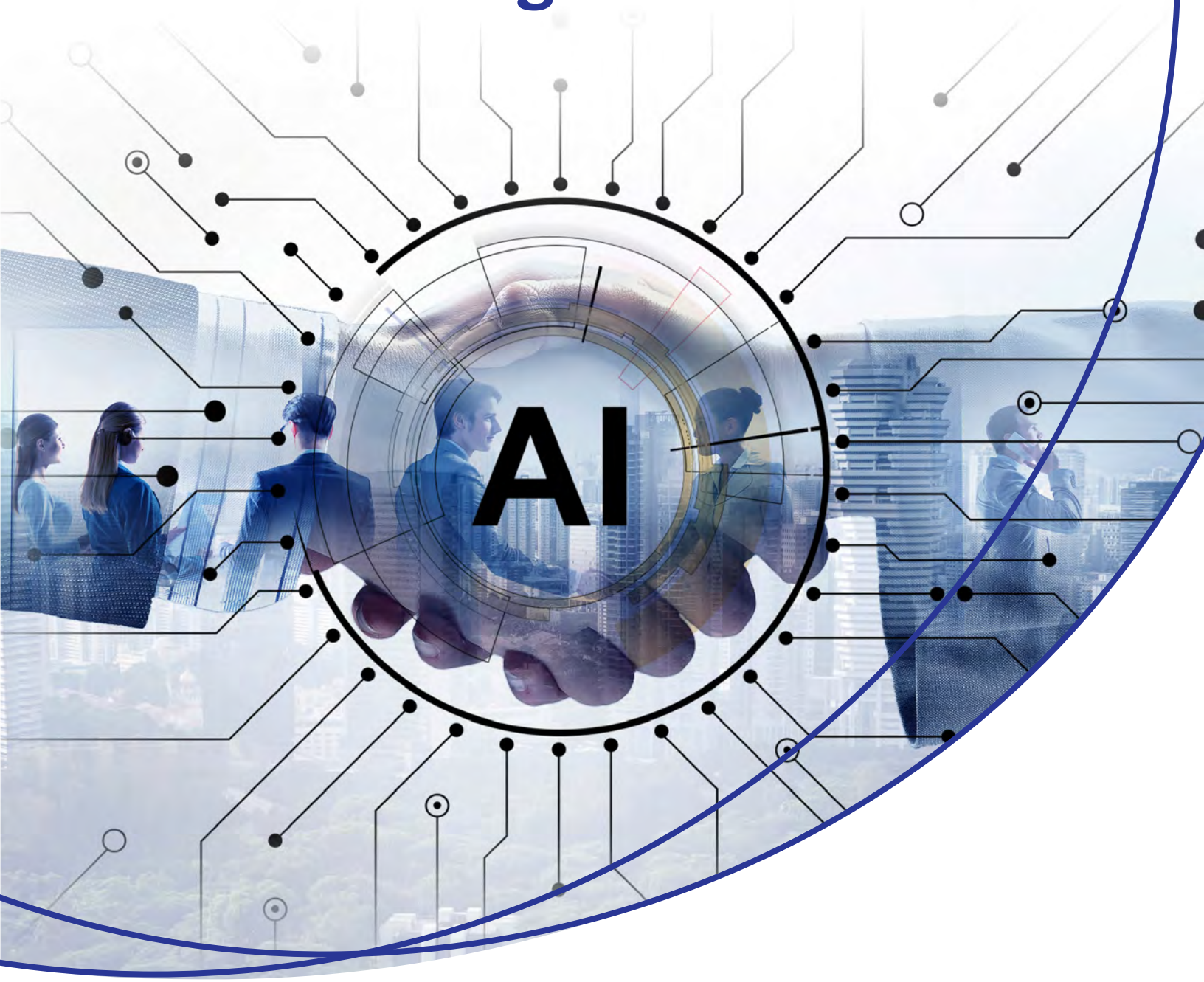
Pathways to Trusted Progress with Artificial Intelligence



Kevin C. Desouza
Queensland University of Technology

Dr. Gregory S. Dawson
Arizona State University

IBM Center for
The Business
of Government

Pathways to Trusted Progress with Artificial Intelligence

Kevin C. Desouza
Queensland University of Technology

Dr. Gregory S. Dawson
Arizona State University

JULY 2023

25TH ANNIVERSARY
IBM Center for
The Business
of Government

TABLE OF CONTENTS

Foreword 4
Executive Summary 6
Introduction 8
Findings from Workshop 14
Recommendations for Building Trusted AI in the Public Sector 17
Conclusion 21
About the Authors 22
Key Contact Information 23
Recent Reports from the IBM Center for The Business of Government 24

FOREWORD

On behalf of the IBM Center for The Business of Government, we are pleased to release this new report, *Pathways to Trusted Progress with Artificial Intelligence,* by Kevin Desouza with the Queensland University of Technology and Gregory Dawson with Arizona State University.

DANIEL J. CHENOK

Artificial intelligence (AI) has proliferated across all sectors of society. National governments have created AI-related strategies, frameworks, and guidelines on the ethical use of AI. Yet while people have faith in AI to produce good and reliable outcomes, they have questions about the safety and security of AI systems. Specifically, this concerns public trust in AI itself, and trust in government to develop mechanisms to successfully deploy and manage such a powerful technology. These issues cover trust in AI in the context of design, development, deployment, and evaluation of public services and public policy.

Information technology leaders in government have the power to play a significant role in future directions for AI that build trust. AI has the vast potential to be a change for good. It can change how governments lead, make decisions, and serve nations for future success. The governance and applications of AI is an important conversation government and industry must all have to help address the needs, security, and progress of delivering services that benefit citizens and industry.

