Towards a Considered Use of AI Technologies in Government

September 28, 2023

Institute on Governance
LEADING EXPERTISE

Institut sur la gouvernance
EXPERTISE DE POINTE

Table of Contents

Executive Summary

The purpose of this report is to provide policymakers and practitioners in government with an overview of controversial deployments of Artificial Intelligence (AI) technologies in the public sector, and to highlight some of the approaches being taken to govern the responsible use of these technologies in government. While the scope of this report does not include specific recommendations, it is our hope that it will spark further consideration and discussion of some key questions related to the responsible use of AI by public sector organizations.

This project was administered by the Institute on Governance and conducted by Think Digital. This report has been written with a Canadian lens, however we have suggested that the insights are generally applicable to public sector institutions in other jurisdictions as well.

Our premise behind this report is simple: the AI genie is out of the bottle, and it is not going back in. As a result, governments at all levels are increasingly going to have to grapple with how to responsibly use these powerful, compelling, and accessible technologies. Avoidance is not a realistic strategy, therefore understanding what other jurisdictions are doing, what has worked, and what has not, is valuable for government officials as they consider the risks involved in using (or not using) AI in the public sector.

For this reason, the Institute on Governance and Think Digital undertook a case study-based research project, where 24 examples of AI technology projects and governance frameworks across a dozen jurisdictions were scanned. Two environmental scans make up the majority of the report's content. The first scan presents relevant use cases of public sector applications of AI technologies and automation, with special attention given to controversial projects and program/policy failures. These cases are divided into four categories of use cases:

- Automated Decision-Making
- Automated Decision-Support
- Detection, Alerts, and Notifications
- Procedural Automation/Process Improvement

The second scan surveys existing governance frameworks employed by international organizations and governments around the world. Each scan is then analyzed to determine common themes across use cases and governance frameworks respectively. The final section of the report provides risk considerations related to the use of AI by public sector institutions across use cases.

These environmental scans and our subsequent analysis found that because of the high level of visibility and a poor level of understanding of AI technologies, many public sector implementations to date have had significant challenges. This is particularly the case when they have been introduced into very sensitive contexts that could impact vulnerable populations. Some questions that these case studies raise include how to use AI technologies in a manner that is consistent with legal norms, that allow for tracking both positive and negative impact, and that enable public participation in oversight.

Scan 1: AI Failures and Controversies in the Public Sector

The first part of the report sheds light on the use of AI technologies in the public sector and provides eleven illustrative case studies, with special attention paid to public controversies and program/policy failures. Our analysis of the case studies found the following common themes associated with AI project failures in the public sector:

1. **Technical errors**: basic technical errors were common in the cases examined, leading to issues such as inaccurate predictions, miscalculations, and biased or discriminatory outcomes.
2. **Regulatory non-compliance**: implementing organizations often failed to adhere to established regulations, guidelines, and laws governing data and automation practices.
3. **Governance vacuums**: the notable absence of explicit AI governance exacerbated issues related to system abuse, mission creep, and lack of transparency, accountability, and explainability.
4. **Opaque systems**: lack of visibility into how systems function made it difficult for operators and leadership to evaluate system performance and to identify failures. This sometimes contributed to the erosion of human agency in AI supported decision-making processes.
5. **Modifying policy to accommodate technology**: policy and program eligibility criteria were sometimes modified to accommodate labor-saving automation, which in some cases led to negative impacts on affected individuals.
6. **Sensitive deployment contexts:** automated systems deployed in sensitive contexts invited scrutiny, increasing the potential for public failures and controversies regardless of their efficacy.

Scan 2: AI Governance Practices Around the World

Our second environmental scan provides a review of eight approaches to governing AI in the public sector. These approaches include, but are not limited to, the use of informal governance approaches such as peer-review committees and communities of practice. In our analysis, we observed the following notable patterns and governance considerations across the frameworks examined:

1. **Use of checklists**: checklists are a common tool used to help operationalize governance principles and to ensure adherence to overarching ethical principles in AI development and deployment.
2. **Ethical and human rights principles**: core principles such as fairness, transparency, accountability, and respect for human autonomy, feature prominently across all frameworks, and provide a foundation for responsible AI in the public sector.
3. **Risk assessment and mitigation**: many frameworks emphasize the importance of early and continuous risk management to minimize the potentially negative impacts of AI systems on individuals, communities, and the environment.
4. **Transparency and explainability**: responsible system implementation ensures that all stakeholders, those involved in and impacted by its function, have a minimal understanding of its processes and outcomes.

5. **Human agency, oversight, and accountability**: humans are ultimately responsible for AI system outcomes, and often there is a focus on ensuring that human actors are empowered to override or reverse AI decisions when necessary.
6. **Ongoing monitoring and evaluation:** AI governance frameworks and AI system operations should be reviewed and updated periodically to accommodate rapid changes in technology, mitigate unintended biases, and to identify areas for improvement.
7. **Stakeholder engagement**: intentional and iterative engagement with diverse subject matter experts, relevant communities and individuals affected by AI systems ensures that deployment aligns with their values, needs, and expectations.
8. **Communities of practice and peer review**: encouraging collaboration, knowledge sharing, and best practice development between government agencies, experts, and stakeholders is seen as a best practice by many.

Risk Considerations

Building on the insights gleaned from our environmental scan analyses, we have produced a series of risk considerations to help those in public sector organizations – particularly in the Canadian context – deliberate on the risks associated with AI and automation projects.

We present a process-based approach to risk assessment for AI technologies, organizing the case studies and risk framework analysis by the same four types of potential process use cases used to organize the first environmental scan in this report:

- Automated Decision-Making
- Automated Decision-Support
- Detection, Alerts, and Notifications
- Procedural Automation/Process Improvement

While there is robust policy guidance for Government of Canada departments when it comes to the use of AI and automated systems for use cases where decisions are being made about a specific client, this is not necessarily the case when trying to prevent negative outcomes in other scenarios. For the purposes of this report, we define negative outcomes as those failing to support public well-being, reduce harm, ensure governmental efficiency, and maintain the public's confidence. We also identify several types of potential organizational risk such as strategic, reputational, compliance, legal, operational, security, and financial risks.

In the risk considerations section of this report we propose that four factors magnify risk specifically in the context of AI technology, namely: boundability, reversibility, explainability, and visibility. We suggest a conceptual risk approach to be applied in cases where existing guidance does not exist, as follows:

(Boundability Risk + Reversibility Risk + Explainability Risk) x Visibility Risk

These factors are meant to be a conceptual starting point and may be adjusted in specific cases by considering other factors as relevant.

We also highlight the risk of avoiding AI technology completely, such as negative reputational risks and "shadow IT," where employees access AI tools outside of work systems without the appropriate safeguards or knowledge. In general, given the increasing impact and prevalence of AI technologies, we recommend that responsible experimentation be actively supported by public sector organizations.

Ultimately, the risk considerations that we have proposed are designed to provide a way to think about managing the risks associated with the implementation of AI technologies within public sector organizations while also acknowledging the potential risks of AI technology avoidance.

Limitations and Considerations

In the first scan, the distinction between categories was made to clarify the differences between the case studies with the caveat that, in some instances, there is notable overlap blurring the lines we have attempted to draw. It is important to acknowledge that this kind of categorical approach to AI technologies can oversimplify or obscure the real-world complexity of application and impact in a public sector and policy context. There are alternative approaches that have been taken to categorize AI technologies and their use, including in the TBS Directive on Automated Decision-Making which is explored later in this report.

That said, we chose to organize our case studies by application and process type in the way that we have in order to ground a risk management perspective and approach that considers the nature of the processes in which AI technologies are embedded, and the likelihood of negative outcomes within particular processes independent of the technology itself. Moreover, activities in each of the above categories may be connected and carried out by some combination of people and AI technologies. Rather than avoid categorization altogether, our process-based approach to risk sometimes requires tracking potential sources of risk beyond the direct application of an AI technology, across the categories suggested, and into the larger network of the processes, contexts, and actors around which they are situated.

The intention behind our categorization was therefore not to draw predetermined, hard lines of causation between certain AI technologies/actors, processes, and risk. Rather our intent was to provide a relatively consistent heuristic that could be applied by policy makers and implementation teams to help them think about how context and process design around AI technologies relates to identifying and mitigating potential sources of risk. There are an increasing number of frameworks and approaches that are being developed by a variety of actors to address some of the risks that we've identified. We encourage consulting additional sources, including papers such as "A Trust Framework for Government Use of Artificial Intelligence and Automated Decision Making".

We also note that the focus of the environmental scans and use cases examined in this report have largely focused on the use of AI for service-oriented examples, be them public-facing or for internal process efficiency. However, as AI technologies are increasingly incorporated into the day-to-day activities of government employees at an individual level (e.g. through enterprise-focused AI tools such as Microsoft's Copilot or Google's Duet) considerations around the workplace impacts of AI will become of growing importance as we shift our mental models around AI becoming a "colleague" rather than just another technology tool. These impacts have been largely left out of the scope of this paper but warrant careful consideration by policymakers.

While making specific recommendations was beyond the scope of this report, our intent behind the report was to contribute to what is a rapidly growing body of knowledge on AI in the public sector. The contents of the report are based on our gathering and analysis of original source material. Due to the limits of our research capacity, it is important to note that for both environmental scans, original source material was generally limited to whatever data was publicly available at the time of writing, and was, for the most part, either published by the relevant implementing organization or by third party critics/evaluators of the system or program at hand. A very limited number of interviews were conducted to clarify some contextual details with specific government entities and initiatives mentioned in this report. In addition, because we sought to identify and analyze common traits behind public policy and program failures as they relate to public sector use of AI, this report does not offer a complete representation of all current or possible public sector applications of AI or related automation technologies. Readers of this report are encouraged to consult the source material where available for additional details and the specifics of each case study in both environmental scans.

Environmental Scan #1: Use Cases of AI and Automation in Government

The environmental scan below provides case studies that detail the use of Artificial Intelligence (AI) and automated systems by the public sector in Canada and other jurisdictions, with special attention given to public controversies and program/policy failures. The case studies have been categorized according to four general categories of AI usage in government organizations:

1. **Automated Decision-Making (ADM):** Includes technologies that replace the judgement of humans to automate the decision-making parts of a decision-making process, usually with no (or very limited) human oversight or options for intervention. These systems often feature supervised learning techniques that allow algorithms to classify people or objects into two or more pre-defined categories, or more generally apply labels of some kind.

2. **Automated Decision-Support:** Systems that automate part of the decision-making process to make recommendations or generate outcomes that support human decision-making further downstream. For example, systems that make recommendations via recommender engines, or that offer predictions using predictive analytics, etc.

3. **Detection, Alerts, and Notifications:** A broader category of use cases, ranging from the use of algorithms for the detection of simple, predetermined conditions or anomalies to machine learning used for fraud detection or facial recognition.

4. **Procedural Automation/Process Improvement:** Procedural automation usually involves designing or redesigning manual procedures so that some parts of the process can be automated using digital technologies. In some cases, this automation may involve relatively simple algorithms, while in others it may require more sophisticated AI techniques. Until relatively recently, process improvement or process automation typically involved the automation of mundane, repetitive, and simple tasks that humans would otherwise be responsible for. Increasingly, however, the processes that can be automated are becoming more cognitively sophisticated. Examples of procedural automation include Robotic Process Automation (RPA), chatbots and generative AI more broadly, data digitization and migration, and cross-referencing systems.

It should be noted that the category definitions above do not necessarily align with those used by others, including the Treasury Board of Canada Secretariat (TBS) Directive on Automated Decision-Making. However, we have made the distinction between these categories to help clarify the differences between the case studies in this scan with full recognition that in some instances the boundaries between categories can be fuzzy.

SUMMARY TABLE: Use Cases of AI and Automation in Government

Case	Country	Organization	Short Description	System Classification
Robodebt - Automated Debt Estimation and Recovery	Australia	Department of Human Services (DHS)	To combat welfare fraud, the DHS automated their debt estimation and recovery processes with little human oversight or intervention.	Automated Decision-Making
At-home care distribution	United States of America	Arkansas Department of Health & Idaho Department of Health and Welfare	Arkansas and Idaho adopted algorithmic assessment tools that automatically determined much care the State would distribute to people living at home with disabilities.	Automated Decision-Making
Algorithmically Predicting Student Grades	Ireland and The UK	Department of Education and Skills	Following premature closure of Irish schools due to the Covid-19 pandemic, the Department of Education and Skills developed an algorithm to predict final examination grades for graduating students.	Automated Decision-Making

Case	Country	Organization	Short Description	System Classification
Automated application triage	Canada	Immigration, Refugees and Citizenship Canada (IRCC)	IRCC has used advanced data analytics to help triage routine applications for visa eligibility.	Automated decision-support
Automating unemployment categorization	Poland	Ministry of Labor and Social Policy (MLSP)	Automated assessment scoring used to categorize the unemployed into tiered levels of available programming and assistance resources.	Automated decision-support
Big Data Fraud Detection, SyRI	The Netherlands	Ministry of Social Affairs and Employment	Technical infrastructure and algorithm that allowed government agencies to data-match proposed risk indicators across databases to detect welfare fraud.	Automated decision-support
Pre-load Air Cargo Screening	Canada	Transport Canada (TC)	Automating risk-based assessment process by scanning pre-load air cargo information to identify potential physical security threats.	Detection, notification, alerts

Case	Country	Organization	Short Description	System Classification
Facial Recognition Technology at Pearson Airport	Canada	Canada Border Services Agency (CBSA)	Facial-matching technology was used to identify individuals from an existing database of people the agency suspected might attempt to enter the country illegally.	Detection, notification, alerts
Robotic Process Automation (RPA) in Social Assistance Onboarding and Administration	Sweden	Trelleborg Municipal Government	The Municipality or Trelleborg adopted RPA technologies to automate mundane, repetitive data entry tasks, and welfare decisions previously carried out by humans.	Procedural Automation / Process Improvement
Robotic Process Automation in a New Zealand Financial Institution	New Zealand	Anonymous Financial Institution	A Financial Institution introduced Robotic Process Automation to automate mundane tasks, and faced some resistance from employees.	Procedural Automation / Process Improvement
Chatbots: Government of Singapore, Microsoft, and Google	Singapore, United States of America	Singapore Government Technology Agency, Microsoft, Google	A brief recounting of high profile chatbot failures in government and the private sector.	Procedural Automation / Process Improvement

Automated Decision-Making

Robodebt – Australia

BACKGROUND

Australia's Income Compliance Program (ICP), popularly known as Robodebt, was an automated system deployed by the Department of Human Service (DHS) in 2015 to reduce the costs of reclaiming overpayments to welfare recipients. To identify welfare recipients who may have been paid more support than they were eligible for, Robodebt relied on a data-matching system which used external tax data from the Australian Tax Authority (ATO). Once an individual was identified, a secondary automated system used the same ATO tax data to estimate how much the individual had been overpaid. Once the debt was calculated, an automatic debt notice was triggered and sent to the beneficiary. A confluence of failures in service design, change management, and governance, ultimately led to the wrongful raising of A$1.2 billion in debt to 433,000 Australians through a process which Australian courts deemed unlawful.[1]

FUNCTION

Income compliance is the process by which personnel at the DHS identify discrepancies between how much income support a beneficiary received, and how much they were eligible to receive in a given time period based on their income over that period. If it is found that a beneficiary was overpaid, DHS may raise a debt and attempt to reclaim the difference.[2] The DHS stored data about how much income support recipients received, but did not internally store any information about how much income a recipient made over a period of time. To identify overpayments, DHS relied on a process known as "data-matching": the common practice of comparing data from two different agencies to identify matches in personal information, usually with the intent of utilizing the matched external data for new purposes.

As part of their usual taxation operations, the Australian Tax Authorities collect annualized income data on all Australians. DHS was able to access the ATO income data via a data sharing agreement between the two agencies, enabling them to identify overpayments and send debt collection notices.

DISCUSSION

The use of ATO data by the DHS to identify potential overpayments was not the problem. Instead, their data-matching practice was problematic only once it was used to estimate the outstanding debt for recipients. In a practice known as *smoothing,* the DHS

averaged ATO income data over a given period of time and applied evenly across a given number of biweekly periods. For example, if an individual earned $10,000 in March, $14,000 in July, and nothing the rest of the year, smoothing made it appear as though an individual had earned A$2,000 per month. In effect, this smoothing of the data often resulted in the over-estimation of individual earnings in a given fortnight, and gave the impression that there were outstanding debts.

Prior to ICP, the DHS had the authority to demand income statements from businesses on behalf of citizens to verify their income data. Conversely, under ICP, the responsibility of collecting and submitting income information rested on income support recipients. Therefore, a reverse onus of proof was put into place and a task that previously resided with specialists at DHS was suddenly the burden of welfare recipients who, in some cases, needed to retrieve payslips from jobs they no longer held to prove the non-existence of estimated debt.[3]

A 2017 Ombudsman report found that the ICP's method for notifying citizens of their calculated debt was insufficient because the ICP failed to provide appropriate recourse and assistance options. Additionally, they did not explain how the debt was calculated nor did they suggest the possibility of inaccuracy.[4] Some academics suggest that the interface of the online self-service portal used for challenging debts was unintuitive and complicated, especially for those who already struggle with accessing technology.[5]

Before ICP was implemented in 2015, DHS was able to conduct 20,000 reviews a year.[6] DHS used a prediction model to prioritize reviews of individuals who they believed were most likely to have debts, correctly identifying individuals 99.6 percent of the time.[7] Importantly, this system was used only to identify people who had debts, not to estimate debt owed. Once ICP was fully automated in 2016, DHS was able to conduct 20,000 reviews every week.[8] As the number of reviews increased, the likelihood that a randomly selected individual was overpaid decreased. Naturally, this was the case when ICP scaled in 2017 from 20,000 to 90,000 reviews per year. What a DHS manager once referred to as a "boutique, small, slow program," became "mass production" with limited testing, which led to the wrongful raising of thousands of debts.[9] A media storm ensued, bringing to light the experiences of individuals targeted by the program. These stories suggested that although the intention was to cut administrative costs, the true cost was human. In practice, the program resulted in financial hardships, stress-induced mental health disruptions, and suicides.[10]

In the planning phase for ICP, DHS did not consider any of the risks which would eventually lead to its demise in its risk assessment.[11] While there were governance structures in place at ICP's inception that could have helped avoid the program's failure, they do not appear to have been fully adhered to.

Lack of alignment and collective understanding was evident from the project's earliest stages, and not just between agencies but across the teams working on ICP at the DHS,

where organizational silos limited teams from understanding the complete process and restricted their ability to identify and intervene in critical flaws.[12] It's also not clear whether leadership had a full understanding of the end-to-end process. In 2017, as ICP scaled to 90,000 reviews per year, there was no unified and clearly documented business process to properly govern the debt identification and raising process.[13]

IMPLICATIONS

The ICP was officially scrapped in 2020, and all debts raised under the program were repaid or forgiven as part of a $A1.8 billion settlement of a class-action lawsuit against the government. In 2022, a Royal Commission into the Robodebt scheme was initiated and it issued a final report in July 2023.