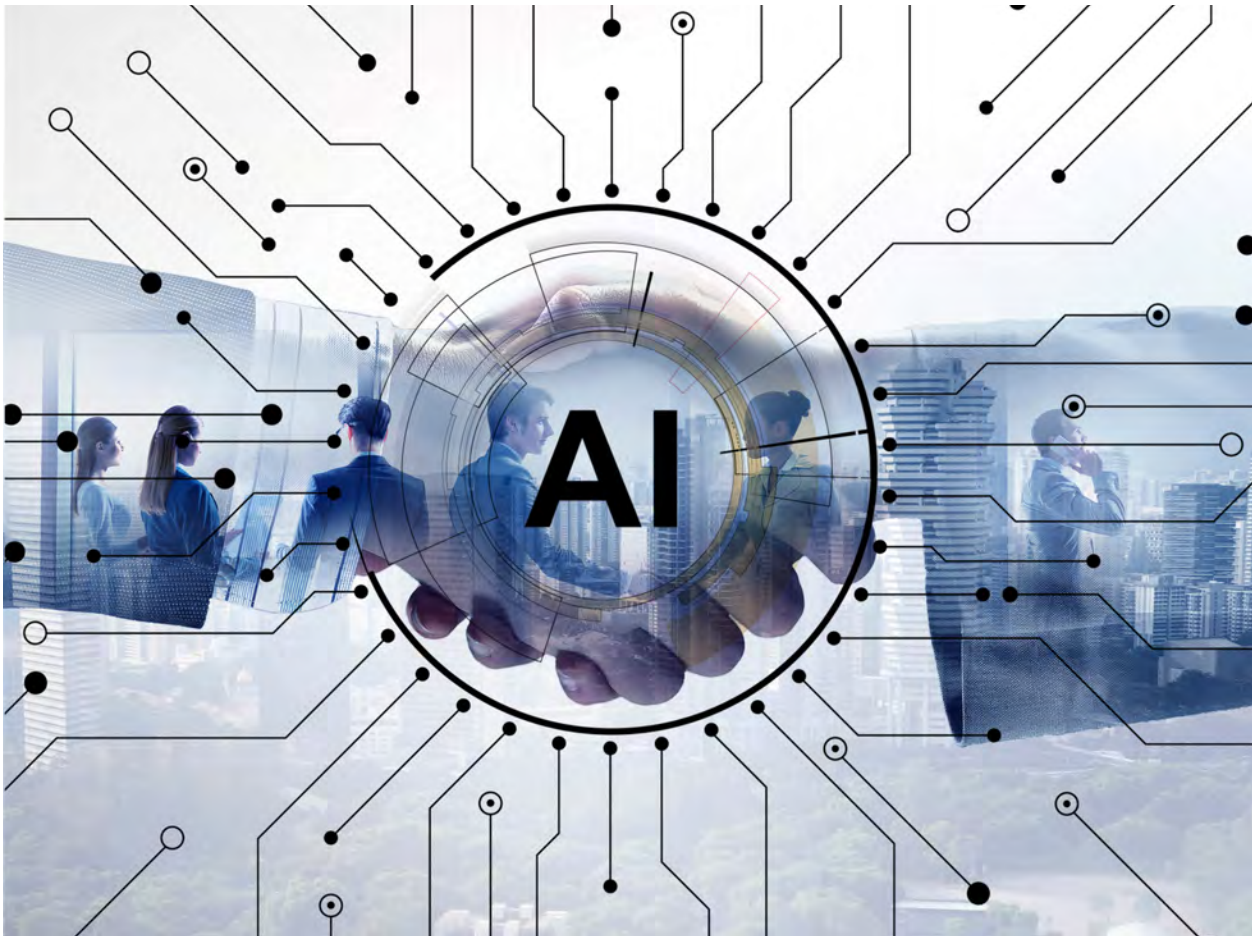
NICHOLAS FLOOD

This report, which distills perspectives from an expert roundtable of leaders in Australia, discusses major questions to help inform government decision making and design principles, including:

- How can governance be an enabler of action and trust, rather than an inhibitor of progress?
- How can AI help to navigate the nuances of meeting government and citizen needs?
- What are best practice insights from other governments? How are these outcomes measured?

Insights from experts as reported in this report focus on how governments need to develop and communicate a framework for the public to understand why AI is being used, what has been done to ensure that the AI is fair, transparent, and accurate, what experiments were done to ensure that the output is reliable, and how public value from AI is being measured and created. By addressing the growth and management of AI, and the governance of data aligned to AI strategies, government can take full advantage of the power of AI.

The authors also explore case studies, addressing the potential that AI has to transform how government agencies interact with citizens, along with risks that can arise when AI is left unchecked.



This report builds on considerable research that our Center has led about how Australia and other governments can implement AI in ways that build trust, including *Charting the Course to Tomorrow's Trusted Digital Services*, *Artificial Intelligence in the Public Sector: A Maturity Model*, *Risk Management in the AI Era*, and *More Than Meets AI*.

We hope this report helps government leaders across the globe implement pathways to use AI in ways that build public trust.

Daniel J. Chenok
Executive Director
IBM Center for
The Business of Government
chenokd@us.ibm.com

Nicholas Flood
Vice President IBM Technology
Country Leader IBM Australia
nicflood@au1.ibm.com

EXECUTIVE SUMMARY

Implementing artificial intelligence (AI) in the public sector is now a certainty, and governments across the world are moving aggressively into AI to realize its benefits. These efforts have shown the promise of AI, as well as challenges that exist when it comes to engendering trust.

In reaction to larger world events, trust in government is on the decline in many nations. This directly impacts how much citizens are willing to trust government in the implementation of any powerful new technology. How can public sector leaders create trust in AI, given declining trust overall in government? To address this challenge, the IBM Center for The Business of Government hosted a recent forum of senior Australian government officials, who identified several major themes of AI in government:

- **Theme 1—Government is in the business of providing services, and AI is simply a tool to facilitate that.** Government should remain focused on providing government services, and not get "techno dazzled" by AI.
- **Theme 2—Government is held to a higher standard of performance regarding AI versus private companies, making explainability and transparency of utmost importance.** Citizens expect government to get things right, and the services facilitated by AI should be sufficiently transparent and fully explained to the citizen.
- **Theme 3—Government needs to work holistically in terms of defining AI standard practices, operating models, etc.** There is too much work and too many risks in implementing AI for standards development to happen only at the departmental level. Rather, this work needs to be coordinated at the highest level of government.
- **Theme 4—Adequate governance is necessary not only for AI technology, but also for the people who build AI systems and the processes used to build them.** Issues emerge not only from the technology itself but also from the people and processes that implement AI.
- **Theme 5—There is a need to distinguish between different types of AI (fully autonomous, semiautonomous, and augmented) in establishing guidelines and approaches.**[1] Not all AI is the same, and costs, benefits, and risks differ for each type of AI. Discussing AI at a more granular can ensure optimal uses.

These themes, coupled with background work done by the authors, gave rise to several recommendations:

- **Recommendation 1—Promote AI-human collaboration when appropriate.** Different kinds of AI call for different levels of human involvement, and citizens are generally more comfortable with a human being involved in providing direct services.
- **Recommendation 2—Focus on justifiability.** Justifiability can be thought of as an outwards-facing business case, and with citizens as a primary audience. The government needs to article why an AI system needs to be developed, the amount of human involvement, and execution strategies.

1. https://newsroom.ibm.com/Whitepaper-A-Policymakers-Guide-to-Foundation-Models.

- **Recommendation 3—Insist on explainability.** Government must be able to explain why the AI came to a proposed decision, including the data that was used for the decision. This becomes increasingly important with decision making for high-stakes outcomes.
- **Recommendation 4—Build in contestability.** Just as citizens can appeal to a person in government about the fairness of a decision, they also need to be able to contest the decisions made with AI. This feedback loop helps ensure that decisions are reasonable and not prone to bias.
- **Recommendation 5—Build in safety.** While AI is deployed, risks can arise that make a safety feedback loop important. Government needs to either create or join an incidents tracking database to capture and act upon feedback.
- **Recommendation 6—Ensure stability.** The machine learning function in AI means that supporting algorithms will be constantly tweaked in response to new information. Not only does the AI system need auditing prior to implementation; regular examinations will ensure that AI provides stable results.

Use of AI will continue to grow, and very likely will become a major delivery mechanism for many government services. Government leaders can act now to implement fundamental recommendations to ensure successful AI delivery.

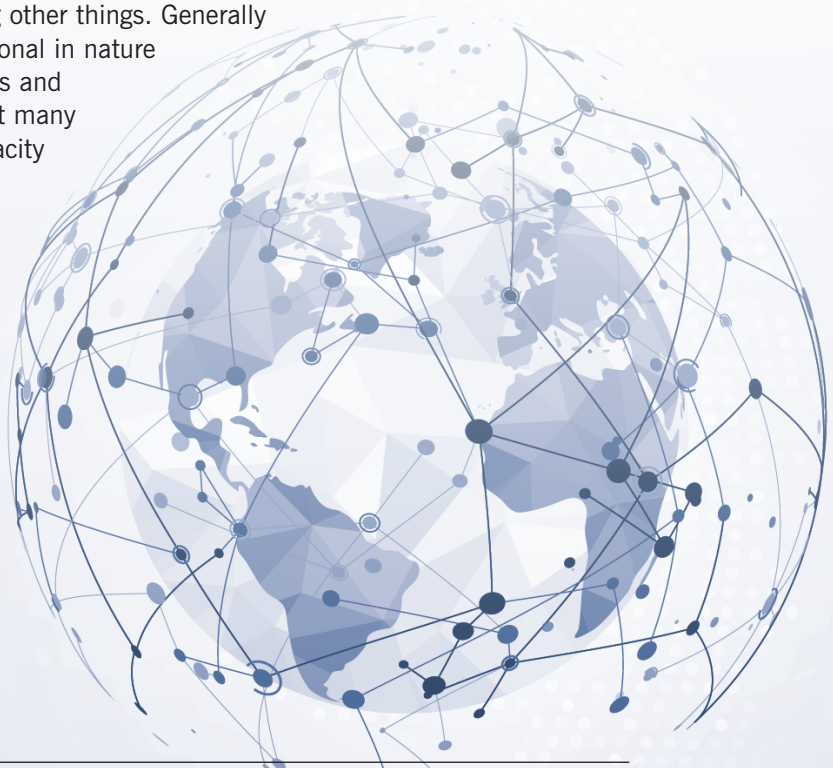
INTRODUCTION

Digital transformation initiatives are revolutionizing all aspects of the public sector.

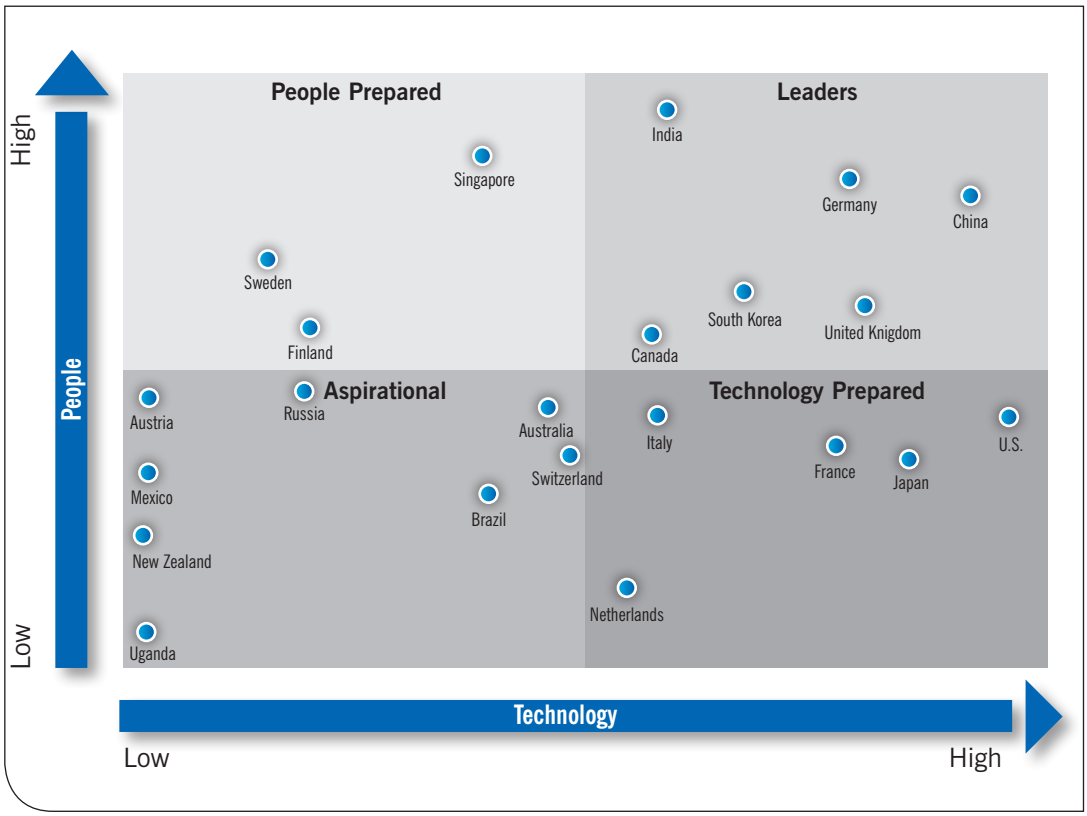
> Data and digital are inextricable from the delivery of government information products and services.
>
> —*Chris Fechner*
> *Chief Executive Officer, Australia's Digital Transformation Agency*

Information systems, from basic such as e-government portals to pay taxes, to more sophisticated technologies such as robotic process automation (RPA) that have automated manual tasks of sorting and classification of artefacts, are now the primary vehicles through which any public agency achieves its mission objectives. Over the last few years, the public sector's interest in a particular class of information systems—artificial intelligence (AI)—has received significant attention.