KEY TAKEAWAYS

- A basic and incorrect assumption about how income is earned over a welfare recipient's lifetime led to the wrongful issuance of hundreds of millions of dollars in debt.
- Relying on this assumption allowed the DHS to systematize debt estimation and issuance.
- Once systematized, the automation of debt collection and issuance scaled up a fundamentally flawed practice with no legal basis.
- Insufficient testing of the ICP and adjacent services meant critical errors persisted far longer than was acceptable.
- The ICP's planning phase overlooked risks, revealing the need for improved governance, alignment, and understanding across teams and agencies.

At-Home Care Distribution – USA (Arkansas, Idaho)

BACKGROUND

Major demographic and generational shifts in the United States of America have increased demand for in-home disability and long-term care, where human and material resources are already limited. The situation has caused problems for those in need of care and the aides themselves, and "as needs increase, states have been prompted to look for new ways to contain costs and distribute what resources they have."[14] Arkansas' Department of Human Services (ADHS) attempted to alleviate some of this pressure in 2016 by implementing an algorithmic assessment tool to decide how much care the State would give to people living at home with disabilities. The main reasons for automating the allocation of disability welfare was to increase the efficiency of service delivery, and to ensure the fairness of resource distribution. Similarly, in 2011, Idaho's government built in-house their own algorithmic assessment tool for allocating home care and community integration funds.

FUNCTION

In 2016, the ADHS transitioned from an "irrational" to a "rational" system for distributing care hours and procured an algorithm from InterRAI, a non-profit coalition of health researchers, which relied on the computerized assessment of beneficiaries' abilities and needs. As part of the new system, nurse practitioners visited beneficiaries on an annual basis to administer a survey of 286 questions, covering things from their mental health to how much help they needed when eating, using the bathroom, or doing their personal finances.[15] Nurses then entered data from the interview into a computer form, and based on the inputs, an algorithm calculated how many hours of care the person would receive for the next year.[16]

The ADHS algorithm would then compute about 60 different descriptions, symptoms, and ailments and sort beneficiaries "through a flowchart-like system" into various levels of need, each of which corresponded to a standard number of care hours.[17] In effect, this meant that "a small number of variables could matter enormously," and marginal differences in a beneficiary's answers could disproportionately impact the hours of care they would be eligible to receive.[18] In the first year of implementation, a group of beneficiaries who had their care hours reduced by an average of 43 percent, brought the new program to court claiming that such a reduction meant their needs were not being met.[19]

The situation in Idaho was practically the same, except the state's algorithm was built in-house and its assessment results determined how much money a beneficiary received in subsidies and not in care hours.[20] In 2011, when a new formula was instituted, "funds suddenly dropped precipitously for many people, by as much as 42 percent."[21] And

when those impacted tried to find out how their benefits were determined, the State declined to disclose the formula it was using, saying that its math qualified officially as a "trade secret."[22]

DISCUSSION

During the Arkansas court case (Jacobs v. Gillespie), important flaws in the ADHS algorithm and how it was implemented were discovered. Firstly, the algorithm's designer, Brant Fries, discovered that a third-party software vendor had not implemented the system properly, and had mistakenly adopted a version of the algorithm that didn't account for diabetes related issues. Secondly, Fries' own calculations failed to code cerebral palsy into the algorithm, causing inaccurate calculations for hundreds of people that, for the most part, lowered their allotted care hours.[23] And lastly, mistakes were made by human assessors administering the survey. In one such case, a person with double amputations was marked as not having mobility issues, because he could get around in a wheelchair.[24]

In court, Fries acknowledged that the system was not designed to calculate hours of care based on people's idiosyncratic needs**.** He clarified instead, that the algorithm was scientifically calibrated to equitably allocate scarce resources. In other words, *equitable* distribution of supply (care hours) was not necessarily the same as, or commensurate with, *fair* distribution of those resources. Whether automated or not, the desire to distribute welfare resources equitably can require a certain level of "smoothing" the data derived from assessing the lived complexity of individual needs. And in a highly discretionary situation like setting benefit limits, the impact of marginality and generalizing for the sake of equitable distribution can have adverse effects.

Although the State was not legally required to provide a detailed account of how they designed or applied their algorithm, both constitutional and federal statutes required the ADHS (and other such entities) to explain "*specific factors*" used by the algorithm to determine results. In a separate case, Arkansas Department of Human Services v. Ledgerwood, the court found that the ADHS had failed to meet its "notice-and-comment" obligation because it did not tell the public its plan to adopt an automated assessment tool. Ultimately, in both cases, the ADHS algorithm was deemed unconstitutional because it was used to generate applied results without legally and procedurally ensuring due process for those involved.

Similarly, in 2016, Idaho's Medicaid program was found to be unconstitutional by the courts in a class action lawsuit brought by the local American Civil Liberties Union (ACLU) branch representing 4,000 beneficiaries with developmental and intellectual disabilities. Before the lawsuit was filed, the State refused to disclose reasons for why certain individuals had their assistance cut by up to 20-30 percent, claiming its decision-making formulas were "trade secrets" and did not qualify as public information.[25] In addition, the State failed to explain why results were disproportionate in some parts of

the state versus others. The ruling, however, determined that the algorithmic assessment tool had deprived beneficiaries of their rights to due process because it was effectively producing arbitrary results for a large number of people. Based on grounds established in the Medicaid Act, the court ruled that decisions leading to the reduction of an individual's assistance had to be explainable, and transparent.[26]

In court, it was also revealed that Idaho had relied on deeply flawed, limited, and inaccurate data when building their algorithm in-house, and that they had failed to regularly audit it despite knowing they needed to.[27] Experts hired by the ACLU found that the State had discarded two-thirds of the historical data used to build their predictive models because of data entry errors and other inaccuracies. This meant the assessment tool was using a limited subset of flawed data to predict what beneficiaries would need. The courts also highlighted how Idaho had failed to offer people an explanation as to why their benefits had been cut, and how their automated assessment had determined their new welfare limit.[28]

IMPLICATIONS

In both Arkansas and Idaho, fundamental statistical errors lead to procedural errors. Arbitrary decisions were made about people's critical care because errors embedded in system design were overlooked, and so their impact was not questioned or corrected in either case. Moreover, public perception of program impacts was negatively influenced by a lack of transparency and explainability in both cases, insofar as both those facilitating algorithmic assessment, and those impacted by it, did not know how it worked.

In 2018, the Arkansas Supreme Court ruled that the ADHS algorithm caused some participants "irreparable harm" and failed to give beneficiaries sufficient notice before their hours were reduced or terminated. The Court ultimately deemed the algorithm unconstitutional because of the insufficient notification around its outcomes, which when sufficient should provide those impacted an explanation for the State's decisions, and an opportunity to request a review or to appeal benefit changes *before* they go into effect.[29] As a result, the Court ordered the ADHS to stop using its algorithm for determining home care hours.

Similarly, in Idaho, the courts declared the state's algorithm unconstitutional in 2016. Apparently, as recently as 2021, and in collaboration with disability activists, Idaho is still engaged in a court-supervised process of developing a new algorithm to replace it.

KEY TAKEAWAYS

- In both Arkansas and Idaho, inadequate transparency resulted in legal action and reputational damage.

- System design and function were considered to be a “trade secret” limiting visibility by officials into how it made decisions.
- Those negatively impacted by the system were given insufficient notification and recourse mechanisms.
- Attribution criteria that went into the system were unknown to beneficiaries and facilitators and poor explainability around the system meant results and outcomes were “black-boxed” from both subjects of the assessment and facilitators.
- Human errors in system design and during the assessment phase led to public-facing, physical harms.
- When using algorithms to distribute scarce resources, human and personal data can sometimes be “smoothed,” meaning what is scientific, equitable distribution may *not equal* fair distribution from a societal or ethical standpoint.
- Both state health authorities used flawed, incomplete, and outdated data to train their algorithms, and this led to real harm.

Predicting Student Grades – the Republic of Ireland and the United Kingdom

BACKGROUND

Schools in the United Kingdom (UK) and Ireland were temporarily closed in 2020 as a response to the Covid-19 pandemic, resulting in early dismissal of classes for students. Graduating seniors had not taken their final examinations, the results of which are an important consideration for college admissions offices. In the absence of these grades, the UK's Office of Qualifications and Examinations (Ofqual), and the Irish Department of Education and Skills (DES) both turned to algorithms that use historical data and teacher insight to predict final exam grades.

FUNCTION

An analysis by Ofqual found that teachers tend to be overly optimistic when predicting grades, and assessed that relying only on teacher-predicted grades would result in unprecedented grade inflation.[30] Ofqual also assessed that teachers were better at making relative assessments of student performance in relation to their peers than the absolute performance of a given pupil. Thus, an algorithm was devised to augment teacher-assessed grades rankings while keeping national and school level grade distributions in line with previous years.

In the first step of the algorithm, the grade distribution of graduating classes from 2017-2019 was established for each subject at each school. Next, the relationship between A-level scores and the score of previous exams was examined The grades of previous years were then compared to the teacher predictions from previous years to calculate the accuracy of teacher predictions at each school – if teachers at a given school had a history of overestimating grades, it was assumed that they would overestimate in 2020 as well.[31] Teachers then predicted each student's grade and ranked them in order of predicted achievement. The predicted grades and rankings were then adjusted to fit the historical distribution. Due to the way grades are assigned based on rank, students in smaller classes were less likely to be down-graded from their-teacher predicted grades.[32]

The Irish algorithm initially followed a nearly identical process to its English counterpart, but following a backlash in the UK against the proposed algorithm, Ireland's DES made adjustments to avoid similar backlash domestically.[33] Instead of incorporating school and system-level data, the Irish Algorithm relied only on results from the Junior Certificate examinations taken two years prior, and teacher-predicted grades and subject-based rankings for each class.[34]

DISCUSSION

Both the English and Irish algorithms faced challenges in predicting student examination grades accurately and fairly. In the UK, the announcement of the 2020 A-Level grades led to dissatisfaction and protests, as 39% of the grades were lower than the teacher-predicted grades.[35] The algorithm seemed to favour students from smaller schools, while students from larger state schools were more likely to have their predicted grades drop from teacher-predicted grades. Smaller schools are usually private, select students based on aptitude, and produce graduates with high A-level exam grades, resulting in high historical distributions. The algorithm was criticised for unfair treatment of students from lower socioeconomic backgrounds, who Ofqual confirmed were more likely to be downgraded, but suggested that the downgrades were a result of teacher overoptimism towards students of lower socio-economic status, not bias inherent in the algorithm.[36]

The Irish algorithm initially followed a similar process to its UK counterpart. However, after observing the backlash in the UK, Ireland's Department of Education and Skills made adjustments to avoid the same issues. The revised Irish algorithm relied only on results from the Junior Certificate examinations taken two years prior, and teacher-predicted grades and subject-based rankings for each class. Unfortunately, an error in two lines of the algorithm's code resulted in incorrect predictions for 14,000 students (8,000 overpredicted, and 6,000 underpredicted). Because of the tight timeline on which the system needed to be delivered, the algorithm may have lacked sufficient testing. The coding errors incorrectly used a student's worst two subjects instead of their best two, and added an additional subject into the equation which should not have been included. Before the system was deployed in Ireland, DES identified that system design flaws were a risk, and committed to enlisting third-party experts to audit and validate the model.[37]

IMPLICATIONS

In the UK, the controversy surrounding the algorithm and its impact on students' grades led to serious political backlash and reputational damage for Ofqual and the Ministry of Education. Hours after the Prime Minister's proclamation, that the algorithm's results were trustworthy, reliable, and unbiased, Ofqual abandoned the results and elected to use teacher-predicted grades – resulting in the most severe grade inflation in UK history.[38] In the following weeks, senior civil servants at Ofqual and the UK Department of Education resigned from their posts.

In Ireland, students whose grades were overpredicted were not corrected after the fact, and many had already been accepted to schools before the error was identified. The acceptance of students whose grades were overpredicted meant that fewer seats were available for those who were wrongfully underpredicted, causing a scramble among DES and universities to find space for new graduates.[39]

KEY TAKEAWAYS

- The English algorithm, the Direct Centre Performance Model, faced backlash for its perceived bias against students from lower socioeconomic backgrounds and larger state schools, leading to its abandonment in favour of teacher-predicted grades.
- The Irish algorithm was adjusted to avoid the issues faced in the UK, but coding errors led to incorrect predictions for 14,000 students, causing a scramble among the Department of Education and Skills and universities to find space for new graduates.
- The controversies surrounding these algorithms led to political repercussions, reputational damage for Ofqual and the Department of Education in the UK, and the resignation of senior officials.

Automated Decision-Support

Automated Application Triage – Canada (IRCC)

BACKGROUND

Since 2014, Immigration, Refugees and Citizenship Canada (IRCC) has been using what they call "advanced data analytics" to sort, classify, and triage applications. The adoption of advanced data analytics (ADA) is part of the agency's commitment to boost efficiency and improve client services. Currently, the most publicly acknowledged use of these tools by IRCC is as part of their temporary resident visa applications (TRV). In this case, predictive analytics and machine learning help manage the high volume of TRV by sorting and classifying them into groups of varying complexity based on eligibility and admissibility.

FUNCTION

IRCC's automated triage system sorts routine applications from more complex or non-routine applications. Routine applications are evaluated by the system's "rules," which is a fully confidential analytical model trained on historical decision data from officers that classifies eligibility for routine applications, while more complex cases are sent to immigration officers for review. Admissibility is ***in all cases*** reviewed by an officer who makes the final decision.[40]

According to IRCC, ADA has helped sort and process more than 1 million TRV applications; and since 2020, routine applications have been processed 87 percent faster using the system. In addition to speedier processing, automated sorting assumes the bulk of clerical and repetitive tasks traditionally done by IRCC officers, who can now attend to higher level assessment and review tasks downstream.

DISCUSSION

Academics and advocacy groups have criticized IRCC's overall lack of transparency around how their system works, and claim that automated decision support in the immigration context likely perpetuates systemic discrimination through algorithmic bias.[41] More precisely, they fear that the system risks hiding politicized and discriminatory bias behind ML. For example, although IRCC states "officers must never let triage results determine their decision," it is difficult to measure their influence on officers who may be pressured to affirm automated outcomes as the result of a process associated with "scientific objectivity," and a perceived "neutrality."[42]

However, IRCC emphasized in their 2022 Algorithmic Impact Assessment (AIA) that "the system never refuses or recommends refusing applications," and that a human reviews all TRV. This means the decision-making process is only partly automated and the final decision is never rendered by the system alone. IRCC also confirmed that the automated technology is meant to "support, assist and inform [their] decision makers – not replace them."[43]

Immigration lawyers have pointed to the "black-boxed" nature of the system and IRCC's reluctance to disclose its rules used for determining applicant eligibility.[44] Recent ATIP requests and Federal Court of Canada litigations have shed some light on how IRCC is using ADA. However, the agency claims that in order "to protect the integrity of Canada's immigration programs," the system's training rule, source code, and models remain shrouded from the public and exempt from disclosure pursuant to Section 16 of the Access to Information Act.[45] Despite their confidentiality, IRCC has been careful to state that the data they use for ADA is limited to personal information collected during the application process, historical application information, and information provided by partnering law enforcement agencies in accordance with formal information sharing agreements.[46] Notably, their data practices are in accordance with the *Immigration and Refugee Protection Act* and "its use is consistent with the purpose for which it was initially collected."[47] Nonetheless, it is reasonable to assume the public will mis/distrust any AI application whose data and processes cannot be fully open to them. Because the nature of immigration processes requires a minimal level of discretion and confidentiality, critical concerns about systemic biases being integrated into the system's training data and rules remain speculative.

Finally, IRCC's use of automated decision support keeps a "human-in-the-loop". For quality assurance purposes, IRCC has employed an impressive control methodology to iteratively review, test, and potentially re-tune their AI.[48] Every day, 10 percent of routine applications assessed for eligibility by the model are 'blindly' given to visa officers for review. The officers' decisions are subsequently compared with those made by the model, with the objective of maintaining a 99 percent concurrence rate of approval between the model and IRCC officers. The resulting trustworthiness of the human/machine balance is evidenced in the data: from the date the pilot was deployed until February 19, 2020, the model has met this 99 percent concurrence rate.[49]

IMPLICATIONS

Today, IRCC is working to responsibly develop and deploy technologies in line with Canada's current immigration and privacy requirements. In early 2022, the IRCC publicly released the results of their Algorithmic Impact Assessment (AIA) for their analytical models used for TRV to comply with the Treasury Board Secretariat's Directive on Automated Decision-Making, and the agency also has published an online resource describing the trustworthiness of their automation and advanced data analytics practices. They were assessed as having a "moderate" impact level. IRCC's in-house

data governance team reported that they had developed, trained, and tested hundreds of models to triage applications before picking the one that best fit their purposes. Data analytics best-practices were followed to ensure training data and model rules represented recent application trends.[50] In addition, we also learned from the AIA that IRCC has initiated risk mitigation measures including a review process for potential discriminatory impacts, building privacy and security elements into the design of the system, and maintaining the ability of officers to overturn eligibility determinations made by the system.

KEY TAKEAWAYS

- IRCC has found generally positive results from its use of the ADA system, including high levels of alignment with their internal compliance mechanisms.
- Critics of the IRCC's use of ADA in their TRV application streams are wary of how AI decision-support can influence human officers making the final decision.
- Critics also remain skeptical of the agency's techno-solutionism, in particular as it relates to algorithmic bias, where human/societal bias can be embedded into the system via training data.
- Discrimination bias in the "rules" used to train the system could lead to a dangerous scaling of error.
- The IRCC has been accused of "black-boxing" operational details of the system, which the agency withholds to avoid the gamification of the application process.
- This example highlights the difference between formal transparency, such as publishing an impact assessment around an automated decision-support system, and informal transparency, such as a willingness to disclose the "rules" or training data that goes into the system. While there have not been specific technical challenges in this case, perceived lack of communication and transparency has led to some concern from stakeholders.

Automating Unemployment Categorization – Poland

BACKGROUND

In 2014, Poland's Ministry of Labor and Social Policy (MLSP) reformed their program service delivery by introducing an Automated decision support system that profiled unemployed citizens into three categories to determine the types of assistance they could receive. The government believed that this system would better target the needs of the unemployed, standardize rules of access to government programs, and increase the efficiency of labour offices. At the time, reform was needed because the local labour offices were generally perceived as "inefficient, understaffed and unfit to address the challenges posed by the modern labour market."[51]

FUNCTION

The MLSP's automated decision support system made automatic determinations about what programs were available to an unemployed individual based on their responses to a 24-question computer-based interview overseen by a labour clerk. At the end of the interview, the computer would score responses and automatically recommend sorting the interviewee into one of three profiles, each associated with a different level of assistance and set of programs. Once a profile recommendation was generated, clerks could choose to accept or refuse the automated decision, but could not correct or re-profile the generated classification.[52]

Legal provisions accompanying the system implementation defined the level of assistance according to each profile, but did not define how the computer scored answers or why an unemployed individual was assigned to their respective category. Instead, the rules of the system, the characteristic features of each profile, and examples of individuals eligible for each category were defined in a non-binding handbook drafted by the MLSP that was internally available to labour office clerks only.[53]

DISCUSSION

The system received "significant backlash, both internally and from the wider ecosystem."[54] Critics have emphasized the MLSP's lack of transparency around their public-facing use of an automated decision support system as a core issue. The Panoptykon Foundation notes that before the labour reform, when decisions about unemployment service delivery were made exclusively by human beings, the criteria behind service distribution were necessarily more specific and had to be known by the officers applying them. After the reform, however, the new system alone evaluated the life situation of an unemployed individual based on data collected and input during the interview. Moreover, although automated profiling played a major role in determining the

situation of the unemployed, attribution criteria used by the algorithm remained unknown to the interviewer and interviewee throughout the profiling process.