At present, 44 countries have developed and published a national AI strategic plan. These plans often address the country's public functions (e.g., immigration, education and public safety), industries (e.g., financial, agriculture and defence), and approaches to data (e.g., privacy, regulations, and sharing), among other things. Generally speaking, the plans are lofty and aspirational in nature (e.g., addressing societal inequality issues and intellectual property right protections) but many are also very clear about addressing capacity approaches (e.g., pilot projects, tax incentives, etc.). Global spend on AI is estimated to be $3 trillion (AUD) by 2030.[2] Figure 1 shows the relative positioning of various national governments in their national AI-preparedness.[3]



2. https://www.statista.com/statistics/1365145/artificial-intelligence-market-size/#:~:text=According%20to%20Next%20Move%20Strategy,a%20vast%20amount%20of%20industries.
3. https://www.brookings.edu/blog/techtank/2021/10/21/winners-and-losers-in-the-fulfilment-of-national-artificial-intelligence-aspirations/.



Early this year, Chris Fechner, who leads the Australian Government's Digital Transformation Agency, and is also head of the Digital Profession tasked with uplifting digital capability across the Australian Public Sector, noted:

> The technologies available to us today have infinite abilities for good if they are administered and leveraged by people who are conscious and aware of the implications of how data and digital are to be used. We need to make sure that we've got people who are talking about digital democracy, how we use technology to support better customer experience, how we break down silos in government, how we share information for the benefit of people and businesses, and how we make our policy decisions work much better through using technology while not forgetting to keep people at the centre of our design. These are the things that will make the biggest difference to the people and businesses of Australia. This is how we will transition to a future ready economy.[4]

AI systems will play an important role in transforming government as well as the national economy. Realising AI's potential will only occur if there is a concerted effort to ensure that citizens trust AI systems, the government, and the government use of AI. Consider one category of systems that use AI—facial recognition technology that is commonplace today. Individuals use it to access their mobile phones, the Australian Border Force uses it for screening passen-

4. https://www.dta.gov.au/blogs/transformative-role-data-and-digital-next-gen-public-sector.

gers, and cities use it for law enforcement. These applications are adopted and accepted, because of either low trust concerns (as in the case of individuals who give their mobile phone the permission for facial recognition) or the benefits from using them outweigh the cost (as is the case for the Border Force). However, the nuances when it comes to issues such as ensuring privacy, responsible data use, and data security, are non-trivial. In her keynote remarks to launch the 2023 Privacy Awareness Week, Angelene Falk, Australian Information Commissioner and Privacy Commissioner, noted:

> Facial recognition tools carry heightened privacy risks as they can be used to uniquely identify individuals. Facial features are difficult or impossible to change and they can be used to estimate or infer other sensitive or personal information such as age, sex, gender, and ethnicity. Automated facial recognition systems can collect large amounts of biometric information indiscriminately and without any direct involvement or even knowledge of individuals. This limits the effectiveness of traditional privacy self-management mechanisms such as notice and consent to provide individuals with control over their personal information.[5]

There is a wide assortment of AI systems, and each class of AI systems has their own characteristics. However, at their core, these systems ingest vast swaths of data, employ either supervised or unsupervised learning techniques or both, and can be deployed autonomously, semiautonomously, or in an advisory capacity to augment human decision makers. Consider the following three examples of AI systems successfully deployed in the public sector:

- **Fully autonomous:** The State of North Carolina (USA) uses AI-based chatbots to free up operator telephone lines and customer help desks. Most of the service questions are simple and repetitive (for example, approximately 90 percent of requests are password resets). The use of these chatbots, which require no oversight or human intervention, allows customer service agents to focus on more complex and time-sensitive tasks.[6]
- **Semiautonomous:** At colleges across the world, semiautonomous robots are delivering food to hungry college students. A student orders food on an app. This triggers a robot to be dispatched to the restaurant to pick up the food and then dispatched to the dorm where the food was ordered from. Humans are responsible to "set" the restaurant and delivery location, but the robot makes all of the operational decisions in terms of the route, its speed, avoiding pedestrians and crossing streets.[7] This technology is being examined for use in search and rescue operations.
- **Augmented decision making:** Wildlife rangers use AI to protect native African animals and plants more effectively. In this case, the AI recommends which wildlife territories to patrol based on the AI's prediction about where poachers will set their traps. In this case, the AI provides actionable insights, but it is up to the human to decide what to do.[8]

While there have been plenty of successful deployments of AI systems, there have also been challenges. Consider the following three examples:

5. https://www.oaic.gov.au/newsroom/privacy-awareness-week-2023-launch.
6. Stamatis, A., A. Gerontas, A. Dasyras, and E. Tambouris. (2020). Using chatbots and life events to provide public service information. Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance, 54–61. https://doi.org/10.1145/3428502.3428509.
7. Autonomous food-delivery robots roll out on ASU's Tempe campus (azcentral.com)
8. National Science Foundation, "Outwitting poachers with artificial intelligence." Retrieved from http://bit.ly/2oBRTLy.

- **Rotterdam's fraud detection AI:**[9] In 2017, Rotterdam deployed a machine learning algorithm. The system assigned a "risk score" for welfare recipients that then triggered investigation. Items used to determine the risk score included, among other things: recipient's age (younger is worse), gender (female is worse), romantic relationship status (no partner is worse), appearance (ill-kempt is worse) and number of children (more is worse). Rotterdam used the system for three years until a Dutch court ordered its immediate halt because it violated human rights.[10]
- **Britain's Universal Credit AI:**[11] Great Britain's Universal Credit program is designed to provide financial assistance to citizens at an amount based on how much the person earns. However, due to an overlooked programming error in the AI, the system failed to properly account for those paid multiple times a month (common for lower wage earners), leading to significant underpayments to such individuals. Because the system was not properly developed and overseen, it had the potential to put recipients in poverty until the error was caught and corrected.
- **Risk assessment bias:**[12] The use of risk scores in sentencing criminals is common in the U.S., and Northpointe is one of the largest providers of algorithms to produce such scores. Northpointe uses 137 questions—such as "Do you think it is right to steal if you are hungry?"—to calculate risk scores. Unfortunately, the algorithm, which Northpointe refuses to release, has been found to contain racial biases that heavily influence these risk scores and resulting sentencing of convicts.

Clearly, the above cases have impacted individuals and caused harm. This has led to the question of *how should AI systems be deployed in the public sector?* Central to answering this question is the question of how governments can generate and maintain *public trust* when it comes to the design, development, and deployment of AI systems.

The Problem of Trust

Trust is a multidimensional concept that can be broken down into three components—ability, integrity, and benevolence.[13] See Table 1.