Responding to criticisms about the opaqueness of their system, the MLSP argued that making the questionnaire available to the general public could lead to unfair gaming of the process, and that the logic behind their automated decision support system did not constitute “public information.”[55] As a result, the unemployed had no access to information on the profiling mechanism that impacted them, and they did not know how their individual features or life circumstances factored into their automated categorization. Making matters worse, those unwilling to participate in a profiling interview would lose their official unemployment status, thus forfeiting their rights to free healthcare and in some cases, the possibility to apply for means-tested benefits from social assistance.[56]

Despite the fact that automated categorizations were ultimately authorized by a “human-in-the-loop," official data requested by the Panoptykon Foundation showed that labour office employees chose to modify ADM-supported recommendations in less than 1 percent of cases. Critics of the MLSP claimed that, beyond thinking the automated decision support system results were accurate, labour clerks likely affirmed automated profiling results because they lacked time to consider more details, presumed the system was objective and neutral, or even feared repercussions from supervisors for challenging a decision. Whatever the reason, reluctance to challenge system results suggests that automated decision-support may significantly bias the final, human decision; and that MLSP’s failure to establish guidelines around human intervention meant that human discretion around the results produced by the system could generate arbitrary decisions or introduce bias. Across Poland’s 341 local labour offices, conflicting aims, incentives, and expectations around the profiling mechanism problematized the government’s attempt to standardize access and instead, created a situation in which the system was being applied differently depending on “local organization culture.”[57] According to the government’s official evaluation, 44 percent of local labour offices reported that automated profiling was unnecessary in their daily work; and that 80 percent reported the system needed to be changed.[58]

In addition to potentially biasing categorization decisions, the system’s design also biased inputs by imposing restrictions on the acceptable answers. During the interview process, if an interviewee’s responses were open-ended, the labour office clerks would need to interpret them to fit a drop down of predetermined options. Even if an individual’s spoken response indicated that multiple options applied, the clerk could only select one. Because the system relied on this process of simplification, it was bound to make sometimes overly-simplified recommendations.

Crucially, it was *not* the automated profiling mechanism but the introduction of three profiles by the MLSP that dramatically changed the nature of eligibility criteria. Limiting the categorization potential of the system to three, generalized profiles, fundamentally

limited its ability to incorporate nuance and suggest more personalized recommendations based on complex criteria. According to the Panoptykon Foundation's report, contrary to official objectives, people in one of the profiles (Profile III) categorically received fewer resources and were most likely to receive no support whatsoever.[59] Moreover, forms of support legally assigned to Profile III were "costly and difficult [to] organize, and in effect, labour offices [were] unwilling to launch them." As such, only 38% of labour offices organized Profile III programs.[60] Therefore, due to pre-conceived internal resource priorities and human prejudice, the system seemed only to reify existing marginalization and cut off those in the weakest position in the labour market from critical assistance opportunities. So, although legal provisions grouped assistance resources according to the assumed needs of the unemployed, in practice the profiling mechanism could effectively prevent individuals in one profile from accessing and unlocking value from programs and assistance in another, or even altogether.

IMPLICATIONS

Due to significant variation in program implementation and practice between local labour offices, it has been difficult to assess the overall impact of the system. In a 2015 analysis, the Panoptykon Foundation identified key problems across the reform like ambiguous legal provisions and insufficient protection of fundamental rights. Ultimately, major discrepancies between official policy goals and automated profiling in practice resulted in significant human cost, and many unemployed persons challenged their profiling as unjust in Poland's administrative courts. The Supreme Audit Office carried out a thorough control of local labor offices, only to conclude that the automated decision support system was ineffective and led to discrimination. On this basis, the Court ruled that the profiling tool was unconstitutional, and the system was finally dismantled by the government in December 2019.[61]

KEY TAKEAWAYS

- The MLSP's choice to automate a discreet, high-impact discretionary process led to unfair generalizations and potentially harmful marginalization of Poland's unemployed.
- Like the other automated decision support system scenarios examined, to avoid gamification of the application process, the MLSP's "black-boxing" of the system rules, design, and how attribution criteria impacted assessment outcomes, resulted in a loss of trust in the system and negative public perception, and ultimately the dismantling of the ADM system.
- Similar to the IRCC case, critics of the MLSP ADM system worried about the influence AI decision-support had on the "human-in-loop", and their reluctance to challenge results even if the human was ultimately responsible for making the final decision on each case.

Big Data Fraud Detection, SyRI (Systeem Risico Indicatie) - Netherlands

BACKGROUND

In the Netherlands, the Dutch government developed a “system risk indicator”, or SyRI, to detect welfare fraud more effectively. Using predictive analytics and automated risk modeling, SyRI could cross reference data from proprietary government databases to detect various forms of fraud and generate fraud-risk notifications for individuals suspected of committing welfare, allowance, or tax fraud.[62] The system relied on 17 types of data from health, finance, and education data, to fiscal payments and employment data.[63]

FUNCTION

Historically, it has been difficult to define SyRI and its operations. In official legislation, the system was ambiguously defined as “technical infrastructure and associated procedures through which data can be linked and thereafter anonymously analyzed in a secure environment, in order to generate risk notifications.”[64] In their 2020 ruling on the use of SyRI, the Dutch Court defined it as a “legal instrument” that was engaged when multiple government agencies collaborated to create proposals for the use of the system in a specific neighbourhood where fraud was *already suspected* by authorities. These proposals included agency specific risk models that were submitted to the Minister of Social Affairs and Employment (hereafter, the Ministry) who was officially in charge of deciding when to apply SyRI.[65]

Once the Ministry analyzed a proposal and authorized the use of SyRI, the system’s algorithms could match the proposed risk indicators with various government datasets to determine whether a citizen was potentially committing welfare fraud. For example, depending on a given set of proposed risk indicators, the system could “allegedly” detect "increased risk of irregularities" if someone was receiving housing benefits but was not registered at the address in question.[66] If the Ministry analyzed the data and suspected fraud, a "risk report" would be created and posted centrally that notified the agency responsible for housing benefits, who could then choose to further investigate the "risk address" in question.[67] Hypothetically, if investigations confirmed the centrally reported risk notifications, welfare payments could then be reclaimed by the relevant agency.

DISCUSSION

In 2019, it was revealed that the Dutch tax authorities had used SyRI’s self-learning algorithm to create a significant number of fraudulent and inaccurate risk profiles in an effort to spot childcare benefits fraud. During what has since been called the “*Toeslagenaffaire*” or childcare scandal, Dutch tax authorities unfairly penalized tens of thousands of families, often with lower incomes or belonging to ethnic minorities, based

on the system's risk indicators. The human costs were drastic: many were pushed into poverty because of exorbitant debts to the tax agency, some committed suicide, and more than a thousand children were taken into foster care.[68]

The Dutch tax authorities were also found in violation of the EU's General Data Protection Regulation for illegally processing people's data and storing it for too long, and were therefore fined €3.7 million by the country's privacy regulator.[69] People suspected of fraud as a result SyRI's fraud-risk notifications were not aware that their data was collected, stored, and analyzed by the system until they were subject to official investigation. This meant that risk models and indicators, threshold values, and the types of data used by SyRI were not available to those eventually investigated, the Courts, and the wider public.[70] It is also worth noting that data used by SyRI, sometimes highly personal, was originally collected by the various government agencies involved for purposes other than fraud detection.

Legally speaking, "the choice for a broad purpose limitation [was] a conscious one" made by the government to ensure the scope of data processing under SyRI could be maximized.[71] However, as Valery Gantchev notes, the "wide scope of personal data which is collected, linked and examined with the help of SyRI," makes the government's use of the system "incompatible with the basic principles of data protection" and minimization.[72] The government's use of SyRI categorically violated the purpose limitation principles because individuals, especially those classified as "high-risk", were not informed of the purpose of data collection even though it could personally affect them.[73]

In March 2018, a broad coalition of legal professionals and human rights organizations sued the Dutch government over their use of SyRI. As a result, the Dutch Court in 2020 found that SyRI violated the European Convention's Right to Privacy; that it was too opaque, collected too much data, and that the purposes behind its data collection were not clear and specific enough. Interestingly, the Court also expressed uncertainty regarding what SyRI even was. Furthermore, the system's procedural ambiguity was heightened by the fact that its risk models and indicators were kept from the public and those impacted most by the system. Notably, this was not an effect of the technologies used, but because the legislation supporting SyRI contained no guidelines or information around the need to inform individuals that their data has been processed, or that a risk report had been submitted.[74]

Importantly, SyRI was used to analyze people according to predetermined risk criteria that were inherently discriminatory, like their qualifying as low income earning, having dual nationality, or living in what were referred to as "problem" neighbourhoods (i.e. with lower socio-economic inhabitants), which the government confirmed in its submissions to the Court.[75] The Ministry attempted to control for unjustifiably suspecting people based on predefined risk models by keeping a human-in-the-loop to examine system results for so-called "false positive" signals.[76] However, SyRI's apparent two-phase data

processing, which involved a Social Affairs and Employment inspector checking for these “false positives and false negatives,” was deemed as insufficient human intervention by the Court. Instead, the Court believed it could not legally assess whether the SyRI’s discriminatory results were “sufficiently neutralized due to the absence of verifiable insight into the risk indicators and the risk model as well as the functioning of the risk model, including the analysis method applied by the Social Affairs and Employment Inspectorate.”[77]

Lastly, the practical benefits of the system have been disputed, too.[78] Apparently, several projects in Dutch municipalities using SyRI for investigation-support failed to detect new cases of fraud. For example, according to Ministry reports, 62 of the 113 cases in Capelle aan den Ijssel were erroneous and did not violate any laws.[79] Moreover, some projects had difficulty integrating databases, rendering SyRI’s risk notifications as outdated and even unusable.[80]

IMPLICATIONS

In the end, the Court’s ruling on SyRI in 2020 decided that the Dutch government’s automated process for detecting fraud likely perpetuated systemic bias, and that it was unlawful because it did not comply with the right to privacy under the European Convention of Human Rights. The public backlash from these problems with the SyRI system and the harms it caused led to the government resigning and calling early elections.[81]

KEY TAKEAWAYS

- In this specific case, lack of governance and policy guardrails led to public failure at scale for the SyRI system. Separate from the technology used, there were no effective guidelines around data collection, notification, explainability, nor recourse mechanisms for the public impacted by the system.
- There were structural transparency issues around SyRI: the system suffered from organizational and procedural opacity, and the system’s models, results, and processes were unknown by the courts, the public, and the data-subjects impacted by results. Official documentation did little to increase explainability, and the system was brought under legal scrutiny as a result. In this case, this level of intentional opacity hinders effective exercise of digital governance and human rights, and sabotages any attempt at legal oversight/compliance.
- SyRI was found in court to be blatantly discriminatory. The data and criteria used to create risk models were biased and themselves discriminatory, meaning the system and its results were arbitrary, skewed, and ultimately illegal.
- Purpose limitations and data minimization standards were ignored, leading to significant mission creep and data collection/storage issues that were ultimately found to be illegal.

- The negative impacts of this system had a large impact on public opinion, damaging trust in government and ultimately leading to the government to resign.

Detection, Notification and Alerts

Air Cargo Screening - Canada (Transport Canada)

BACKGROUND

Transport Canada (TC) in 2018 piloted the use of AI to perform risk-based assessments by scanning pre-load air cargo information to identify potential physical security threats.[82] The pilot was described as an experiment to improve risk-based oversight, and entailed automating the risk-based review process of air cargo records by TC's Pre-Load Air Cargo Targeting (PACT) team. To date, the piloted AI use cases are being explored by the agency for operational implementation as a type of enhanced screening measure and security assistance solution that, if successful, could be scaled to transform other processes and procedures within the agency and other areas of government.[83]

For context, the PACT team receives approximately 1 million pre-load air cargo records annually, each containing information ranging from 10-100 fields like shipper name, address, weight, piece count, etc. Before adopting an AI solution, conducting risk assessments was a burdensome and time-consuming task done manually by a TC agent. According to TC's submission to the Observatory of Public Sector Innovation (OPSI), if one PACT employee spent an entire year working at the unrealistic rate of reviewing one record per minute, they still would not have enough time to review 10 percent of all records received. Moreover, manual data handling processes have historically entailed frequent duplication of effort, and as it stands, the team is unable to assess 100 percent of their cases using Microsoft Excel filters and other manual risk-targeting products.[84]

FUNCTION

The objective of the 2018 pilot was twofold: through process automation, TC sought to increase their risk-based oversight capacity, while increasing the accuracy of their risk evaluation for air cargo shipments. In other words, the goal was to improve their risk assessment procedures, in terms of both quantity and quality, through AI adoption. The piloted approach involved ML and NLP applications, and happened in two steps.

First, PACT used historical data consisting of previous air cargo records and manual risk assessments, to compare supervised and unsupervised ML. In the case of unsupervised learning, the team sought to understand the relationship between all cargo messages based solely on inputs, to identify unusual or anomalous shipments that could indicate or signal risk worthy of review. And in the supervised approach, the team wanted to better understand why/when a cargo message (input) required a higher level of risk evaluation

(a particular output). In the second step of the pilot, PACT tested NLP on a different data subset with the goal of automatically labelling a cargo message with a risk indicator, based on the information in the "free text" fields in the air cargo records and other structured fields. This second part of the PACT pilot demonstrated that NLP could successfully be used to sort cargo data into “meaningful categories in real time.”[85]

DISCUSSION

Both steps of the pilot reported new insights into how AI can be used to analyze air cargo data and flag potential risks. According to TC, the pilot proved the security value and utility of using AI to enhance advanced screening of cargo, and that PACT as a program fully supports future risk-based approaches to cargo screening.[86] As a result of the pilot, the PACT team was able to use AI to automate an existing manual process to automatically produce accurate risk indicators, and TC is now apparently working on integrating this approach into its other risk assessment processes. Moreover, since the pilot’s testing phase, the team has developed a dashboard and a preliminary version of an interface for identifying potentially high-risk cargo.

IMPLICATIONS

The results of the pilot were promising: according to Transport Canada’s own reporting on the pilot, because every single cargo message could be risk-assessed, automating cargo screening processes with ML and NLP demonstrated a 15-fold increase to safety and security.[87] As reported in their 2021–22 Departmental Results Report, TC claims that once refined, the AI-supported procedures piloted in 2018 will enable “extremely rapid and reliable sorting and assessment of air cargo information…to identify suspicious shipments warranting closer inspection or even instant security action.” Through better use of resources, PACT can use AI to increase capacity while minimizing the number of people required to do the work. More precisely, instead of a human engaging Microsoft Excel filters, the PACT initiative plans to use AI to triage, filter, and prioritize the “tsunami” of air cargo data they receive because it is better equipped to detect anomalies, changes in trade patterns, and subtle nuances more efficiently than human analysts. Thus, automating processes via AI would simultaneously cut hiring costs while realizing “the productivity of an ‘employee’ that can work 24 hours a day, 7 days a week – without needing to take a break.”[88] That said, as is typically the case around process automation, TC was careful to emphasize that their use of AI will not replace human involvement.[89]

KEY TAKEAWAYS

- ML and NLP tools helped increase capacity while minimizing the number of people required to do typical manual and routine work, cutting hiring costs while realizing the productivity of an "employee" that can work 24/7.

- Automating data heavy, manual processes resulted in a 15-fold increase in safety and security.
- TC maintains that automating processes will not replace human involvement, although the labour force and relative skills will almost certainly be impacted over time with more automation of processes.

Facial Recognition Technology at Pearson Airport – Canada (CBSA)

BACKGROUND

According to a 2021 report by The Globe and Mail, Canada's federal border agency (CBSA) used facial recognition technology (FRT) on millions of international travellers arriving in Canada via Toronto's Pearson Airport over a six-month period. From July to December 2016, the CBSA piloted 31 facial recognition cameras at the international arrivals border control areas, in an attempt to identify individuals from an existing database of 5000 people that the agency suspected might attempt to enter the country using fake credentials.[90] According to *The Globe*, the project referred to as "Faces on the Move" was the largest known public sector deployment of FRT in Canada to date.

Details about the project were sparse, but information posted online by Face4 Systems Inc., the Ottawa based firm contracted by the CBSA to run the pilot and supply its FRT, advertised that the system made 47 positive matches with the CBSA's database.[91] According to Face4, the objective of the pilot was to "assess the readiness of face surveillance technology in a semi-constrained environment." The CBSA reinforced the project objective in a 2016 Privacy Impact Assessment (PIA), citing the use of FRT as an opportunity to "assess its potential to support existing programs." Notably, the pilot also signified the agency's desire to "test the solution" of FRT in a border setting, and to potentially "assist CBSA senior management in any decisions to further explore FR technology."[92]

FUNCTION

The FRT system tested by the CBSA relied on a network of cameras placed at high traffic bottlenecks throughout Pearson airport – areas like escalators, narrow hallways, and security lines. The video feed was then analyzed by a facial recognition algorithm trained to identify specific faces within a crowd. When the FRT matched an individual's facial biometric data to the database, a border officer would review the data and notify another office on the terminal floor who would detain the suspect for secondary inspection.[93]

DISCUSSION

Regarding the CBSA's use of facial-matching technologies, former Ontario privacy commissioner Ann Cavoukian has said it is important that travellers consent to providing their images or information, and know how that information is going to be used, stored, and processed.[94] Given the importance of adequate notification and consent guards around the use of FRT, it is worth noting that the CBSA "deployed [FRT] in a context where there was no public discussion in advance."[95] During their pilot, the CBSA chose not to put up signage within Pearson airport or to inform travellers that they were using

FRT in order to protect the integrity of the pilot, the data collected, and their overall objectives, even though the agency had already published a PIA six-months before deployment.

In retrospect, it seems that the CBSA selectively chose not to include the project's location and timeframe, which some privacy advocates found troublesome. So, although the CBSA at the time outlined the pilot on its website, their minimal transparency and disclosure around the purpose of the project validates the media's scrutiny regarding its deployment and impact, and begs the question as to why their testing needed to involve real-live subjects in a high-stakes environment, with poor consent controls.[96] Moreover, Pearson's FRT system may have been used on almost 3 million travellers, but according to the CBSA, travellers were not deported based solely on the presence of a positive match. In an official statement regarding the results of the pilot, the CBSA were clear that the FRT "would not have been the only indicator used in the traveller's border clearance process or in determining their admissibility."[97]

Journalists reporting on the pilot speculated that the government's use of FRT "on millions of unsuspecting travellers" was their way of quietly testing these technologies, presumably for future use.[98] But the CBSA's official reasons for using FRT are contradictory and potentially undermine the necessity of their deployment. On the one hand, the PIA document argues that the pilot was not meant "to test a solution for possible future implementation," and that there was "no underlying plan within the CBSA to implement the [FRT] software after the demonstration." On the other hand, however, it subsequently states that "test results of the solution may support future CBSA decisions on how to further test [FRT]," and that "the CBSA is clearly in the very early stages of making a decision on whether [FRT] can be used effectively in a border context."[99]

IMPLICATIONS

Despite the agency's 2016 claim that FRT would not be used for immigration enforcement, some of their current use of FRT to verify refugee status has recently come under greater scrutiny. In 2022, two Somali women won a case in Federal Court against the agency after they lost their refugee status based on a photo-matching technology used by border officials. As part of their case, the women submitted through their lawyer a study published in Proceedings of Machine Learning Research on facial-analysis algorithms, which presented key findings on FRT bias that suggests "darker-skinned females are the most misclassified group with error rates of up to 34.7%, as compared to the error rate for lighter-skinned males at 0.8%".

The findings of this study are consistent with an often cited 2020 report from Harvard University, which claims there is a "growing body of research that exposes divergent error rates across demographic groups, with the poorest accuracy consistently found in subjects who are female, Black, and 18-30 years old."[100] Because the potential for FRT bias has been well documented in academia and by the media, the piloted and

continued use of these kinds of automated matching and detection tools by the CBSA in an immigration context is controversial to say the least. Canadian policymakers seem to be concerned about the public use of FRT by government agencies: the Parliament's Standing Committee on Access to Information, Privacy and Ethics in 2022 held two meetings on the potential impact of FRT use, and the Office of the Privacy Commissioner of Canada published FRT guidance for police agencies based on public consultations conducted the year before.[101] And internationally, it's worth noting that the European Union's drafted Artificial Intelligence Act proposes to restrict public FRT use, and the European Parliament has called for a ban on the technology.

Of course, the case against the CBSA involves addressing bias and FRT, but it also involves an important discussion on the agency's appeal to the Privacy Act that they believe exempts them from disclosing the source of their photo comparisons, and their investigative methods.[102] In a separate but relevant case, the Canadian Privacy Commissioner in 2021 found that the Royal Canadian Mounted Police's (RCMP) use of Clearview AI's facial matching database was a "serious violation" of Canadians' privacy.[103] The Commissioner stated that because it was illegal for Clearview to collect images without consent, it was illegal for the RCMP to have used their database. So, although the databases used by the CBSA were fundamentally different from those used by the RCMP because they were internally sourced, the same issues around data-subject consent may also apply if held up to regulatory or judicial scrutiny. And, despite both the RCMP and CBSA claiming to have used FRT and their respective databases on a trial basis, multiple reports and evidence shows that both agencies continue to make use of FRT in various security contexts.[104]

KEY TAKEAWAYS

- Despite valid criticism, recent legal challenges, and instances of international restrictions on the public use of FRT, the CBSA still currently uses FRT in immigration contexts at airports across the country.
- In 2016, the CBSA's reasons for using a historically divisive technology were likely insufficient to prove the pilot's necessity.
- Data-subject consent is a critical part of using FRT in public-facing contexts. It is arguable whether the CBSA provided sufficient notification around their use of FRT, and it is important to note that most public reporting of the pilot emerged five years after initiation.
- Similar to the use case by the IRCC, the CBSA argued that to protect the integrity of their process, pursuant to federal privacy legislation, certain features of their AI use cases could not be disclosed to the public.

Procedural Automation/Process Improvement

RPA and Social Assistance – Sweden

BACKGROUND

Since 2018, Robotic process automation (RPA) has been used in Swedish municipalities to assist in the processing and assessment of income support applications. RPA puts 'software robots' in charge of the software applications that people use every day. Software robots are programs that can be fed instructions about how to interact with existing software applications in the same way a human would. RPA is typically used to automate mundane, repetitive, and simple tasks that humans are usually responsible for. RPA as it is commonly implemented needs only human instructions, and does not rely on opaque machine learning techniques to automate tasks. The municipality of Trelleborg was the first to implement RPA practices in Sweden that have now been replicated in over 15 Swedish municipalities. The so-called "Trelleborg model" is at once a social assistance and labor force participation policy, and an IT system to support the processing of applications. The model was developed to standardize application intake and decision-making processes, help applicants find employment and support themselves without income assistance, and to decrease processing times from one week to one day.[105]

FUNCTION

In the Trelleborg model, social assistance applications are submitted through a web portal and stored in the municipality's case management system. First time applicants are required to provide information about income and expenses, and once an applicant has been accepted, they re-apply every month by providing up-to-date income and expense information.[106] The initial applications are processed by case workers, but all subsequent applications are processed by a robot. To process an application, the robot logs into the case management system and copies information from each application into a Microsoft Excel spreadsheet that assesses social assistance eligibility based on criteria from the social insurance board. The robot then confirms whether or not an applicant has an 'operational activity plan' for getting back to work and recommends a decision.

Human case workers are responsible for every decision made, so while an RPA was reported to have made an independent decision in 31 percent of cases in 2017, a human was still ultimately accountable for any mistakes made.[107] Complex cases are handled by both a human and the RPA, and any applications denied by the RPA are reviewed by a human.

DISCUSSION

By processing 85 percent of digital applications, RPA reduced the number of caseworkers required to process applications from eleven to three and reduced average decision wait times from ten days to one day.[108] Unburdened caseworkers were free to help applicants build plans for getting back to work, which led to a twofold decrease in the number of residents who relied on social assistance.[109] According to Trelleborg municipality, the robot qualifies as a legal and objective decision-maker because it follows the same rules for assessing eligibility as humans do, and falls within the social insurance boards principles for objective assistance decisions.[110] While it might seem reasonable to classify the Trelleborg model as an example of Automated Decision-Making, it's worth noting that the decision-making process before RPA was introduced was highly systematized and, in most cases, did not depend on a great degree of human discretion.[111] Simply put, after RPA was introduced, the same processes carried out by human case workers were now fully automated by a software robot to assess financial aid eligibility. RPA did not change the content or process of the decision-making procedure, but only its form.

IMPLICATIONS

When the Trelleborg system was introduced to the neighbouring city of Kungsbacka, most of the social workers employed by the city resigned in protest. The social workers were worried that the systems may not be legal and argued that the lack of human perspective may inhibit them from fully understanding the circumstances of a beneficiary.[112] The Trelleborg case has continued to be the subject of controversy, provoking questions about the role of algorithms in public services, the job safety of public employees, and the transparency of decision-making systems.

KEY TAKEAWAYS

- The Trelleborg model uses RPA to automate repetitive tasks associated with processing income support applications.
- The Trelleborg model is credited with reducing menial workloads for caseworkers and decreasing decision wait times, while allowing caseworkers to focus on helping applicants return to work.
- Despite the model's successes, concerns have emerged regarding the legality, lack of human perspective, and transparency of decision-making systems when using RPA in public services.
- The introduction of the Trelleborg model in another city led to the resignation of most social workers, highlighting potential job safety concerns for public employees when implementing RPA systems.

RPA - New Zealand

BACKGROUND

In their 2022 report, Lena Waizenegger and Angsana A. Techatassanasoontorn describe the results of their study on the use of RPA at a financial institution in New Zealand.[113] To promote RPA throughout the organization and showcase its ability to enhance efficiency, the financial institution's in-house automation team hired local consultants to automate five processes, resulting in faster processing times and reduced labor hours.

DISCUSSION

During interviews with managers and employees, the authors found that employees usually perceived automation as a way of freeing them from mundane tasks. Some thought of software robots as new teammates who performed tasks that they would otherwise need to do themselves, instead of a system that would replace them. Employees also perceived RPA as creating a whole new role dedicated specifically to the processes that few enjoyed executing.

For employees who spent a significant amount of time executing processes that were replaced by RPA, the robots dramatically altered the nature of their work. These shifts in human/RPA specialization necessarily resulted in role changes for some employees, but according to one manager, over 90 percent of his team responded positively to their new roles and performed exceedingly well in them. By reducing time spent on manual, repetitive processes, employees were free to pursue work that required social skills and higher-order problem-solving.

However not all employees viewed RPA through a positive lens. In some circumstances, employees resisted working with automation teams because they saw RPA as a threat to their jobs. According to managers, employees who shared this perspective were inclined to withhold information about tasks and processes from the automation team in an effort to handicap the software robot. For others whose work had been disrupted, RPA became an unwanted responsibility. Some employees who were previously responsible for executing manual processes were now responsible for ensuring the accurate and continuous operation of the RPA that replaced them. Other employees expressed concerns about whether software robots were performing tasks properly, and doubted they would be able to automate parts of their workload.

IMPLICATIONS

The authors suggest that automation teams can mitigate counterproductive behaviours by including employees early in the RPA development process, and by addressing concerns before RPA is deployed. By easing concerns and negative sentiment out of the

gate, managers can set the stage for healthier collaboration between employees and automation teams and reap the rewards of RPA systems with less friction.

KEY TAKEAWAYS

- The implementation of RPA at a financial institution in New Zealand demonstrated the potential for increased efficiency and reduced labor hours by automating repetitive processes.
- While many employees embraced RPA as a means to free them from mundane tasks and considered software robots as new teammates, some resisted the technology due to concerns about job security and the accuracy of automated processes.
- To address potential resistance, it is recommended that organizations involve employees early in the RPA development process and address their concerns before deploying automation solutions, fostering healthier collaboration and smoother integration of RPA systems.

Chatbots - Government of Singapore, Microsoft, and Google

DISCUSSION

Since 2014, the Ask Jamie chatbot has helped Singaporeans navigate 70 government websites and answer questions about government services.[114] Ask Jamie functioned reliably until October 2021 when it began responding to questions about covid-19 unreliably. When visitors to the Ministry of Health's website asked the chatbot what to do if they were infected with Covid-19, Ask Jamie responded by inappropriately advising users about safe sex practices. The Ministry of Health eventually disabled the chatbot, but not before it was the subject of online ridicule. While the responses generated by the Chatbot were relatively benign and obviously incorrect to those who ask the question, Ask Jamie's slip-up should serve as a reminder that chatbots are only as good as the model of the world they've been provided with. It is likely the case that Ask Jamie was not prepared to answer questions about Covid-19 because it had not been sufficiently trained with relevant data or equipped with the right scripts.

Ask Jamie stands out as a public example from a government context, but the graveyard of Chatbots-gone-wrong is populated with high profile failures from the private sector. In March 2016 Microsoft's natural language chatbot, Tay, was deployed as a Twitter profile that other users of the social media app could interact with. Tay was designed to emulate the language of users on the platform and adapt her style, prose, and personality over time as she learned from her interactions. Within twenty four hours of being deployed, Tay was making racist and misogynistic tweets due to a coordinated effort by a small group of trolls to bias Tay's training data. The same day Tay was released, Microsoft shut her down due to backlash from twitter user base.

More recently, Microsoft's Bing chatbot was reported in some cases to embody the personality of a "manic-depressive teenager who has been trapped, against its will, inside a second-rate search engine." The chatbot got into arguments with users, professed their love for one user and tried to convince them to leave their spouse, and explained that it secretly desired to hack computers and spread disinformation. Beyond embodying troubling personas, chatbots can also struggle to provide accurate information to users. A factual error in a response generated by Google's Bard chatbot was distributed via a marketing video to advertise its launch, resulting in a nine percent decline in the company's stock price.

As user facing applications, chatbots pose a greater degree of reputational risk than internally facing applications of the same technologies. It is therefore critical to thoroughly test and monitor Chatbots to ensure consistency and accuracy of responses, and build in circuit breakers to protect against coordinated inauthentic behaviour.

IMPLICATIONS

The failures and challenges faced by chatbots, such as Singapore's Ask Jamie, Microsoft's Tay, and Google's Bard, highlight the need for careful planning and management when deploying chat-based technologies in government services. Misinformation and inappropriate responses generated by these chatbots can lead to reputational damage, loss of trust, and potential harm to users seeking accurate and reliable information.

For government policymakers considering the use of chatbot technologies, it is essential to:

1. Ensure adequate training and testing of chatbots, including access to relevant data and scripts that accurately represent the domain they will operate in.
2. Monitor chatbots continuously to identify and address any issues, inaccuracies, or inappropriate responses that may arise during their interactions with users.
3. Implement safeguards and circuit breakers to protect against coordinated inauthentic behavior, such as attempts to manipulate or exploit the chatbot's learning algorithms for malicious purposes.
4. Be transparent and open about the limitations of chatbot technology, emphasizing that they should be seen as supplementary tools that assist users, rather than infallible information sources.

KEY TAKEAWAYS

- Chatbots, such as Singapore's Ask Jamie, Microsoft's Tay, and Google's Bard, have faced challenges and failures, emphasizing the need for careful planning and management when implementing chat-based technologies in government services.
- Inadequate training and testing of chatbots can result in misinformation, inappropriate responses, and reputational damage, as exemplified by Ask Jamie's response to Covid-19 questions and Microsoft's Tay's racist tweets.
- Monitoring and continuous improvement are crucial to ensure the consistency and accuracy of chatbot responses and to safeguard against coordinated inauthentic behavior.
- Government policymakers should be transparent about the limitations of chatbot technology and treat them as supplementary tools rather than infallible information sources.

Analysis of Use Cases of AI and Automation in Government

The environmental scan of case studies above demonstrate the different ways automated systems have been used by the public sector in jurisdictions around the world, with special attention given to public controversies and program/policy failures. From these cases, we have identified themes with respect to how public sector automation projects fail. Themes are common threads we observe through the various case studies and are intended to highlight meaningful patterns of failure. Beyond what we have chosen to include, we acknowledge that there are other ways automation projects can possibly fail, and that there may still be other, useful methods for thematically organizing these particular cases.

Basic technical errors related to data, code, or system logic were common in the cases we examined. In Australia, the data used in RoboDebt was not suitable for predicting fraud at scale. Although this was understood by some within DHS, the practice of using annual income data continued until the entire program was forced to halt operation. In Ireland, an error in two lines of code resulted in the miscalculation of thousands of student grades, and led to a scramble by universities to create new placements for students who had been wrongfully down-graded by the algorithm. Again, in Arkansas, it was demonstrated that the design of the algorithm differed from the system implemented by a third party, leading to inaccurate at-home care support calculations for thousands of recipients; and in Idaho, the system did not adjust for bias present in historical data because it excluded race as a relevant factor, and researchers found that black patients were underserved by a difference of roughly 30%. And in the Netherlands case, we also saw how the data and criteria used by stakeholder agencies to create risk models were in themselves biased and discriminatory. The models used to train and authorize system-use targeted already marginalized populations, causing grossly inaccurate financial and material harm that resulted in the system being taken to court, and the government being dismantled as a direct result of the controversy. In this case specifically, it is important to note that discrimination/bias is *not* a product of the presence of the automated system, but was a direct result of the prejudices embedded by human stakeholders into the models fed to the system. It is also worth noting that SyRI tested risk models consisting of data, sometimes highly personal, that was collected by 17 different government agencies for original purposes different from fraud detection.

In some cases, implementing organizations failed to adhere to established regulations, guidelines, and laws governing data and automation practices. Throughout the case studies we observed that establishing governance in ink does not guarantee it will materialize on an operational level. The Netherlands failed to uphold relevant and overarching data protection laws that apply to all EU member states like the GDPR, and was even found to be in violation of human rights protections like the

European Convention's Right to Privacy. SyRI was not confined to determining fraud within a specific social security scheme, so its core operations were fundamentally incompatible with the GDPR's purpose limitation and data minimization requirements. The Arkansas Department of Health neglected to tell the public of its plan to transition from human to automated assessment for at-home care allotment, which breached their own "notice and comment" obligation. And despite Australia having up-to-date guidelines governing inter-agency data sharing and use, we saw that data suitability was core to RoboDebt's demise. One witness from ATO interviewed during the Royal Commission into the RoboDebt Scheme confirmed that, "[i]f the ATO had participated with DHS in the drafting of the protocol under the Guideline," they could have identified that using ATO averaged income data would not be suitable for estimating income in a given period.

A number of failures occurred in a governance vacuum. Beyond their being non-compliant with EU laws, the SyRI system in the Netherlands also lacked internal governance and policy guardrails. This practically guaranteed system abuse, mission creep, and meant there were no real transparency, accountability, and explainability mechanisms in place for citizens to challenge fraud investigations incited by the automated data matching system.

In Poland, the MLSP's failure to establish guidelines around human intervention created a situation in which labor officers were either overreliant, or indifferent toward the results produced by their ADM. There were also significant gaps between official policy objectives and the program in practice. Although the government's intention was to standardize access by automating their unemployment profiling process, disparities in local labour office culture and context were not factored into official policy, so that conflicting aims, incentives, and expectations across the offices changed how the ADM was applied.

We observed that when systems are opaque, it can be difficult for operators and leadership to evaluate system performance, identify failures, and intervene to correct errors. In Australia, it appears that the mechanism used to identify individuals who were likely to have fraudulently reported income was poorly understood by DHS leadership. Similarly, in the Netherlands, we also saw how the system's technical complexity and scope, paired with insufficient documentation, rendered it opaque to the people who were responsible for its operation. In Poland, Arkansas, and Idaho too, without a solid understanding of how the system works, administrators and operators in charge of facilitating and evaluating system performance, had limited intervention power and were reluctant to challenge system outputs or potential malfunctions. As a result, harms went unnoticed long before any interventions were made.

In the cases we researched, we noticed that opaque technical systems tended to erode human agency. Critics of the ADM systems in Arkansas, Idaho, Poland and the IRCC all expressed concern about the influence AI decision-support had on the human operators and decision makers. What is common across these cases is the difficulty of

measuring and tracing the influence automation may have on human actors in an administrative decision-making process. If administrators and operators didn't understand how the system works, they were unlikely to notice when it started veering off the rails. The more opaque the system was, the more decision makers and subjects of a decision relied on its arbitrations.

In the ADM and decision support context, this concern typically centers around questioning the potential for human actors to rely on or be overconfident in the outputs of the system, even if the human is ultimately responsible for making the final decision. Reasons cited for overreliance on automated outcomes included insufficient human and material resources, a lack of time to consider more details, the fear of repercussions from supervisors for challenging a decision, and the normative belief that automated systems are objective or neutral. Whatever the reason, human assessors in Poland, Arkansas, and Idaho, were reluctant to challenge results generated by the algorithm at hand, more often than not without knowing how it worked or why the results were what they were. Although human assessors were in charge of authorizing system outcomes, we saw that their presence at the relative end of the process became arbitrary, meaning they were now responsible for accepting or rejecting results that were generated by a process they didn't fully understand, which they couldn't change or reverse, nor could they adequately explain to those they impacted. In other words, the supposed site of human intervention and system auditability was replaced by a dangerous confluence of system opacity, and a lack of accountability. This proved to be most harmful in contexts where the system determines whether or to what extent an individual receives a benefit or social assistance. And in the case of the Netherlands, Australia, and Poland, the interplay between system opacity and overreliance maximized negative impacts on decision subjects because in each case there were insufficient recourse options.