Table 1: Trust Elements

Trust Element Definition	Example in Government	Example in AI
Ability—Belief in the competency of the trust target	Belief that government can provide national security	Belief that the AI can correctly and consistently give the correct answer
Integrity—Belief in trustee's ability to adhere to a set of ethical principles	Belief that government will treat all people equally regardless of their gender or ethnicity	Belief that the AI will mirror society's view of ethical principles
Benevolence—Belief that the trustee wants to do good to the trustor	Belief that government will act in the best interests of the citizen	Belief that the AI has good intentions (or not negative intentions) in its functioning and outcomes

9. https://www.wired.com/story/welfare-state-algorithms/.
10. https://www.theguardian.com/technology/2020/feb/05/welfare-surveillance-system-violates-human-rights-dutch-court-rules.
11. https://www.thelondoneconomic.com/politics/poorly-designed-universal-credit-algorithm-forcing-people-into-hunger-and-debt-203635/.
12. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
13. Mayer, Roger C., James H. Davis, F. David Schoorman, "An Integrative Model of Organizational Trust," Academy of Management Review, Vol. 20, No. 3 (1995), pp. 709-734.

As noted above, trust in government and trust in AI can play out independently, and each of the trust elements can play out individually. But the interaction between these trust elements and these targets of trust (government and AI) presents a challenge. For example, do citizens trust that government has the integrity to build an AI with sufficient ability to achieve its objectives? Through these intersections trust becomes real.

Consider the case of the Dubai Electricity and Water Authority (DEWA) that implemented a chatbot, Rammas, to respond to queries from residents. The chatbot can respond to queries both in English and Arabic, and through its use, Dubai could capture and analyse resident requests more holistically. A year after implementation, the chatbot had responded to over 700,000 requests resulting in an 80 percent drop in physical visits.[14] Another example of a system that demonstrates benevolence is the tax filing tool provided by the Australian Tax Office. Leveraging behavioural insights, the tool both enables taxpayers to claim work related expenses, and provides insights when an expense claim might be greater than other taxpayers who have similar profiles, allowing individuals to ensure the accuracy of expenses claimed.[15]

As noted above, trust in government in general has seen as steady decline over the last few years, including in Australia. According to the 2022 Edelman Trust Barometer,[16] only 52 percent of Australians trust government to do the right thing (down 9 points from the previous year), which parallels their lack of trust in business (58 percent trust, down 5 points) and media (43 percent trust, down 8 points). Interestingly, 55 percent of Australians say that their default tendency is to distrust something until they see evidence that it is trustworthy. Factors attributable to declining trust in Australia mirror trends around the world, and include decreasing interpersonal trust, perceptions of corruption, and deeply seated economic worries stemming from COVID-era policies.

However, few of these distrust factors appear to directly involve the design or use of information systems, including AI systems. Moreover, in the private sector, most individuals use AI systems in their everyday life, including auto-completing emails, intelligent search for information, recommender systems on digital platforms, and even chatbots for service triaging. According to a recent study by IBM on AI adoption,[17] the global AI adoption rate is now at 35 percent, a four-point increase from the previous year, and 44 percent of companies are currently working to embed AI into current applications and processes. China, the most aggressive implementer of AI, reports that 84 percent of companies in their country are actively implementing AI.

There are many examples of successful AI implementation in the private sector. Zzapp Malaria is using AI to identify potential sources of malaria-carrying mosquito breeding grounds to reduce the spread of malaria. Vistra, a power producer in the U.S., had been using human workers to monitor hundreds of power indicators (e.g., pressure, oxygen level, pump speeds etc) to optimize operations. Using an AI-powered tool (a heat-rate optimizer), they improved efficiency and generated millions in savings in addition to lower greenhouse gas emissions.[18] Wayfair, an e-commerce company, invested early in AI, and when COVID hit they changed their shipping logistics and generated a 7.5 percent reduction in inbound logistics costs.

Individuals use AI without thinking in their daily life from opening their phone with face ID, to enjoying AI-curated recommendations using social media, to using digital voice assistants (e.g.,

14. https://www.brookings.edu/blog/techtank/2018/09/28/dubai-offers-lessons-for-using-artificial-intelligence-in-local-government/.
15. https://www.afr.com/life-and-luxury/health-and-wellness/how-to-use-nudge-theory-to-get-fit-and-save-more-money-20230210-p5cjmu.
16. https://www.edelman.com.au/trust-barometer-2022-australia.
17. https://www.ibm.com/downloads/cas/GVAGA3JP.
18. https://hbr.org/2022/02/what-makes-a-company-successful-at-using-ai.

Siri, Alexa, etc), to smart home devices.[19] Clearly the public has embraced AI adoption. However, despite this trend and seeming lack of trust concerns around commercial use, there is widespread belief that adoption of AI in the public sector presents more challenges.

As with unpacking the three categories of AI, two main categories of service offered by government reveal a deeper look at trust.[20]

- **Specific services:** These services are explicitly requested by the citizen and focus on direct interaction, with an impact on a limited number of citizens. Examples include receiving a traffic ticket or requesting a camping pass at a government park.
- **General services:** These services are provided by government without a specific request and concern all or most citizens. Examples include the sales tax or broad-scale health programs.

One study has shown that the willingness to trust government use of AI is impacted by both the type of AI and the type of services, as seen in Table 2.[21]

Table 2: Willingness to Accept Government Ai Usage

Types of AI	Specific Service	General Service
Fully Autonomous AI	Highly unwilling	Somewhat willing
Semiautonomous AI	Somewhat willing	Highly willing
Augmented AI	Highly willing	Highly willing

As shown, citizens accept all forms of AI for general services by the government, but are only willing to do so for specific services when augmented AI. Undergirding each of these cells are the three trust elements that loom large (but differently) across the types of AI and the types of service.

Governments across the world recognize that further discussion and action are necessary to address the trust issues faced by public sector AI implementation. Prior to the 2023 G7 meeting, Taro Kano, minister of Digital Affairs in Japan, said: *"All governments need to consider how we can keep the trust of the people towards democracy. . . . all democratic governments now feel an urgency in dealing with AI, so that's why, at the G7, it is on everyone's mind."*[22]

19. https://www.forbes.com/sites/bernardmarr/2019/12/16/the-10-best-examples-of-how-ai-is-already-used-in-our-everyday-life/?sh=6eab30851171.
20. Gesk, T. S., and M. Leyer. (2022). Artificial intelligence in public services: When and why citizens accept its usage. *Government Information Quarterly*, *39*(3), 101704.
21. Gesk, T. S., and M. Leyer. (2022). Artificial intelligence in public services: When and why citizens accept its usage. *Government Information Quarterly*, *39*(3), 101704.
22. https://www.ft.com/content/6a6b91ca-62d0-43ac-a1c5-717ee218a2e6.