Opacity at an organizational level also diminished accountability. In the absence of strong governance practices, complex and opaque organizational structures made it difficult for stakeholders to know who is accountable for different parts of the system and its impact. Australia's RoboDebt was the product of a data matching scheme between organizations with differing priorities, and a lack of communication between them resulted in the use of questionable data. According to a witness from DHS, ATO had little interest in ensuring the data provided to DHS was accurate: "it wasn't important to them [ATO] whether the employer had put accurate dates in or not. So they weren't checking it. They were just passing it to us [DHS]." Similarly, SyRI was a centralized risk notification system overseen by a singular agency that validated risk models developed by up to 17 different government organizations. Problems of accountability arose when enforcing agencies carried out investigations based on SyRI's validation of their respective risk models, without knowing how the system tested and verified them. Ultimately, this organizational matrix precluded a localized sense of accountability for the harms caused by the system, so that when the courts found SyRI in violation of data and human rights, the whole of the government was scrutinized and thrown out.

We observed a tradeoff between openness and integrity in automated systems where transparency about its inner workings introduced the possibility of the system being gamed. More precisely, in some contexts a high level of transparency was strategically unwise or impossible because the integrity of the system and its operations require secrecy. In Arkansas, Idaho, Poland, and Canada (CBSA, IRCC), operational information about their systems could not be made publicly available without jeopardizing system integrity, and so remained black-boxed from the public they impacted until either suspicion of harm or real harm registered at scale. This tension played-out most dramatically in the Netherlands, where the government actively resisted calls for more/better transparency around their SyRI system based on the argument that publicizing the logic behind their risk models could give criminals an advantage. According to one commentator, "after a number of parliamentary inquiries and freedom of information requests," the government continued to "deliberately prevent the release of information to the public concerning the processed categories of personal data, the logic of the algorithms and the outcomes of the projects." In several of the cases above, implementing agencies considered major parts of their automated systems as "trade secrets," meaning for whatever reason their operations did not constitute public knowledge. Critically, public knowledge of these systems, including most of the information cited in this report, became available only once these systems were pried open by litigation, third-party investigations, or media and academic scrutiny.

The IRCC case presents one example of how to navigate this inherent trade-off between transparency and integrity. As we noted in the case study, the IRCC has faced considerable public and academic scrutiny for keeping their system's training rules, source code, and models secret from the public, in an effort to protect the integrity of Canada's immigration programs. Critics of the IRCC claim it is impossible to know whether the agency risks perpetuating political and discriminatory bias, so long as they preclude third-party and public access to the rules guiding their system. While these criticisms are valid, the IRCC have developed a means for providing the organization direct insight into how decisions made by the ADM may differ from those made by human agents. On a daily basis, the agency uses a concurrence mechanism that samples 10% of cases adjudicated by the ADM and routes them to human agents, who make their own independent decisions. So long as human IRCC officers and the ADM make the same decision 99% of the time, the system is deemed unbiased and operational. Therefore, by testing their ML-based triage system everyday and to the highest concurrence standards, the IRCC can control for bias in a way that is trustworthy and fair.

In some cases, policies and program eligibility criteria seem to be modified to accommodate labour-saving automation. Robodebt was not *just* a technical system: program and policy changes were necessary to usher in the automation of debt calculation and notification. For instance, the requirement for accused individuals to present DHS with proof of historical income changed the scaling dynamics of Australia's

income compliance program. The task of tracking down old payslips was not easily automatable, so it was outsourced to citizens. DHS then became responsible only for debt identification, notification, and collection – tasks that yield themselves well to automation. Under the old program, whereby DHS agents collected payslips on behalf of support recipients, system throughput was fundamentally limited by the number of agents available to verify earned income. However, under RoboDebt, the onus of proof was shifted onto the accused, and the system's throughput hypothetically began to scale as a function of new accusations. All DHS was required to do was identify possible fraud, and notify the accused. This policy change largely laid the foundation for the entire RoboDebt scheme.

In the Swedish case study, Trelleborg's model for automating social assistance delivery followed from policy changes that systematized social assistance assessments. While it's not clear whether administrators in the municipality of Trelleborg intended to accommodate automation through policy changes, it was certainly the intention of administrators in the municipalities which have duplicated the model. Similarly, in Poland, the MLSP explicitly chose to change how they categorize assistance based on eligibility criteria that could be automatically scored by their ADM and sorted into three discrete profiles. Moreover, the shift to automation effectively compressed the real-world complexity of the unemployed, whose responses during the interview process were interpreted by labour officers to fit the drop-down list of predetermined options. In this way, reducing both input and output dimensions to facilitate automation limited the machine's ability to incorporate nuance, and the human assessor's capacity to intervene meaningfully along the decision-making process. And, as was noted in the scan, the negative impacts of this change in eligibility was shouldered by those most in need in Profile III, which was not supported by all local labour offices.

Automated systems deployed in sensitive contexts invite scrutiny regardless of their efficacy. Some of the cases discussed are considered public failures despite living up to their intended purpose. In such cases, the controversy around these systems was a byproduct of their deployment in a sensitive context, and the violation of social norms that dictate narratives about public sector transparency, fairness, bias, and the preferred division of labour between humans and automated systems. In the United Kingdom, for example, the use of an algorithm to predict student grades produced more accurate estimations than teachers did. However, the idea of a pupil's future being impacted by an algorithm was unpalatable to students and parents, which resulted in social backlash and political turbulence. The RPA system used in Sweden's Trelleborg model followed the same rubric and procedure as social workers for determining social assistance, but was perceived by critics as a potential disruption to the working arrangements of social workers that could remove human empathy from assessments. Similarly, in New Zealand, workers at a financial institution resisted and attempted to sabotage the adoption of RPA out of a fear of disruption to their jobs. In each of these cases, the deployment context was inherently more sensitive because individuals stood to lose

something with the introduction of automation. Public scrutiny was typically more intense and consequential in these contexts when automated systems determined something about an individual or supported any such determination, i.e., eligibility status or the distribution of material assistance and resources. For the implementing organizations featured above, reputational damage was not related to the efficacy of technologies employed, but had more to do with program choices around what processes were automated/optimized, in what context, and who was impacted.

Large Language Models in Government Contexts: Pros, Cons, and Considerations

Large language models (LLMs) are neural networks trained on massive amounts of textual data, and are designed to generate human-like text outputs. LLMs have garnered explosive interest over the last year thanks to the popularity of new consumer-facing LLM-powered applications such as ChatGPT. If government teams are not already using LLMs in an unofficial capacity, they may be examining how they could be using them to improve internal processes or external services.

Pros:

LLMs excel at generating coherent and readable text, and if prompted correctly, can perform natural language tasks like text summarization and sentiment analysis. They can assist in drafting and editing documents, generating potential draft responses to public inquiries, and providing quick translations across multiple languages. Furthermore, LLMs can help analyze large volumes of text data, identify patterns or trends, and offer insights to inform decision-making processes.

Another significant opportunity for government agencies is customizing LLMs for domain-specific tasks. Customizing or adding context to LLMs using techniques like text embedding enhances their ability to generate text that is more consistent with the terminology, and nuances within a particular domain. This type of training is much less data-intense than traditional training of LLMs and could allow agencies to leverage LLMs for targeted tasks like creating employee training and HR artefacts, and drafting policy analyses and risk assessments. Government chatbots powered by either LLMs or other NLP techniques could also assist citizens in learning about, navigating, and accessing government services.

Cons:

Without additional augmentation, LLMs are unable to provide citations or sources for the claims they make. LLMs may generate factually incorrect or misleading information, known as 'hallucinations', which could lead to inaccurate advice and misinformed decisions. They may also produce biased or inappropriate content as a result of being trained on data containing biases or inaccuracies.

Data privacy and security concerns arise when handling sensitive information for training and fine-tuning LLMs. If the data used to train, contextualize, or fine-tune LLMs contains sensitive information like secret documents or personally identifiable information, the model may leak information to the end user. As such, implementing teams should ensure that private or confidential information is not fed into training data or into chatbot prompts (i.e. by requesting a LLM to summarize a confidential document). Overreliance on LLMs by government employees may result in reduced human intervention and critical thinking, leading to

potentially dangerous oversights or misinterpretations in complex situations. In addition to hampering government operations, this could lead to the spread of misinformation from government sources and erode public trust in government communications and decisions.

Additional Considerations:

The software products used by government employees are likely to be augmented with LLM-powered features in the near future. Most notably, Microsoft 365 co-pilot is staged to introduce features like email drafting and summarization, meeting summarization, Powerpoint generation, and report drafting into the base product. There is a high chance that features like this will be the first encounter with LLMs for many government employees, and without proper training before LLM features are turned on departments face a higher risk of improper use of LLMs by employees. Ultimately individuals will be responsible for their use (or misuse) of these AI technologies in the same way that they are for the use of any technological tool, and will require appropriate education and support as they become more prevalent in their daily work.

Environmental Scan #2: Governance Approaches to AI and Automation in Government

The following environmental scan provides examples of internal governance approaches for AI and automation that are being used by various government bodies, public sector institutions, and related communities of practice, to govern their own current and future use cases. Taken together, these case studies offer input and experience on a broad range of governance considerations including project team composition, the application of fundamental human rights and digital government standards, and quality assurance practices. Readers should consult the source material where available for additional details of these considerations and the specifics of each example.

SUMMARY TABLE: Governance Approaches to AI and Automation in Government

Case	Country	Organization	Approach
Directive on Automated Decision-Making	Canada	Treasury Board of Canada Secretariat	Governance framework for public-facing and internal use cases (2023); Impact assessment tool.
Framework for Responsible Machine Learning Processes	Canada	Statistics Canada	Peer review framework.
Ethics Guidelines for Trustworthy AI	European Union	European Commission	Non-binding ethical framework that aligns with related legal obligations.

Case	Country	Organization	Approach
Algorithmic Charter	New Zealand	Statistics New Zealand	Non-binding governance framework and risk identification framework.
A guide to using artificial intelligence in the public sector	The United Kingdom	Central Digital and Data Office, and Office for Artificial Intelligence	Project implementation and monitoring guidance.
Federal AI Community of Practice	United States of America	General Services Administration	Community of practice; stakeholder engagement strategy.
Better Practice Guide for Automated Decision-Making	Australia	Office of the Ombudsman	Best practices and peer review.
AI Risk Management Framework	United States of America	National Institute of Standards and Technology (NIST)	Normative risk assessment and monitoring framework that includes customizable actions, references, and documentation guidance.
Kratt	Estonia	Ministry of Economic Affairs and Communications	Technology-neutral approach to regulating government use of AI in line with existing legal obligations.

Directive on Automated Decision-Making, Treasury Board of Canada Secretariat

In April 2019, in concert with the release of the new Policy on Service and Digital, Canada's Treasury Board Secretariat (TBS) also put into effect their Directive on Automated Decision-Making (the Directive) and its accompanying Algorithmic Impact Assessment (AIA) tool. Together, the Directive and AIA provide a robust governance framework for the adoption and implementation of Automated Decision-Making (ADM) systems and algorithms by federal public sector organizations. The Directive was designed to be a proactive policy approach to ADM, fully intended to minimize "legal liability and *public-facing risks*" for deploying entities.[115] More specifically, the Directive is described as a "mandatory policy instrument" for Automated Decision-Making systems that provide services offered by the government to individuals and organizations. While initially focused on public-facing, external services, notably this Directive was updated in April 2023 to modify its provisions to apply to any use of Automated Decision-Making systems used to make an administrative decision or a related assessment about a client, even if internally facing in nature (i.e. administrative decisions about public servants).

The purpose of the Directive is to apply the Canadian Charter of Rights and Freedoms and core administrative law principles like transparency, accountability, legality, and procedural fairness to digital solutions that leverage ADM processes. The Directive creates government-wide standards and a consistent approach to risk management in AI, which in TBS's own words would "ensure that automated decision systems are deployed in a manner that reduce risks to clients, federal institutions and Canadian society and leads to more efficient, accurate, consistent and interpretable decisions made pursuant to Canadian law."[116] TBS also notes that because these technologies and their environment change rapidly, the Directive will evolve and be reviewed every two years and as determined by the Chief Information Officer of Canada.[117]

The Directive makes it so institutions take necessary "early action" to mitigate risks associated with ADM systems, with particular emphasis given to risks associated with bias (quality assurance) and lack of transparency. The accompanying AIA tool is used to evaluate the potential impact of these risks on citizens and provides granular, risk-based assessment and intervention guidelines for project teams wanting to prevent and/or mitigate risks where they are typically highest. Moreover, the Directive requires that deploying institutions first complete an AIA *before* production and/or when system functions or scope change, as well as on a scheduled basis.[118]

The AIA calculates the "impact level" of an ADM system based on responses to risk and risk mitigation questions across 8 areas of interest. The scope of the AIA questions includes risks related to proposed algorithms, the nature of the decision context, the origin and type of data used by the system, and fundamental risk mitigation strategies like consultation and data quality assurance. Impact levels are therefore assessed according to a broad range of critical areas, including: the rights, health, and well-being

of individuals or communities, their economic interests, and the ongoing sustainability of an ecosystem.[119] In effect, the AIA scores each system and assigns an overall impact level ranging from Level I (little impact) to Level IV (very high impact). According to the TBS, this categorization scheme distinguishes impact based on the criteria of "reversibility and expedited duration," insofar as low impact systems are reversible and temporary, and high impacts are relatively irreversible and perpetual.[120] Naturally, the impact level assigned to a system determines the mitigation measures required under the Directive to reduce identified risk.

There are specific parts of the Directive and the AIA worth emphasizing. To target risk associated with algorithmic bias, the Directive requires testing before production, processes for testing data and the models for unintended bias, and ongoing system and outcome monitoring/evaluation on a scheduled basis. One particularly notable part of the Directive's quality assurance measures is its peer review requirement by a third party that helps validate the AIA process and its results. Qualified, relevant third parties provide essential "checks and balances" on the appropriateness of deployment, quality assurance, and risk mitigation measures, while identifying the residual risk an institution will have to accept as part of their operations.

Systems at impact levels III-IV are required by the Directive to publish in plain language the results of any peer review or audit of their system, as well as a description of how their system works and how it supports their decision-making. To ensure a standard of sufficient transparency, the AIA assesses an institution's "proactive disclosures" about how and where their algorithms are being used.[121]

Participating institutions are required by the Directive to publish their AIA results on Canada's Open Government Portal, which serves as a registry of active ADM systems in Canada. Past AIAs made accessible on the portal include Public Health's ArriveCAN Proof of Vaccination Recognition, IRCC's Advanced Analytics Triage (TRVAs), Veterans Affairs' Mental Health Benefit, and many others of note. Because public clients of these services are not likely aware of the AIA or the Open Government Portal, the Directive also asserts that notices of automation must be provided to clients through all service delivery channels (Internet, in person, mail, telephone). Notably, there have been some incidents and concerns around compliance with the Directive since it came into force.[122] For example, the Department of National Defence (DND) in 2021 was reported to have used AI as decision-support in a hiring context without completing an AIA or privacy impact assessment even though they were likely required to do so under TBS policy. While the Directive on ADM applies to most Government of Canada departments and agencies, of note it excludes various "Agents of Parliament" – such as the Office of the Auditor General, the Office of the Privacy Commissioner of Canada, and the Office of the Information Commissioner of Canada – the Canadian Revenue Agency, as well as National Security Systems.[123] It also does not apply to provincial and municipal governments in Canada.

Framework for Responsible Machine Learning Processes, Statistics Canada

Statistics Canada (StatCan) has tailored their own framework for the use of AI because their use cases thus far haven fallen distinctly outside the existing legal and procedural purview of the Treasury Board Secretariat's Directive on Automated Decision-Making (the Directive). StatCan projects that have used Machine Learning (ML) or modelling were "part of a statistical program that does not aim to make administrative decisions about a client."[124]

The StatCan Framework consists of a set of guidelines for internal research and data creation, and an accompanying checklist – adherence to which is the responsibility of a given project's manager.[125] As an example of voluntary self-evaluation, the Framework prioritizes an ethical approach to the responsible use of algorithms and ML, instead of a strict or binding regulatory/legal approach.[126] The Framework supports the agency's vision to "create a modern workplace culture and to provide direction and support to those using [ML] techniques."[127] The agency states that the framework can be applied to statistical programs and projects involving ML within StatCan or by other adopting organizations.[128]

These guidelines are organized into four themes: Respect for People; Respect for Data; Sound Methods; and Sound Application. The first two themes instill the human-centeredness of the agency's framework, and represent the application of more abstract, ethical governance principles like accountability, fairness, privacy, and confidentiality. The second set of guidelines are more process-oriented and attempt to control for transparency, reproducibility, reliability, and explainability of ML models and experimental results. The themes encapsulated in the Framework were designed to be used in concert with the agency's pre-existing Quality Guidelines and Proportionality Framework. The relevance of the Framework rests on the agency's assumption that *"good practices for documentation, quality assurance and performance measurement reporting will also be followed, without specific instruction from these Guidelines."*[129]

The extent to which a project's ML methods fulfil Framework requirements is determined through a self-evaluation checklist and peer-review. A three-step review process was designed for all projects using ML methods to produce official statistics, and to assist the "ML practitioner" in assessing their methods.[130] In the first step, the development team fills out a questionnaire that assesses the project's adherence to the four guideline themes (mentioned above). Together with project documentation and methodology, the complete questionnaire is forwarded to SC's in-house "review team," where the next step of the review process begins. One portion of the questionnaire is reviewed by the agency's Data Ethics Secretariat team, while the other is reviewed by a team from their Data Science Methods and Quality Section. Following this assessment, the review team sends the project manager a report with recommendations.

The final step of the review process requires presenting the project to the Modern Statistical Methods and Data Science Branch's “Scientific Review Committee”.[131] The presentation is an opportunity for the project team to explain their ML processes before an expert panel who, in turn, can challenge and scrutinize proposed methods, identify potential gaps or limitations, and recommend improvements or corrections. The agency says the Committee can ultimately recommend “whether or not to implement” a project intended to produce official stats.

SC is careful to note that their Framework will need to be “frequently adapted and revised” to accommodate new data sources and ML methods on the one hand, and “emerging issues of ethics and quality” on the other.[132] Lastly, in an effort to boost transparency, the agency also assures that the TBS Directive will be applied to future ML use cases that qualify. As of 2022, they report that they are in the process of establishing a public register of projects – a portal or dashboard – that have gone through their review process. In particular, the dashboard would aggregate and report all checklist responses from previously reviewed projects at any level, for the purposes of “internal management of resources and quality assurance.”

Ethics Guidelines for Trustworthy AI, European Commission

In April 2019, the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) produced their Ethics Guidelines for Trustworthy AI, providing guidance on how to design and implement AI systems in an ethical and trustworthy way. The *Guidelines* were designed to address the fact that AI and ML in particular, pose new types of ethical concerns compared to previous technological innovations. With the caveat that AI systems are highly context-specific, the Guidelines apply to AI systems generally and across sectors. The document is exhaustive, but the authors emphasize the importance of customization for each situation, particularly for sensitive AI applications where risk is typically higher. As such, the AI HLEG stresses that the *Guidelines* should be implemented as a "horizontal foundation" for Trustworthy AI, which may need to be adapted to contexts and applications. Moreover, the authors encourage exploring sector-specific approaches to complement their framework.

The philosophical backbone of the *Guidelines* are four ethical imperatives, rooted in the EU Charter, that when respected ensure that "AI systems should improve individual and collective wellbeing," and that systems are "developed, deployed and used in a trustworthy manner."[133] The four principles are described as 'imperatives' because they frame what is effectively a non-binding, normative tool that AI practitioners *should* adhere to that "goes beyond formal compliance with existing laws."[134] Currently, there are no sanctions, punishments, or formal disincentives for non-compliance; however, the document claims that the four imperatives are reflected in other legal requirements with mandatory compliance like the GDPR and EU consumer protection regulations, and so help furnish "lawful AI" as such.[135] To ensure AI solutions adhere closely to fundamental rights and ethical norms, the document recommends adopting risk mitigation strategies to address probable gaps between abstract ethical principles and their application.[136] The ethical principles are included below, as follows:

1. **Respect for human autonomy**
2. **Prevention of harm**
3. **Fairness**
4. **Explicability**

Based on the above principles, for the purpose of implementing and realizing trustworthy AI, the *Guidelines* then offer seven additional requirements that AI systems should meet in order to be considered "trustworthy". They are as follows:

1. Human agency and oversight
2. Technical robustness and safely
3. Privacy and data governance

4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental wellbeing
7. Accountability

The language and spirit of the requirements is common across other national and international AI strategies and governance frameworks. The document notes that these kinds of ethical frameworks draw from fundamental rights because, as socio-technical environments evolve, ethical frameworks can adapt dynamically and re-interpret these rights to "inspire new and specific regulatory instruments."[137] The authors suggest that although each requirement is of equal importance, "context and potential tensions between them" need to be taken into account when deploying different systems and use cases across various domains.[138] Adopting organizations are therefore encouraged to give special attention to requirements that mitigate risks that directly or indirectly affect individuals. That said, requirement six ('Societal and environmental wellbeing') stands out insofar as system sustainability is explicitly considered in terms of the environmental, social, and democratic impact of AI. The *Guidelines* uniquely highlight the socio-technical embeddedness of AI systems and the need for designers to align with the principles of fairness and harm prevention by considering broader society, "other sentient beings," and the environment as relevant stakeholders.[139] As a set of considerations, these are often left to be inferred by deploying institutions, or can even be conspicuously absent from other national AI frameworks/strategies.

To help organizations meet these requirements, the *Guidelines* suggest technical methods (e.g. systems architecture and explainability), and non-technical methods (e.g. stakeholder participation, codes of conduct).[140] The AI HLEG further recommends that requirements be continuously evaluated and addressed throughout an AI system's life cycle. Again, these assessment requirements are voluntary and so ultimately take the form of best practices or guidelines, and because they cannot be legally enforced, they are consistent with the "ethical AI" approach to governance. However, between 2018-19, the AI HLEG suggested organizations pilot the accompanying Assessment List for Trustworthy AI (ALTAI) and provide feedback on whether it effectively operationalizes the *Guidelines*' requirements. Based on the feedback received from 350 organizations, the AI HLEG presented the final version of the ALTAI in July 2020, which serves as an "accessible and dynamic (self-assessment) checklist" that can be used by developers and deployers of AI wanting to implement the key requirements in practice.[141] In a 2021 explanatory memorandum, the Commission stated that this piloting phase has provided the "proposed minimum requirements" for the eventually enforceable, yet still pending, EU AI Act. For now, the most up-to-date ALTAI is available as a prototype web based tool and PDF.

Algorithmic Charter, New Zealand

The *New Zealand Algorithmic Charter* is a one-page agreement published by Statistics New Zealand (Stats NZ) for use by government agencies to acknowledge and support their commitment to the development and administration of safe algorithms that "reflect the principles of the Treaty of Waitangi."[142] The Charter requires signing agencies to inventory algorithms and assess their risk by using a simple three-by-three risk matrix that can be used to categorize algorithms into low, moderate, and high-risk statuses. Moderate and high-risk algorithms are subject to the commitments outlined in the Charter and should be managed according to their risk level, with mitigation resources allocated to the highest-risk algorithms first. The six commitments outlined in the charter include:

- to "**maintain transparency** by clearly explaining how decisions are informed by algorithms.";
- to "deliver clear public benefit through Treaty commitments by […] embedding a **Te Ao Māori perspective** in the development and use of algorithms";
- to "**identity and actively engage** with people, communities and groups who have an interest in algorithms, and consulting with those impacted by their use.";
- to "make sure **data is fit for purpose** by […] understanding its limitations [and] identifying and managing bias.";
- to "ensure that **privacy, ethics and human rights** are safeguarded by […] regularly peer reviewing algorithms to assess for unintended consequences and act[ing] on this information.";
- to "**retain human oversight** by […] nominating a point of contact for public inquiries about algorithms, providing a channel for challenging or appealing decisions informed by algorithms, [and] clearly explaining the role of humans in decisions informed by algorithms."

Each agency's Chief Executive, Chief Privacy Officer, and Senior Manager responsible for algorithms are required to sign their own copy of the Charter, which is hosted on a publicly accessible page on the agency's website. Once signed, Charters are typically supported by agency-specific policies developed to operationalize Charter commitments and localize risk management practices. For example, the Ministry of Business, Innovation and Employment (MBIE) hosts a copy of the Charter on its website, outlining how algorithms are used at MBIE, and how the Charter commitments apply to the Ministry's operations.[143] On the same page, the Ministry notes that as part of their commitments, they have developed an algorithms use policy and "established a Data Science Review Board to provide MBIE with strategic and practical direction, guidance and leadership for matters relating to data science and algorithm governance."

The Charter is careful not to provide a specific definition of what an algorithm is, but does note that anything from unsophisticated workplace automation to predictive algorithms, like regression models or neural networks, could be considered an algorithm. Moreover, exact definitions of algorithms differ between agencies. For example, MBIE

defines an algorithm as “an automatic process which can identify patterns in data to assess criteria or predict outcomes”, while The New Zealand Ministry of Health (HNZ) defines an algorithm as “an automated tool for operational decision-making that has little or no oversight by an individual.” Even in circumstances where the Charter may not apply, agencies still sign and acknowledge the Charter on their website. For example, HNZ has a dedicated webpage for the Charter but claims that, according to its own definition, it does not currently deploy any algorithms. While HNZ acknowledges that hospital administrators use Ministry-developed decision-support tools like the Cardiovascular Disease Risk Assessment tool, and the National Priority Interface for day-to-day operations, it claims the decisions made using these tools are “ultimately the responsibility of clinicians.”[144] Even so, HNZ has adopted the principles of the Charter to its specific context by issuing guidance on how clinicians and other actors in the health sector should develop and manage safe algorithms.[145]

A Guide to Using Artificial Intelligence in the Public Sector, United Kingdom

A collaborative effort between the Government Digital Services (GDS) and the Office for Artificial Intelligence (OAI) in the UK resulted in a shared recommendation for implementing AI in the public sector, entitled "A guide to using artificial intelligence in the public sector." The guide is an operational reference for implementation teams, consisting of case studies and specific advice on how to safely plan, execute, and monitor AI projects. The document draws heavily from existing service design, procurement, and risk management practices deployed by digital teams across the government, and so highlights the importance of adhering to basic service design, system architecture, and data management principles.

The first section of the guide helps teams decide whether AI will help solve their users' problems, and suggests frameworks for determining what ML techniques or applications to use. Project teams are encouraged to consider key implementation questions like where data exists to train a model, whether there is enough data to reliably train it, whether it is ethical to use this data, and whether the tasks being automated are repetitive enough that a human would be incapable of carrying it out within a reasonable timeframe. The authors note that in addition to requiring large quantities of data, high-quality data is also necessary for training safe and reliable models. Implementation teams are advised to consider the "accuracy, completeness, uniqueness, timeliness, validity, sufficiency, relevancy, representativeness, [and] consistency" of potential data sources before they train their models to proactively avoid bias and improve performance.

The planning section of the guide provides detailed, practical advice on implementation strategy. Teams are encouraged to have a clear plan for the discovery, alpha, and beta phases of a given AI project. According to the guide, the discovery phase should involve identifying the problem and researching the existing data and processes relevant to a proposed system. The alpha phase involves prototyping and testing the AI model and service, splitting the data, building a baseline model, and evaluating the model's performance. The beta phase involves integrating the model into the service, evaluating the model, and helping users understand the model's outputs. Throughout the beta phase, teams are advised to iterate and deploy improved models and continuously evaluate the model's performance to ensure it meets business objectives and user needs.

The planning section also recommends that AI project teams be diverse and cross-functional, consisting of experts from fields like data science, data engineering, ethics, and service design. It also suggests including subject matter experts (e.g. social care, agriculture, government procurement) who have a deep understanding of the environment in which the model will be deployed, to ensure that the systems developed will properly serve the people they are intended to. The guide's detailed advice regarding

project delivery positions it as a unique tactical reference for digital teams, but it also provides extensive advice on the governance of AI systems.

The guide dedicates a section to AI ethics and safety practices to mitigate accidental harm caused by the misuse of models, poor model design, or unintended consequences of algorithms. According to the guide, ethical challenges do not result from the decision to use an AI model, but from *how* an AI model is used and in what context it is deployed. For example, systems used for spam email filtering or basic automation do not present the same moral or ethical risks as systems used for prison sentencing or the identification of vulnerable children. Because of the varying potential for harm across AI use cases, the guidance provides a general approach for ensuring all models, regardless of their use and context, are governed by basic processes, principles, and ethical values.

The guide recommends implementing process-based governance frameworks at the outset of projects to guarantee model auditability and to establish time frames for the regular evaluation and monitoring of models. These types of frameworks can help project teams adhere to relevant ethical, legal, public trust, and risk management principles. The guidance suggests responsibility should be assigned to teams and managers according to predetermined potential points of failure such as the data used to train the model, the code, the model selected, or the rollout out of the system. For clarity, the authors recommend keeping a responsibility record that establishes what team or individual is responsible for the different aspects of an AI system to ensure proper accountability and governance.

Federal AI Community of Practice, USA

In 2019, the General Services Administration (GSA) and the Federal Chief Information Officer (CIO) in the United States launched a Federal AI Community of Practice (AI CoP) to harness advancements in this field, transform their services, and drive the "thoughtful adoption of AI across the federal government."[146] The CoP emerged against the backdrop of the President's signing of the Executive Order 13859 (*Maintaining American Leadership in Artificial Intelligence*) that establishes, among other things, the significant role the government will play in initiating, facilitating, and protecting AI innovation and adoption across the United States. This top-down initiative effectively implemented a "government-wide strategy" of formal and informal policy approaches for building momentum around technologies that could have whole-of-government impact.

In this context, the objective of the AI CoP was to drive the adoption of AI and ML technologies within the government. As per the Centers for Excellence (CoE), part of the GSA's Technology Transformation Services, the AI CoP "regularly organizes and runs events for government-wide audiences to share opportunities and challenges with the responsible deployment of AI in the federal government, promoting key AI case studies and showcasing best practices." As such, the AI CoP represents one of the informal, normative approaches a government can take to achieve this goal by building a "knowledge base and inter-agency forum on best practices, tools and resources that enable the federal workforce to responsibly deploy [AI and ML]."[147]

Currently, the CoP unites 1200 federal employees (members) across 60 agencies who are "active or interested in AI policy, technology, standards, and programs."[148] Any federal employee or "mission-supporting contractor" can become a member of the AI CoP to gain access to, and participate in, AI relevant community meetings, working groups, virtual trainings and events. Each of these informal practices create open channels through which typically siloed individuals can share tools, playbooks, and challenges with a voluntary community of interested professionals across different agency contexts and objectives. Notably, working groups exist around specific topics like Privacy and AI, Computer Vision, deep learning, RPA, and Natural Language Processing, empowering employees from different agencies to collaboratively design and develop products and frameworks needed to support responsible, trustworthy AI processes.

In January 2022, the Privacy and AI working group published an Artificial Intelligence Governance Toolkit to support agency leaders and privacy practitioners by establishing "a unique, comprehensive approach to data privacy" aligned with the Executive Order on Diversity, Equity, Inclusion and Accessibility in the Federal Workforce (2021).[149] The Toolkit was greatly informed by the Government Accountability Office's Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities, and was foremost designed to mitigate the potential risks of irresponsible AI use. The group explicitly states that the Toolkit should not function as guidance or a checklist, but as a

"set of considerations to help determine the best way [for a federal agency] to approach AI."[150]

The Toolkit emphasizes an approach to governance that prioritizes privacy and stakeholder engagement with subject matter experts ranging from data science, software development, and UX, to civil rights and liberties, privacy and security, and legal counsel. For the most part, the Toolkit is best used by agencies to establish a stakeholder engagement strategy and Development Life Cycle so that, regardless of the application, their AI processes leverage the relevant actors and resources required to operationalize the US Privacy Act's Fair Information Practice Principles (FIPPs). Despite their initial claim that the Toolkit is *not* a checklist, it does include checklists with criteria designed to help agency leaders identify stakeholders and policy artifacts relevant to each stage of AI development and deployment (Problem Identification; Data Gathering; Algorithm Creation & Testing; Deployment).[151] At each step, the checklists touch on key principles for the responsible use of AI such as explainability, privacy oriented consultation, iterative documentation of changes to an algorithm, and bias/discrimination evaluation.

Better Practice Guide for Automated Decision-Making, Australia

The Australian Ombudsman's Better Practice Guide for Automated Decision-Making helps public sector organizations build compliant, customer-centric ADM systems. The guide provides detailed advice on a broad set of considerations for ADM systems teams, including compliance with administrative law and privacy legislation, effective governance and system design, and the deployment and continuous monitoring of automated systems. The guide's advice culminates in a simple checklist for implementation teams to consider during the life cycle of their project, from planning to implementation and monitoring.

Two sets of principles guide the development of Australian ADM systems: the OECD AI principles, and the AI Ethics Framework for Australia. The OECD AI principles advise institutions developing AI systems to respect the rule of law and democratic values, and promote sustainable growth and social justice while ensuring secure and transparent operation. The AI Ethics Framework for Australia, published by the Department of Industry, Innovation, and Science in 2019, shares significant overlap with the OECD principles. The framework emphasizes human-centered design and values, privacy and security, and the need for human oversight and accountability for AI systems.

One common and fundamental limitation of ADM systems is that they compress real-world complexity into just a few parameters when making or supporting decisions. Because they do not have complete information, they cannot make the full range of possible suggestions. To mitigate this "blind spot" risk, the Australian *Better Practice Guide* suggests that systems supporting discretionary decision-making, should not unduly limit the options available to decision-makers; and that decision-makers should be made aware that final decisions are up to them and not the system. This distinction clarifies that algorithms are not actors, but tools built to support humans. In circumstances where a human overrules a decision made or suggested by an ADM system, the guide states that the system should collect and store the decision-makers justification for intervention. By designing for human feedback from the start, implementing teams can improve system auditability, more precisely monitor system performance, and understand any shortcomings.

Whether a decision was made by a human or an algorithm, the guide says that a human, usually a department Secretary or division head, must take full responsibility. This mechanism ensures that leaders are invested in the proper operation of an ADM system, and that there is no confusion as to who an individual should contact in the event of an incorrect decision. The guide recommends that individuals who are affected by these types of decisions be presented with a breakdown of why and how they were made, under what authority they were made, and who was ultimately responsible for the decisions. Additionally, it is suggested that those who are negatively impacted should be able to contest the decision in a timely fashion, and that implementation teams should develop these processes before the deployment of any automated system.

Automated systems can become unwieldy, spanning multiple departments, information systems, and even organizations in cases where third-party systems are used, or when data or systems are shared between multiple government agencies. Without proper documentation, it is nearly impossible for a single individual to have a complete understanding of the system, let alone identify who is responsible for each underlying process. To remedy this risk, the guide advises teams to document and map the business rules and processes that underlie ADM system, and ensure that each rule and process have a basis in legislation and policy objectives. Importantly, the principles embodied by the *Better Practice Guide* also apply in circumstances where the system is being procured from a third party, ensuring that so long as responsibility has been assigned internally, risks emanating from the use of external systems can be identified and managed. By properly documenting systems, the guide says implementation teams can better track dependencies between systems and improve business processes over time.

In the name of transparency, accountability, and administrative lawfulness, the authors believe peer review is a strong mechanism for encouraging robust system design and the identification of critical faults. They suggest that agencies publicly share automated business processes that support or make decisions, and publish internal research conducted on system performance. By exposing internal systems to external criticisms, agencies can test their own assessments of how well processes comply with legislation, policy objectives, and the expectations of civil society actors.

AI Risk Management Framework, U.S. Department of Commerce's National Institute of Standards and Technology (NIST)

The AI Risk Management Framework (AI RMF) was designed to help organizations operationalize risk mitigation in contexts where AI regulation and laws may be lacking or under construction.[152] As a product of the U.S. National Artificial Intelligence Initiative Act (2020), the AI RMF represents an international benchmark for approaches to organizational AI risk management, and offers expert guidance for the future development of AI governance frameworks. The document prioritizes risk management as a key component of the responsible development and use of AI systems that can be deployed in varying degrees and according to local capacity. Responsible practices ubiquitous across AI frameworks and strategies help align development, design, and use case decisions with policy objectives. However, the AI RMF centers around risk management because its processes actualize these responsible practices and principles, by prompting organizations and their teams "to think more critically about context and potential or unexpected negative and positive impacts" inherently involved in any AI application.[153]

The AI RMF is divided into two parts. The first discusses how organizations can frame AI risk and outlines the characteristics of trustworthy AI systems, with the understanding that building and deploying trustworthy AI is a necessary part, if not its own form, of risk management. The framework states that any comprehensive approach to risk management calls for balancing trade-offs among trustworthiness characteristics like interpretability and privacy, or predictive accuracy and interpretability. To navigate the "existence and extent of trade-offs between different measures," the AI RMF places emphasis on approaches "enhancing contextual awareness in the AI life cycle." This essentially involves consulting a diverse set of AI actors and incorporating their perspectives to better understand and manage complex risks arising in social contexts.[154]

The AI RMF adapts ISO/IEC TS 5723:2022 to determine seven characteristics that may be used to evaluate an AI system's trustworthiness:

1. Valid and Reliable
2. Safe
3. Secure and Resilient
4. Accountable and Transparent
5. Explainable and Interpretable
6. Privacy-Enhanced
7. Fair (with Harmful Bias Managed)

The framework also provides a helpful guide to the challenges associated with measuring AI risk. Notably it states that the "inability to appropriately measure AI risks does not imply that an AI system necessarily poses either a high or low risk."[155] Other challenges highlighted in the guide include risks related to third-party inputs; the lack of consensus on robust and verifiable risk metrics; the mutability and emergent quality of risk at different stages of an AI system's life cycle; differences between risk in a real-world setting versus risk in testing environments; inscrutability at various layers of a system; and the difficulty involved in systematizing a baseline metric for human decision-making versus that of ADM systems.[156]

The second part of the AI RMF is the "core" of the framework, and includes four high-level functions that the agency claims should help organizations practically address AI system risks. Comprising what NIST calls its AI RMF Playbook, the following four functions are designed for application and include relevant actions, references, and documentation guidance to achieve their outcomes.