Findings from Workshop

The Australian Government has recognised the need to take a leadership role regarding responsible innovation with AI to enhance public value.[23] Australia is a member of the Global Partnership on Artificial Intelligence (GPAI) along with 28 other nations and the European Union, whose goal is to "foster responsible development of AI grounded in these principles of human rights, inclusion, diversity, innovation, and economic growth."[24] In 2019, a set of voluntary AI ethical framework that consists of eight principles were published by the Australian government:[25]

- **Human, societal and environmental well-being:** AI systems should benefit individuals, society, and the environment.
- **Human-centred values:** AI systems should respect human rights, diversity, and the autonomy of individuals.
- **Fairness:** AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities, or groups.
- **Privacy protection and security:** AI systems should respect and uphold privacy rights and data protection and ensure the security of data.
- **Reliability and safety:** AI systems should reliably operate in accordance with their intended purpose.
- **Transparency and explainability:** There should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI and can find out when an AI system is engaging with them.
- **Contestability:** When an AI system significantly impacts a person, community, group, or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system.
- **Accountability:** People responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.

These principles are aspirational—few countries have expended the massive effort to fully operationalize and implement these valuable concepts. To help answer questions around trusted AI use by government and support implementation of these principles, the IBM Center hosted a workshop with senior executives of the Australian public sector in Canberra on May 11, 2023. This meeting provided a first-hand perspective from Australian government officials on the status of AI, issues associated with AI, and the roadblocks and accelerators to implementing. Attendees at the workshop touched on five major themes.

Theme 1—Government is in the business of providing services, and AI is simply a tool to facilitate that.

Attendees stressed that, despite the opportunities and challenges that underscore AI, government's first responsibility is to serve citizens. While AI can be helpful in enabling some of these services, it is merely a tool to facilitate government provision. As such, the goal of government is not to be "techno-dazzled" but instead focus on making AI investments to provide essential services. The business case for use of AI needs to be clear. Moreover, how the use of AI contributes to increased public value should be specified. Public value can arise from doing things more efficiently (i.e., lowering costs) and effectively (i.e., 24/7 availability of services).

23. https://www.industry.gov.au/science-technology-and-innovation/technology/artificial-intelligence.
24. https://www.gpai.ai/about/.
25. https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles.

Theme 2—Government is held to a higher standard of performance regarding AI versus private companies, making explainability and transparency of utmost importance.

One attendee noted that government was held to a higher standard of performance as compared to technology companies. As such, government needs to start with a clear understanding of the performance of AI systems. Humans must review the model to judge its effectiveness. Thus, humans need to be in the loop when AI executes, until the quality of the model is known and (virtually) flawless. Accordingly, the "black box" era of AI needs to end, as government employees and citizens must be able to understand what the model does (explainability and transparency). This feedback loop will go a long way to solving the explainability and transparency issue that can dramatically increase trust by Australian citizens.

Theme 3—Government needs to work holistically in terms of defining AI standard practices, operating models, etc.

AI technology is emerging in all departments and functions within government. Leadership at the highest levels within government can ensure that AI is developed based on a common and ethical approach for standard practices and operating protocols. Doing so will enhance trust by citizens and will also support a consistent feedback loop to analyse AI-centric decisions. Interagency collaboration and taskforces should take a broad look at AI applications and their affordances across the public sector.

Theme 4—Adequate governance is necessary not only for AI technology, but also for the people who build AI systems and the processes used to build them.

While citizens commonly think of AI as a technology, humans build AI with a set of processes and those humans and processes need to be governed just as much as the application of the technology. Solely focusing on the technology ignores the fact that humans inherently contain biases that could unintentionally slip into the AI. By having adequate governance over all parts of the AI (people, processes and technology), there is less chance that unintended elements will arise. Further, effective governance will support citizens having higher trust of both the AI as well as the government.

Theme 5—There is a need to distinguish between different types of AI (fully autonomous, semiautonomous, and augmented) in establishing guidelines and approaches.

The vast differences in types of AI make it important that decisions occur in a manner consistent with each type's needs. For example, citizens will be much more wary of fully autonomous technology versus augmented technology. Government needs to consider the optimal type of AI for each situation. This decision making should focus on both the costs and benefits of each type of AI, and the type of services rendered by the AI. While trust in AI is still in its early stages, the default should be towards more human oversight of AI decisions (e.g., augmented).

Recommendations for Building Trusted AI in the Public Sector

Based on the discussions from the meeting with senior government executives and independent research on AI deployments in the public sector, a list of recommendations to support building trusted AI in the public sector follow.

Recommendation 1—Promote AI-human collaboration

As noted earlier, AI systems can be deployed in several modes including fully autonomous, semiautonomous, and augmented. It is vital for agencies to design work processes adequately to take full advantage of AI affordances, while ensuring that humans are in the loop *as necessary* so that processes and outcomes are executed in a responsible manner.

For general services (e.g., emailing/mailing out property valuation notices), AI systems can be deployed in a fully autonomous manner (see Table 2 in previous section). These sorts of general services rarely require a human to oversee the steps taken; assuming the AI has correctly been built and tested, governments can do this in the most efficient and expeditious manner. With that said, for higher impact projects, there may be a reason to use semiautonomous or augmented approaches instead.

For specific services (e.g., requesting a code waiver from the building department for a specific building), AI can still provide high value in terms of identifying potential issues with the requested variance (e.g., additional risk of fire). However, citizens will generally insist that a human make the final decision (e.g., augmented AI)—they should be supportive of AI that augments but does not replace the human decision.

In general, the presumption for human interaction should be to err on the side of more involvement. While the economics and decision making may improve with less human involvement, government is far less likely to run into substantial citizen resistance if humans remain integral to the process.

Recommendation 2—Focus on justifiability

For most projects, including implementing a traditional information system, a business case is done to show the costs, benefits, and risks of undertaking the project. Sufficient rationale needs to be given for executives to make a supportable decision.

While a business case still needs to be done for an AI system, AI systems also need to be justified to the public. As such, justifiability can be thought of as an outward facing business case. Justifiability needs to be based around public value, which refers to the greater societal benefits that occur. Additionally, this focus on justifiability helps build legitimacy for the use of AI systems. Legitimacy is not simply what is legally possible but is also about earning trust and the social licence to do more with AI to enhance public value.

The Government of Canada has created an algorithmic impact assessment (AIA) tool for public agencies to better understand and mitigate risks associated with AI systems.[26] The tool calculates impact across six domains: *project*, *system*, *algorithm*, *decision*, *impact*, and *data*. For example, within *algorithm* the disclosure and explainability are assessed, and within *impact* issues associated with the decision such as reversibility and duration of decision outcomes are considered. Two mitigation areas are also assessed—*consultations* (to learn more please refer to the section on contestability) and the *de-risking and mitigation measures* in place such as privacy safeguards.