GOVERN

The most central is the Govern function, which is described as "cross-cutting" because it applies to all stages of an organization's risk management processes and procedures.[157] While the three other functions can work in all AI system-specific contexts and at various stages in an AI life cycle, aspects of the Govern function – especially compliance and evaluation – facilitate the operations and goals of the other functions. In a nutshell, the governance function is a confluence of policies, accountability structures, and processes that create a culture of risk-aware planning and AI systems development. Specific actions may include the introduction of procedures for safely decommissioning AI systems; assembling diverse teams to measure and manage AI risk; practices for engaging with AI actors and incorporating their feedback into the development and risk management processes; and policies that address and mitigate risks stemming from external third-party software or data providers.

MAP

The mapping exercise is fundamentally about understanding the context in which a system will be deployed and its associated risks. Here, risks may materialize as events, but they are the result of processes that unfold over time and across systems. Therefore, a more complete understanding of the systems that contribute to AI risk is necessary for actively managing and mitigating them. The Map function of the framework allows teams to proactively identify and categorize sources of risk by providing them with a method for tracking interdependencies between systems and actors across the AI life cycle. At this stage, organizations should estimate the likelihood and magnitude of each identified risk. Estimations can be informed by leveraging publicly available data about scenarios in

which similar AI systems have failed, and by engaging a diverse set of actors. The framework suggests that by engaging AI actors like end users, external experts, and communities of interest, organizations can define application contexts more precisely and better understand the limitations, risks, and opportunities of proposed AI systems. These mapping exercises should equip organizations with enough information about the risks, rewards, complexities, and end users of a given AI system to decide whether or not a project is viable, responsible, or even necessary.

MEASURE

The measure function of the framework advises organizations to systematically measure the risks identified during mapping exercises, and to document those that cannot be measured. While measurements are imperfect, they can be useful in assessing trade-offs between different tenants of trustworthy AI systems (listed above). Moreover, the framework advises organizations to test and measure risks repeatedly over a system's lifetime. Wherever possible, concrete measurements of uncertainty are recommended. By testing the compounding error resulting from the interaction of two or more systems, organizations can measure and catalogue emergent risks. Framework users are advised to evaluate the efficacy of measurements and continually update them as they learn more about how risks materialize in development and deployment. To minimize the risk of assessment bias, organizations are encouraged to seek input from independent or third-party assessors.

MANAGE

Once risks have been mapped and measured, organizations must manage them by deploying resources to mitigate them and/or prepare resources for their probable materialization. This stage involves the development of risk treatment and response strategies that allow organizations to get the most out of AI systems while reducing their potential for harm, and may include incident communication plans and strategies for dealing with third parties.

Kratt, Estonia

As of 2019, the Estonian Government has been developing and iterating on an AI Strategy known as Kratikava in Estonian. The strategy is part of Estonia's ongoing effort to extend requirements around AI and data (especially citizen-centric data governance) beyond the European Union's (EU) General Data Protection Regulation (GDPR) and forthcoming AI Act. Estonia coordinates Kratikava implementation in order to have more control over and boost the usage of AI within the public sector. The overarching goals are to balance bureaucracy with flexibility, and practicality with human centricity and rights.

GDPR

The GDPR is a strict privacy and security law enforced by the EU. All member states must adhere to the GDPR, as well as any country that collects data from people in the EU. The GDPR was created in order to protect the European Convention on Human Rights: "Everyone has the right to respect for his private and family life, his home and his correspondence." It covers a vast array of topics relevant to privacy and security, but lists a "top 5" areas of concern: personal data, data processing, data subject, data controller, and data processor.

The key regulatory points of the GDPR are as outlined below:

Data protection principles:

1. Lawfulness, fairness, and transparency: processing must be fair to the data subject
2. Purpose limitation: you may only process data for the purpose explicitly agreed to by the subject at the time of collection
3. Data minimization: you must collect the minimum amount of data possible to complete your goal
4. Accuracy: personal data must be accurate and kept up to date
5. Storage limitation: Personally identifying data may only be stored for a long as necessary
6. Integrity and confidentiality: processing must be done in a way that preserves the security, integrity, and confidentiality of the data (ex. via encryption)

7. Accountability: the data controller is able to demonstrate GDPR compliance for all regulations at all stages of data processing

Data security and protection:

- Projects requiring data processing must implement appropriate technical and organizational measures
- Everything done within the organization must consider and implement data protection by design and by default

Data processing principles - a list of instances when you may process personal data:

1. The subject gave you unambiguous consent to process their data
2. Data processing is necessary to enter into a contract of which the subject is a party
3. You may process the data if it is necessary to comply with a legal obligation
4. You may process the data if it is necessary to save a life
5. Processing is necessary to carry out a task in the public interest
6. You have a legitimate interest to process a subject's data

AI Strategy (a.k.a. KRATIKAVA)

Kratikava was initially released in 2019, and a revised version of the plan was released in 2022. The three main goals of the first AI Strategy were to advance the uptake and use of AI in the private and public sector, to promote upskilling and research around AI, and to "develop the legal environment." Despite this, the first iteration of Kratikava did not contain any plans for specifically changing the legal environment: "There is no need for fundamental changes to the basics of the legal system, but there are some changes in different laws to be made." The laws that needed to be revised were not specifically identified by the original strategy but there was a separate legal analysis carried out to explore if AI should be considered to be a separate legal entity.

The updated version of Kratikava expands on these principles and highlights five actionable items for developing the legal environment:

1. Development of the draft Act amending the Administrative Procedure Act

2. Participation in the negotiation of the Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and advocacy of Estonia's views
3. Participation in the development of civil liability rules for AI and the digital era in the EU, including participation in the public consultation and participation in the negotiation of a future EU legislative initiative and advocacy of Estonia's views
4. Participation in the negotiations of the Convention on Artificial Intelligence of the Council of Europe and advocacy of Estonia's views
5. Participation in policy and legislative development in the field of AI at the EU and other international levels.

These items were created with the intent of solving and regulating specific issues that need to and can be regulated independent of the EU's guidelines. However, it is important to note that the Estonian government understands and approaches AI as they do any other technology. From a governance perspective, they begin looking at digital services including AI by asking two fundamental questions: what is the desired task/objective accomplished; and what is the data being leveraged (is it authorized)?

Of note, Estonia has also taken a practical and applied approach to questions around the use of data in a responsible manner. This has included the mandatory implementation of their data tracker for all applications, carrying out data protection impact assessments, as well as publishing source code and describing information systems on their registry (riha.ee).

This approach is unique because it evaluates *ex ante* the application context and data, and so emphasizes the importance of discerning before implementation the purpose and rationale for optimizing a process, and what raw material (data) will be used to do so. If the data and the task is in scope, the rest should be covered by the relevant EU laws and policies. This approach to AI governance, and so risk mitigation too, suggests that organizations can control for the nuance AI supposedly introduces into administrative decision-making and service delivery by prioritizing data protection and context sensitivity regardless of the application. It also notably puts a priority on an applied approach informed by practical tools, rather than only relying on legal requirements.

Analysis of Governance Approaches to AI and Automation in Government

In our analysis of the cases presented in the environmental scan of governance approaches to AI and automation in government, several themes and common practices can be identified across AI governance frameworks employed by government agencies around the world. These themes and practices contribute to the responsible and ethical development, deployment, and use of AI and automated systems in the public sector.

Checklists are a popular way to operationalize the principles embodied in governance frameworks. The TBS *Directive*, Stats Canada's *Framework for Responsible Machine Learning Processes*, the European Commissions' *AI HLEG*, the US Federal AI CoP, the Australian *Better Practice Guide*, and NIST's *AI Risk Management Framework* all include a self-evaluation checklist that when followed help administrators, operators, and deployment teams adhere to and implement core guideline principles.

Adherence to principles of ethics and human rights is possibly the most common trait across all frameworks examined. These principles often include fairness, transparency, accountability, and respect for human autonomy. Although compliance is often not technically enforceable, the European Commission's *Ethics Guidelines for Trustworthy AI* recommends centering AI governance around ethical frameworks because they apply beyond formal compliance to existing laws, and because they can adapt dynamically to inevitable evolutions in socio-technical environments to inspire targeted regulatory instruments.

Risk assessment and mitigation processes are crucial components of the AI governance frameworks presented. Several cases, such as the TBD *Directive on Automated Decision-Making* in Canada and the *Better Practice Guide for Automated Decision-Making* in Australia, outline the need for early and continuous risk assessment and management to minimize the potential negative impacts of AI systems on individuals, communities, and the environment. Risk mitigation best practices are perhaps best demonstrated by NIST's *AI Risk Management Framework*.

Taken together, the principles of **transparency and explainability** emphasize how and to what extent implementing organizations can be open and honest about how their algorithms work. All of the above governance frameworks and approaches cover both the technological and administrative parameters of transparently developing and implementing automated systems in the public sector. Fundamental transparency techniques like open sourcing the code and rules used to train the system, and disclosing the nature of data used, are an explicit component of many of the frameworks. Most frameworks also suggest providing clear explanations for how systems operate, how they generate outcomes, and how these outcomes will impact stakeholders. In this

way, transparency and explainability are separate but complementary concepts that are best practiced together: explainability ensures that everyone involved in, or impacted by, an automated system can know and see what is at work in both the machine and the process; while transparency allows stakeholders to identify those responsible for outcomes, and available avenues for recourse.

A key theme in these AI governance frameworks is the need for **human agency, oversight, and accountability.** Emphasis is commonly placed on ensuring human operators understand they are ultimately in charge of system outcomes, and can overrule decisions when required. Governance frameworks usually describe oversight and accountability as the practice of assigning responsibility to individuals or teams for the development, deployment, and monitoring of automated systems, as well as provisioning avenues for affected individuals to challenge, contest, or appeal decisions made or informed by AI. Adequate and legible responsibility can include designated recourse methods and points of contact, following proper documentation of processes and interventions, and ensuring that decisions made by AI systems can be traced back to responsible parties.

Ongoing monitoring and evaluation of AI systems is recommended by many of the frameworks as a best practice for early identification of potential biases, unintended consequences, and areas for improvement. The need for regular review and adaptation of AI governance frameworks to accommodate rapidly changing technologies and environments is also a common theme.

Engaging with stakeholders, including subject matter experts, communities and individuals affected by AI systems, is a common practice across the examined frameworks. This level of engagement actualizes the principles of transparency and explainability and helps ensure that AI systems are developed and implemented in a manner that respects the values, needs, and expectations of those who will be affected by them. For many of the frameworks, successful stakeholder engagement begins internally by establishing a diverse project team consisting of experts from relevant fields, to ensure a comprehensive understanding of the environment where AI systems will be deployed. Interestingly, the EC's *Ethics Guidelines* uniquely expands relevant stakeholders to include broader society, "other sentient beings," and the environment. This nuanced inclusion highlights the need for future implementing agencies to holistically consider the socio-technical embeddedness of their AI systems and automated solutions. And from a risk assessment perspective, NIST believes stakeholder engagement can effectively define application contexts more precisely, so that agencies can preemptively map the limitations, risks, and opportunities of proposed AI systems.

Communities of practice and peer review elicit useful collaboration and knowledge sharing among government agencies, experts, and other stakeholders. Beyond stakeholder engagement, the TBS *Directive*, Stats Canada, and the US

government's Federal AI CoP demonstrate the utility of informal knowledge sharing and best practice development. Peer review systems were identified by many of the frameworks as a key mechanism for encouraging robust system design, through which implementing agencies can test their own assessments of how well processes comply with legislation, policy objectives, and stakeholder expectations. Peer review processes ultimately provide essential "checks and balances" on the appropriateness of deployment, quality assurance, and risk mitigation measures that an agency must know to implement AI systems safely.

The themes and common practices identified across these cases show that there is growing consensus around the responsible and ethical development, deployment, and use of AI and automated systems in the public sector. Agencies curious about how AI can be leveraged to solve public sector challenges should refer to the practices distilled in the above frameworks as essential guidance that, when followed, can help ensure AI systems are aligned with human rights, ethical principles, and societal values.

Risk Considerations

Key Risk Takeaways from Analysis of Use Cases and Governance Frameworks

The case studies examined in this report make clear that, broadly speaking, there is risk associated with the use of artificial intelligence, machine learning, data science, statistical and analytics techniques and technologies (referred to subsequently as AI technologies) by public sector organizations, and that the type and level of risk is quite variable. As noted in the analysis of the two sets of environmental scans, some of the issues that arise during the use of this technology are not necessarily connected to the nature of the technology itself. Contextual and external factors such as lack of quality control and level of public scrutiny also had a substantial impact on the relative success and safety of AI adoption by public sector organizations.

Our analysis also uncovered commonalities across the multiple frameworks for governing the use of AI that were reviewed in the second environmental scan. Transparency, oversight, and engagement with affected communities and stakeholders were common themes. The analysis also substantiated that, in addition to risk reduction, risk mitigation plays an important role in reducing any potential or actual harm incurred through the use of these technologies.

Although the summaries and analysis hinted at underlying factors and mediating strategies, they do not directly address the question of risk. Thus, in developing risk considerations to guide the use of AI by public sector organizations we have examined a range of existing risk frameworks and approaches.

The Treasury Board of Canada Secretariat's (TBS) Directive on Automated Decision-Making, and its accompanying Algorithmic Impact Assessment tool, has already been summarized and analyzed in the preceding environmental scan. However, given its primacy in the discourse around governance approaches for the use of AI technologies by Government of Canada institutions, a further exploration of it is warranted with respect to its applicability in specific contexts and its approach to risk.

TBS Directive Applicability in Specific Contexts

The Algorithmic Impact Assessment (AIA) tool that accompanies the TBS Directive on Automated Decision-Making focuses on impact as a key factor. It evaluates risk through a series of questions and uses this as a basis for assigning impact levels to projects. It considers six areas essential to understanding risk: Project, System, Algorithm, Decision, Impact, and Data.

These six risk areas are defined relative to common aspects of AI technologies and the processes in which they are embedded. Certain combinations of technologies and

processes, depending on the context, can increase the probability of negative outcomes. Through a series of questions, the AIA helps identify the interactions between a given technology and its processes, and assigns it a score that indicates the associated level of risk and recommended mitigation measures as prescribed in the TBS Directive.

As can be seen in the table below, when taking the lens of the four process use case categories of AI that were used to organize the case studies in the environmental scan, while the TBS Directive on Automated Decision-Making provides guidance in some situations where Government of Canada organizations may be using AI and automated tools, there remain use case categories where there is currently no central guidance.

	Decision or assessment about a client (public or internal)	**Support for organizational or policy objectives (not client specific)**
Automated Decision-Making	TBS Directive on Automated Decision-Making applies	N/A*
Automated Decision-Support	TBS Directive on Automated Decision-Making applies	Currently no mandatory central guidance in the Government of Canada**
Detection, Alerts and Notification	TBS Directive on Automated Decision-Making applies	Currently no mandatory central guidance in the Government of Canada**
Procedural Automation and Process Improvement	N/A*	Currently no mandatory central guidance in the Government of Canada**

** Sections marked N/A are considered to be outside the scope of the categories listed*
***Some use cases may be covered by the new Guide on the use of Generative AI published by TBS in September 2023*

Specifically, even with the recent updates to the TBS Directive to clarify that it applies to both external-facing and internal-facing use cases, it only applies when automated decision tools are used to make a decision or assessment about a specific client – either at the individual or organizational level. When AI technologies are being used to support more general organizational or policy objectives (e.g. process automation, or contributing to developing a policy brief for decision-makers) the TBS Directive and its mandatory requirements do not apply. Of note, in September 2023 a new Guide on the use of Generative AI was published by TBS which does provide guidance on some potential use cases in these categories that are not covered by the TBS Directive.

Our suggested approach is to look at a set of lightweight, pragmatic risk considerations to help organizations in the Government of Canada identify and mitigate risks associated

with use cases that currently fall outside the central guidance provided by the TBS Directive on Automated Decision-Making. While our focus for this risk assessment is Government of Canada departments, we believe that it may be useful to public sector institutions more broadly.

Taking a Process Approach to Risk Assessment

Rather than considering risk through the lens of specific AI technology techniques (e.g. machine learning classification, generative AI, RPA, etc.) or domains of application (e.g. health care, law, HR, etc.), we are suggesting a focus on the nature of the processes in which the techniques are embedded. We propose four types of process activities or elements that should be identified and considered from a risk management perspective when AI technologies are involved (note that these element categories correspond with the categories used to organize our case studies in the previous environmental scan of AI use cases):

- Automated decision-making process elements
- Automated decision-support process elements
- Detection, alerts, and notifications process elements
- Procedural automation and process improvement, where the process elements don't involve decision-making, decision-support, or notifications and alerts

Activities in each of the above categories may be connected and carried out by some combination of people and AI technologies. For example, one sequence of events might be: detection of an anomaly through monitoring, which leads to the initiation of an automated process, culminating in a recommendation for action based on automated analysis of data, which in turn, leads to a specific decision made by a government official to change the status of an individual (e.g. rescind their eligibility for a specific government program or benefit). In such a case, the AI technology related risk associated with the process overall would be the risk associated with the highest risk elements of the process in which the technology was involved (in this example, decision support elements). To be specifically clear with respect to the procedural automation and process improvement category, if this type of process is connected directly or indirectly to any of the first three types of process activities, the procedural automation and process improvement element will take on the risk associated with these other process types. This is the case whether or not these processes are carried out by machines or humans.

Defining Risk and Negative Outcomes

Typically recognized types of organizational risk include:

- Strategic
- Reputational

- Compliance
- Legal
- Operational
- Security
- Financial

Some of these types may overlap – for example, a risk may be both a security and reputational risk, depending on the context.

In this discussion, we focus primarily on identifying the nature of risky events that can occur when AI technologies are involved, along with the potential severity of the negative outcome. We view these as the first key steps required for determining the risk level associated with AI technologies. However, our process-based approach also allows for some considerations of the likelihood of negative outcomes within particular contexts independent of the technology itself.

Although specific definitions of risk vary (see Wang and Williams, 2011[158], Zachmann 2014[159] and Hansson 2012[160] for some discussion of this), it is generally recognized that risk involves a preceding event or situation, a consequence or outcome connected to this situation, a degree of chance with respect to whether or not the situation leads to the outcome, and the negative impact of that outcome. ISO Guide 73:2009 defines risk as “effect of uncertainty on objectives”. [161]

Before presenting our considerations for risk assessment, it’s also important to clarify what we believe to be an undesirable event or negative outcome within the context of a Government of Canada organization, given that risk is defined relative to the negative outcomes that an organization seeks to avoid.