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

Recommendation 3—Insist on explainability

Explainability is critical to articulate how an AI system arrived at a particular outcome. While explainability is easier to do with simple AI models, things get messy quickly once AI modules are deployed at scale. But even with complex systems, one should be able to specify data used to train the algorithm, why the algorithm being deployed is fit for purpose, and how the system is maintained. Absent these elements, which are likely to increase the likelihood of citizen acceptance the system should be reconsidered or cancelled.

Explainability is not a one-time event, particularly if the AI system takes advantage of machine learning. With machine learning, the AI continues to learn and adapt as new data is encountered. For example, an AI that predicts structure fires constantly receives new data as more structure fires are encountered. This may lead the AI to change algorithms based on the new data. While this improves the quality of the AI, care must be taken to ensure that explainability is then reexamined.

A feedback loop is an important part of explainability. Government should be able to explain the data and the algorithms to the public, and the public should have a feedback loop to challenge erroneous or problematic findings.

Recommendation 4—Build in contestability

Agencies need to have governance protocols and administrative processes that allow for contestability of AI systems. Governance protocols should be clear on how citizens can engage with AI systems not only after they are deployed, but also in their design and development. Contestability allows humans to intervene and interrogate critical elements (e.g., datasets, learning algorithms, use cases, etc.) of the AI systems from conception and through ongoing deployment.

> Just as human-centred design has benefited systems development by placing people in central focus, embracing an open-source community mindset can help increase trust levels.
>
> —*Workshop participant*

During the early stages of conceptualising AI systems, public agencies should solicit stakeholder input where possible. For example, property development projects must publicize what is being developed, the impact to the community, etc., and have a timeframe for residents to provide input. Similarly, when AI systems are conceptualised, especially when they have direct impact on citizens as they experience and interact with public services, there should be an opportunity for consultation. This not only allows the agency to engage the public to seek input but is also a critical trust building mechanism.

When AI systems are being designed, public should be able to contest what datasets are being used, seek evidence that datasets are fit for purpose, and even inspect the performance data on algorithms. This early feedback loop will surface problems that developers may be unaware of and will build trust that the developers are working with the public instead of at cross purposes. Finally, when systems are deployed, citizens should have the right to question algorithmic decisions, and the necessary administrative processes for recourse need to be readily accessible.

In 2020, the cities of Amsterdam[27] and Helsinki[28] created AI registries that lists algorithms used to deliver public services. Specifically:

> Each algorithm cited in the registry lists datasets used to train a model, a description of how an algorithm is used, how humans utilize the prediction, and how algorithms were assessed for potential bias or risks. The registry also provides citizens a way to give feedback on algorithms their local government uses and the name, city department, and contact information for the person responsible for the responsible deployment of a particular algorithm. A complete algorithmic registry can empower citizens and give them a way to evaluate, examine, or question governments' applications of AI.[29]

Recommendation 5—Build in safety

A critical mechanism to build trust in AI systems is ensuring that the public understands safety concerns with their use. While ensuring that AI systems have explainability and contestability is highly desirable, these goals are not always fully possible due to the way AI systems ingest and learn from large datasets and the way the AI is deployed.

Safety science is a well-established discipline and one that can offer principles and practices to improve the safety of AI systems, and to provide insights on communicating safety performance metrics to relevant stakeholders. The most important safety principle involves having an incident tracking database where government staff and citizens can log incidents.[30] The incident tracking database should have an automatic notification process to system administrators and senior government officials, as well as feedback to the person reporting the incident.

The incident tracking database should be fully public facing with regular reporting. At minimum, the database should track the system with the problem, specific description of the problem, ways to know why the AI is incorrect, when the system was last audited, and when the system was last updated.

Recommendation 6—Ensure stability

Stability addresses how algorithms can provide consistent responses over time and across cases. Put differently, a stable algorithm is fair and unbiased and can meet the demands of varying cases and interaction modes. Interaction modes are an interesting element to examine. For example, chatbots must interact with a wide assortment of citizens, those that have digital expertise and those that do not—stability calls for ensuring consistency in how both sets of users have a similar experience.

The other aspect of stability ensures that algorithms perform as expected within the expected bounds of conditions. When conditions change, the AI should change as well. When boundary conditions change, the performance of the AI should not result in drastic failure. A stable AI system should track graceful degradation of performance, which if monitored adequately, should trigger human operators to conduct appropriate audit and maintenance procedures.

27. https://algoritmeregister.amsterdam.nl/.
28. https://ai.hel.fi/en/ai-register/.
29. https://venturebeat.com/ai/amsterdam-and-helsinki-launch-algorithm-registries-to-bring-transparency-to-public-deployments-of-ai/.
30. https://incidentdatabase.ai.

CONCLUSION

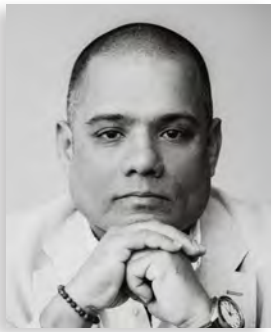
> Machine learning is already bringing a step change in the ability to sift through data and test solutions. . . . Machine learning can also be used to simulate real-world scenarios and model outcomes, for example to test therapies. It sped the development of a coronavirus vaccine. . . . You can just imagine what a step change it will be when we can build these kinds of twinned systems for the human body. It's not difficult to see the potential. It's also obvious that these digital tools are not always a force for good. We've all seen the negatives that connectivity has brought us. You don't need me to describe the harm done by algorithms that actively reinforce our perspectives rather than exposing people to diverse viewpoints. . . . But the simple reality is there's no turning back. More than that, we don't want to turn back. This fourth digital revolution is here. The trick is not to hide. But to get the safeguards right and get that balance right. This is where we need a concerted national and international effort. As we prepare for the ever-increasing use of machine learning, and of artificial intelligence and robotics, the answers lie in preparation and regulation, understanding the pitfalls and the potential, and shaping the technologies for good.
>
> *—Remarks by Australia's Chief Scientist Dr Cathy Foley during a keynote speech in 2022*[31]

AI systems will continue to shape public sector operations and impact democracies. Consider the current activity around misinformation on various social media platforms. Much of this information is facilitated and propagated with AI technologies leading to undesirable outcomes, including loss of trust in government. Governments must play an important role in combatting such actions to continue to retain its legitimacy and maintain social cohesion. Doing so will require governments to take a more active role in the use of AI. Recently, there has also been significant interest in generative AI, which is not new—but the debut of ChatGPT has many now considering how to create responsible innovation frameworks. Governments cannot remain behind when it comes to addressing AI responsibly (or doing the same for any other emerging technologies, such as quantum computing). A successful approach will require a whole-of-government concerted effort, and collaboration with industry and academia.

The emergence of AI in the world, and specifically in the public sector, makes this an exciting era. Given the frantic pace of AI development, government has a responsibility to be more proactive around the design, development, and deployment of AI systems to advance national goals. By adopting the recommendations presented in this report, the Australian government can encourage the growth of AI and realize its benefits while ensuring that adequate guardrails are in place to protect the citizens of Australia.

31. https://www.chiefscientist.gov.au/news-and-media/science-thats-shaping-our-future.

ABOUT THE AUTHORS


KEVIN C. DESOUZA

Kevin C. Desouza is a professor of Business, Technology and Strategy in the School of Management at the QUT Business School at the Queensland University of Technology. He is a nonresident senior fellow in the Governance Studies Program at the Brookings Institution. He formerly held tenured faculty posts at Arizona State University, Virginia Tech, and the University of Washington and has held visiting appointments at the London School of Economics and Political Science, Università Bocconi, Shanghai Jiao Tong University, the University of the Witwatersrand, and the University of Ljubljana.

Desouza has authored, coauthored, and/or edited nine books. He has published more than 150 articles in journals across a range of disciplines including information systems (*Journal of MIS*), information science (*Journal of the American Society for Information Science and Technology*), public administration (*Public Administration Review*), political science (*Studies in Conflict and Terrorism*), technology management (*Technology Forecasting Social Change*), and urban affairs (*Cities*). Several outlets have featured his work including *Sloan Management Review, Financial Times, Stanford Social Innovation Research, Harvard Business Review, Forbes, Businessweek, Wired, Governing, Slate.com, Wall Street Journal, BBC, USA Today, NPR, PBS*, and *Computerworld.*

Desouza has advised, briefed, and/or consulted for major international corporations, nongovernmental organizations, and public agencies on strategic management issues ranging from management of information systems to knowledge management, innovation programs, crisis management, and leadership development. Desouza has received over $2.25 USD million in research funding from both private and government organizations.


DR. GREGORY S. DAWSON

Dr. Gregory S. Dawson is clinical professor in the School of Accountancy in the W. P. Carey School of Business at Arizona State University. Dr. Dawson was awarded his PhD in Information Systems from the Terry College of Business at the University of Georgia.

Prior to becoming an academic, Dr. Dawson was a partner in the Government Consulting Practice at PricewaterhouseCoopers, joining PwC (formerly Coopers & Lybrand) in the Washington, D.C., office and later relocating to Sacramento, California. Dr. Dawson was a leader in the field of public sector outsourcing as well as information systems implementation. He has worked extensively with the federal government (including Central Intelligence Agency, Department of Defense (Army, Navy, Air Force and Marines), the Federal Deposit Insurance Corporation (FDIC), and the Bureau of the Census, among others, and with a variety of state governments (including Virginia, North Carolina, Pennsylvania, New York, and California). After leaving PwC, Dr. Dawson was a director at Gartner, working in the state and local government practice.

Dr. Dawson is also the former president of the Association for Information Systems Special Interest Group on IS Leadership and co-leads a track on IS leadership at a major IS conference. His research is primarily focused on information systems leadership and innovation in the public sector. His work has been published in a variety of top academic and practitioner journals. His research has been published in *Journal of the Association for Information Systems, Decision Support Systems, Organization Science, Journal of Management Information Systems, ACM Transactions on Management Information Systems, Communications of the Association for Information Systems, InformationWeek* and numerous Brookings Institution reports.

In addition to his role as a professor, Dr. Dawson also actively consults with governments around the world on their use of technology.

KEY CONTACT INFORMATION

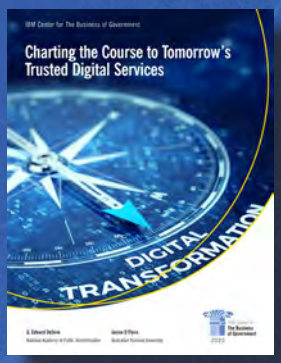
Kevin C. Desouza
Professor of Business, Technology, and Strategy
QUT Business School
Faculty of Business and Law
Queensland University of Technology
Brisbane, Australia

Email: kevin.c.desouza@gmail.com

Dr. Gregory S. Dawson
Clinical Associate Professor
Center for Organization Research and Design
Arizona State University
300 E Lemon St.
Tempe, AZ 85287

Email: GregorySDawson@gmail.com

RECENT REPORTS FROM THE IBM CENTER FOR THE BUSINESS OF GOVERNMENT

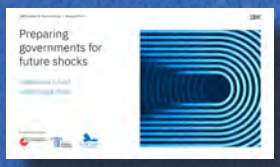


Charting the Course to Tomorrow's Trusted Digital Services

by G. Edward DeSeve and Janine O'Flynn

Preparing governments for future shocks: An action plan to build cyber resilience in a world of uncertainty

by Tony Scott

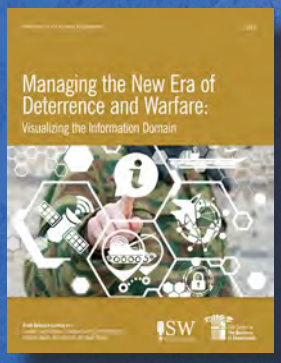


Preparing governments for future shocks: Collaborating to build resilient supply chains

by Robert Handfield Ph.D.

Government Procurement and Acquisition: Opportunities and Challenges Presented by Artificial Intelligence and Machine Learning

by Mohammad Ahmadi and Justin B. Bullock



Managing the New Era of Deterrence and Warfare: Visualizing the Information Domain

by Brian Babcock-Lumish

Leveraging Data for Racial Equity in Workforce Opportunity

by Temilola Afolabi

A Guide to Adaptive Government: Preparing for Disruption

by Nicholas D. Evans

Mobilizing Cloud Computing for Public Service

by Amanda Starling Gould

For a full listing of our reports, visit www.businessofgovernment.org/reports

About the IBM Center for The Business of Government

Through research stipends and events, the IBM Center for The Business of Government stimulates research and facilitates discussion of new approaches to improving the effectiveness of government at the federal, state, local, and international levels.

About IBM Consulting

With consultants and professional staff in more than 160 countries globally, IBM Consulting is the world's largest consulting services organization. IBM Consulting provides clients with business process and industry expertise, a deep understanding of technology solutions that address specific industry issues, and the ability to design, build, and run those solutions in a way that delivers bottom-line value. To learn more visit ibm.com.

For more information:
Daniel J. Chenok
Executive Director
IBM Center for The Business of Government

600 14th Street NW
Second Floor
Washington, D.C. 20005
(202) 551-9342

website: www.businessofgovernment.org
e-mail: businessofgovernment@us.ibm.com

Stay connected with the IBM Center on:

or, send us your name and e-mail to receive our newsletters.

25TH ANNIVERSARY

IBM Center for
The Business of Government