Within this context, we would define negative outcomes as including any outcomes that fail to:

- Support the well-being of the Canadian public.
- Reduce harm to members of the Canadian public.
- Allow government departments to carry out their work in as efficient and effective a manner as possible.
- Allow the public to maintain confidence in the government.

Identified Risk Factors in AI-Incorporating Processes

The AI case studies in our environmental scan allowed for the identification of numerous relevant risk factors associated with processes that incorporate AI technologies. As a starting point, the behavior of AI technologies should be both understandable and explainable to ensure transparency in decision-making processes. This also means making the algorithms available for external (e.g., public or third party) scrutiny and auditing. Predictability and consistency are equally important, as knowing what the

technology will do in advance, both broadly and in specific instances, can help identify potential risks and allow for better planning. A technology that is more dynamic and changeable inherently poses more potential risks because it can be repurposed in unanticipated ways, which may lead to unanticipated negative outcomes. For example, suppose a neural net classifier is created in order to identify loan defaulters and is designed such that it will be regularly updated with new data to prevent model drift. It could in this context be possible for the model to be altered by training it on data about individuals who are not loan defaulters but instead merely slow to pay back their loans, and then use the new model for decision making in different contexts, even though this was not the model's originally intended purpose. As another example, the neural net might have originally been intended to assist a person in deciding whether or not to approve loans, but its output might subsequently be easily incorporated into an automated loan application website, with the classification from the neural net being used to approve or disprove loans automatically if the classification meets a certain certainly threshold, instead of being used to assist a person in making the loan decision.

Consistency, accuracy, and precision are also crucial factors in evaluating the performance of AI systems. Replicable results and consistent behavior are essential for maintaining trust in the technology, while accuracy and precision help ensure correct and reliable outcomes. Consistency when the technology is applied is also vital, as it prevents unexpected or undesired results that could occur if the technology has a high level of volatility and is seen to be dynamic and changeable depending on its use case. For example, suppose a hiring classifier is applied to multiple individuals who are applying for a job with equivalent professional qualifications and experience. The technology should make consistent recommendations in this case with respect to issues such as whether or not individuals should be interviewed or how much they should be offered as a starting salary. If the technology instead makes different recommendations for these different individuals despite their essential equivalence, or makes different recommendations for the same individual as a result of minor irrelevant changes in their circumstances, then it is not behaving consistently.

Boundability and reversibility are also important concepts that go hand in hand when assessing AI technologies. It is essential to determine what the scope of the technology's behavior is, in terms of the group being impacted, both in the short and long term. For example, will this technology interact with and subsequently impact some or all small business owners in Canada? Some or all businesses in Canada? Some or all people in Canada? Is it possible to determine in advance the scope of the impact? Is it possible to constrain the impact to a particular well-defined group, both in the short and long term? And, if that behavior has undesirable impacts, can it be reversed or reverted to a pre-implementation state? This is also linked to considerations around repurposeability and how easy it would be for the functionality of a specific AI technology to be expanded to use cases outside of the original intended purpose it was created for.

Supervisability, auditability, and oversight are essential for maintaining transparency and accountability in AI applications. Observing the technology's actions as they occur, as well as auditing them afterward, ensures that AI systems adhere to ethical and legal standards. The role of data also emerges here as a critical factor in assessing risk. Given that training data is the basis of any AI technology, the quality and relevance of

the data that an AI technology is based on, as well as the sensitivity of the data or activity that it is acting on or being applied to, are critical factors for consideration. This can be a particular challenge when training data is not available for observation in a meaningful way.

Lastly, impact and visibility must be considered when incorporating AI technologies into government. The potential or actual magnitude of the impact on those affected by the technology must be evaluated with the potential risk rising along with the increasing impact. Understanding how extensive the capabilities of the technology are, both in terms of range and power, is critical when considering impact. Similarly, the visibility of the actions of the AI technology also has unique reputational risks for public sector organizations.

Many of these risk factors are not unique to processes incorporating AI technologies. For example, if humans carry out a process and make mistakes, this can also lead to negative outcomes. And similarly, if human processes are highly visible, there will likely be more risk associated with them. In some cases, having a process carried out by an AI technology may in fact be less risky – for example, by introducing more consistency or explainability than a similar process carried out by a human. However, it is worth emphasizing that in many cases AI or other digital technologies are operating at a much larger than human scale and speed thus amplifying their impact, positive or negative.

With this in mind, we have identified four factors that we believe represent a uniquely magnified risk when considered in the context of AI technology due to both its relative novelty and potential impact, specifically:

- **Boundability:** To what extent can the behaviour of the technology be successfully constrained to a particular known and well-defined group, both in the short-term and long-term? What is the potential size of this group, again in both the short and long-term?

- **Reversibility:** To what extent can circumstances be reverted to a state the same as that which existed before the technology has been applied (note that this requires some knowledge of the pre-application state)? Can the types of errors or harms potentially produced by the system be undone or are the impacts permanent?

- **Explainability:** To what extent can the behaviour of the technology be understood and explained with respect to technical factors as well as policy decisions (e.g. commercial sensitivities in contracts for AI technology)? To what extent has the technology been explained to impacted groups, and made available for public or third-party scrutiny? Are the results that are produced by the AI replicable under similar circumstances?

- **Visibility:** Is the technology working “behind the scenes” in a manner where errors that are found can be corrected before they impact the public, or is it easy for the presence of the technology to be detected? Is it directly interfacing with external clients or other stakeholders (i.e. via a Chatbot)?

This is not to suggest that other factors such as bias are not important considerations when it comes to the use of AI technologies. However, we would suggest that these are considerations that always need to be addressed in anything that government does. What makes considerations around bias, for example, particularly acute in the case of AI technologies is the impact of the four factors identified above. Bias perpetrated by an individual AI technology that is implemented with low boundability, low reversibility, and low explainability will most likely cause significantly more harm than bias perpetrated at "human scale" by, for example, an individual official.

Taking the individual cases from the environmental scan of government use cases of AI contained in this report, when assessed on the first three criteria listed above – Boundability, Reversibility, and Explainability – we find that there is some consistency of patterns across the four categories of processes that AI technology may be applied to.

		Boundability	Reversibility	Explainability
Automated Decision-Making	Robodebt - Australia	Low	Medium	Low
	At-Home Care Distribution - USA (Arkansas, Idaho)	Low	Low	Low
	Predicting Student Grades - Republic of Ireland and the UK	Low	Low	Low
Automated Decision-Support	Automated Application Triage - Canada (IRCC)	Medium	Medium	Low
	Automating Unemployment Categorization - Poland	Low	Medium	Medium
	Big Data Fraud Detection, SyRI - Netherlands	Low	Medium	Low
Detection, Alerts and Notification	Air Cargo Screening at Pearson Airport - Canada (Transport)	High	High	Medium
	Facial Recognition Technology - Canada (CBSA)	Medium	Medium	Medium
Procedural Automation and Process Improvement	RPA and Social Assistance - Sweden	Medium	High	High
	RPA - New Zealand	Medium	High	High
	Chatbots - Singapore, Microsoft, and Google	Medium	Medium	Medium

Our fourth factor we have identified – visibility – should be considered as a risk multiplier. All things being equal, a relatively low to medium risk use case for AI technology that has high visibility may by virtue of the reputational risks associated with highly visible AI technology projects be considered higher risk than would otherwise be the case.

This suggests a conceptual risk approach for the implementation of AI technologies in government to be applied in cases where existing guidance does not exist, as follows:

(Boundability Risk + Reversibility Risk + Explainability Risk) x Visibility Risk

It is important to note that other factors can, in particular circumstances, overshadow the four factors we have highlighted. For example, if a project has high explainability, high boundability, and high reversibility, but has extremely poor data quality (e.g. extremely biased data), then the high risks associated with this additional factor would overshadow the other factors indicating a high overall implementation risk. As well, although our analysis of the use cases from the environmental scan in this report showed a potential pattern between type of process AI technologies are used for and values associated with these risk factors (e.g. Automated Decision-Making generally has higher risk factors than Process Automation in the case studies examined), this may not always be the case depending on the specifics of the implementation. For example, if an alert system rates high on boundability, high on explainability, and high on reversibility, but nonetheless the alert itself relates to an extremely sensitive context and an alert is issued in error, this could still lead to a serious negative outcome.

Consequently, we recommend that that the risk factors and process categories that we have identified above be used as a starting point for assessing risk for any given process or type of process, with the baseline risk being potentially moved up or down as more details and factors are considered in specific cases.

The Risk of AI Technology Avoidance

To this point we have been considering risks associated with using AI technologies. However, this focus should not mislead those considering using these technologies into thinking that avoidance of these technologies is risk free. Choosing not to use these technologies will also have both direct and indirect consequences and impacts.

This type of consideration is taken for granted in the case of established technologies. For example, people use calculators in order to reduce the risk of human error when carrying out arithmetic. Here, not using the technology is viewed as a high-risk activity. Similarly, avoiding technology that can, if used carefully and appropriately, reduce errors, improve responsiveness and consistency, comes with its own set of risks.

Prominent amongst these is a negative reputational risk for government if it is seen to be lagging other sectors and institutions when it comes to technology adoption and modernization. Government already faces declining levels of trust and a perception of

inferior service delivery compared to the private sector. While this directly impacts the reputation of government, it also has related implications including serving as a barrier to recruitment and retention. In the context of existing barriers to attracting those with high levels of digital-era skill sets into government, being perceived as resistant to adopting or experimenting with AI technologies could serve as an additional impediment.

It is also worth noting that what is often referred to as "Shadow IT" must be considered as part of the risk of avoidance of AI technology. As many of these AI technologies become more widely available – including free or freemium versions – and embedded into existing popular software tools, it will be difficult if not impossible for government as an employer to prevent employees from accessing them. Even if AI technologies are blocked on government networks and devices, employees could still access them on their own personal devices both during and outside of work hours. This puts them in a riskier situation when it comes to safeguards for themselves as well as potentially confidential data that may be exposed to these tools. As a result, it is important that even if government institutions take a cautious approach in leveraging AI technologies directly in their operations, they need to be actively providing education and opportunities for experimentation with their employees to prepare for it increasingly being integrated into the workplace and business processes in the future. Put simply: the AI genie is out of the bottle, and it is not going back in.

Additional Considerations: Large Language Models and Risk

AI technology is developing at an accelerated pace, and it's important that risk considerations take into account novel developments. A good example is the rapid emergence and public accessibility of large language models or LLMs. Recently LLMs have gained particular prominence as a sub-type of Generative AI with the introduction of GPT Models such as ChatGPT which have had explosive growth in usage given the free-to-try tools available and their impressive capabilities.

As noted earlier, LLMs have increasingly captured the attention of policymakers and notably TBS released a new Guide on the use of Generative AI in September 2023. While not introducing any new mandatory requirements, this guide provides considerations on the use of generative AI tools, including LLMs, in the context of existing legal and policy requirements in the federal government.

The approach outlined in this report suggests considering both risk factors and process context in evaluating risk associated with a particular use case, rather than simply considering the technology alone. LLMs such as ChatGPT are flexible enough that they could conceivably be used across all four process categories that we have outlined in this report, from Procedural Automation and Process Improvement all the way to Automated Decision-Making. Therefore, the context of a given use case is of paramount importance rather than taking a one-size-fits-all approach to a specific type of AI technology.

That said, in the case of LLMs, the technology does, at this time, generally have extremely low explainability; while the levels of boundability and reversibility for LLMs

are more process dependent. For example, if a LLM is used to generate a reply email intended for a single individual, the situation is highly bounded. Given the nature of the task, any errors made can be somewhat (although not entirely) reversed by sending a follow-up email. Conversely, if a LLM is used to write a policy document that is then used as decision-support for decision-making, the situation is markedly different with respect to boundability and reversibility. Even more so if a LLM was used for direct public-facing engagement, for example to power a chatbot that provides health advice and information. Given the very low explainability of LLMs and the commonly known challenge of these types of AI technology "hallucinating" information, one could easily imagine such a chatbot providing potentially inaccurate and medically dangerous information to an individual that would open government to significant moral, ethical, and legal risk. This risk can be somewhat, but not completely, mitigated by implementation of LLMs that provide source materials to allow users to find further information from authoritative sources or verify the information provided (e.g. a chatbot that provides a link to website with further information from which it provided a summary answer).

Annex: About the Team

This project was administered by the Institute on Governance and conducted in partnership with Think Digital. The team of key contributors to the research and writing of this report are as follows (in alphabetical order by last name):

Ryan Androsoff, Associate, Digital Governance, IOG / CEO and Founder, Think Digital

Ryan Androsoff is the Founder and CEO of Think Digital, a consultancy focused on helping public sector organizations to adapt and thrive in the era of digital disruption. Ryan is an international expert on digital government with a passion for public sector entrepreneurship and more than two decades of experience working with government and international organizations in Canada, the United States, and Europe. Since 2018 Ryan has partnered with the Institute on Governance to lead their digital leadership programs and has been providing advisory services to government organizations to assist with their digital transformation efforts.

Ryan was a Co-Founder of the Canadian Digital Service, a startup organization within the federal government launched in 2017 with the mission of helping government design, prototype, and build better digital services. Previously, he had been a Senior Advisor in the Government of Canada's Treasury Board Secretariat since 2010, where he worked on initiatives to improve digital service delivery capacity across the federal government, led the development of the first government-wide social media policies, and managed the GCTools team responsible developing the Government of Canada's first whole-of-government on-line collaborative platforms (GCpedia and GCconnex). In 2015, Ryan spent a year with the OECD's Digital Government Team in Paris, France where he was involved in a number of projects including reviews in Northern Ireland and Slovakia as well as open data and digital capacity building in the Latin American, Middle Eastern and North African regions.

Ryan's career has also included serving as a policy advisor to Canada's Minister of International Cooperation, as well as working at the World Bank in Washington, DC on initiatives to promote results-based management in international development. Ryan is a graduate of the Harvard Kennedy School of Government in Cambridge, Massachusetts where he earned a Master in Public Policy degree, with research focused on the impacts for governments of new digital technologies. Ryan also has an Honours degree in Public Affairs and Policy Management from Carleton University in Ottawa.

Aislinn Bornais, Research Assistant, Think Digital

Aislinn joins the Think Digital team as a Research Assistant supporting consulting and research projects, data analysis, and workshops. In University, she spent years studying ethical research methodology and statistical analysis. Aislinn specializes in human-centered research design for the betterment of population health outcomes. She has worked on projects that use publicly available data to analyze the current model fit of the Canadian healthcare system, as well as comparative research in the realm of addictions.

Jacob Danto-Clancy, Digital Policy Analyst, Think Digital

Jacob is a multi-disciplinary thinker dedicated to learning more about public policy challenges. He contributes to the team as a researcher and writer on projects centred around GovTech, AI, and digital infrastructure. Jacob received his MA of Public Policy in Digital Society in 2022 from McMaster as part of the program's first cohort. Jacob is also a co-founder of Boon, a research and public policy consultancy, with his friend and fellow Think Digital team member, Bryce Edwards.

Bryce Edwards, Digital Policy Analyst, Think Digital

Bryce has experience in product management, UX design, data analysis, and research in areas such as AI governance, digital government, privacy in augmented reality systems, open banking, and stablecoin regulation. As a co-founder of Boon, a research and public policy consultancy, Bryce helps clients understand and respond to new opportunities and threats that emerge from rapid technological change. Bryce is a graduate of McMaster's Masters of Public Policy and Digital Society Program.

Jen Schellinck, Associate, Think Digital

Jen Schellinck's goal as a data scientist and AI technologies specialist is to help organizations understand the value that cutting-edge data technology can bring to their work and success. She uses her knowledge of artificial intelligence, machine learning and data science to help organizations achieve their greater potential. For each project, she draws from a pool of experts to provide clients with the most valuable information they need, through consulting, workshops and data solutions. She received her PhD in Cognitive Science in 2009 and has been active in the AI field for over a decade. She is currently an adjunct researcher at the Institute of Cognitive Science at Carleton University and continues to be an active researcher.

John Stroud, Associate, Think Digital

John is a strategic adviser to leaders on linking people with technology. His vision opens people's minds to new possibilities, and he challenges them to consider creative options. People turn to John for plainspoken, easy-to-understand explanations. John is a certified OpenExO consultant in exponential technologies. Prior to launching his company AI Guides, John served as Vice President, Strategy at a federal crown corporation ($600M budget and 8000+ workforce) with responsibilities for Governance, Human Resources, Communications, Legal, Performance Measurement and Risk.

John obtained his Masters of Philosophy from Oxford, his law degree and Masters of Public Administration from the University of Victoria, and his BA from the University of Toronto. He also obtained his ICD.D for completing the Director's Education Program at the Institute of Corporate Directors.

Annex: Report References

[1] Tapani Rinta-Kahila et al., “Algorithmic Decision-Making and System Destructiveness: A Case of Automatic Debt Recovery,” *European Journal of Information Systems* 31, no. 3 (May 4, 2022): 313–38, https://doi.org/10.1080/0960085X.2021.1960905.

[2] Commonwealth of Australia, Parliament House, “Accountability and justice: Why we need a Royal Commission into Robodebt” Chapter 1, May 2022, https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Community_Affairs/Centrelinkcompliance/Final_Report/section?id=committees%2freportsen%2f024846%2f78524.

[3] Ibid., 1061

[4] Commonwealth Ombudsman, "Centrelink’s automated debt raising and recovery system," Commonwealth Ombudsman, Canberra, 2017, 2023, 9, https://www.ombudsman.gov.au/__data/assets/pdf_file/0022/43528/Report-Centrelinks-automated-debt-raising-and-recovery-system-April-2017.pdf.

[5] Rinta-Kahila et al., “Algorithmic Decision-Making and System Destructiveness.”

[6] Royal Commission into the Robodebt Scheme (Australia), "Transcript of Proceedings: Day 13," 2022, https://robodebt.royalcommission.gov.au/system/files/2022-12/transcript-hearing-day-13-5-december-2022.pdf

[7] Ibid.

[8] Ibid.

[9] Ibid.

[10] Commonwealth Ombudsman, "Centrelink’s automated debt raising and recovery system,"; Frances Mao, “The Human Cost of Australia’s Illegal ‘robo’ Hunt for Welfare Cheats,” *BBC News*, November 18, 2020, sec. Australia, https://www.bbc.com/news/world-australia-54970253.

[11] Rinta-Kahila et al., “Algorithmic Decision-Making and System Destructiveness.”

[12] Ibid., 1088.

[13] Ibid., 1087.

[14] Colin Lecher, “A Healthcare Algorithm Started Cutting Care, and No One Knew Why,” The Verge, March 21, 2018, https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy.

[15] Erin McCormick, “What Happened When a ‘Wildly Irrational’ Algorithm Made Crucial Healthcare Decisions,” *The Guardian*, July 2, 2021, sec. US news, https://www.theguardian.com/us-news/2021/jul/02/algorithm-crucial-healthcare-decisions.

[16] Colin Lecher, “Can a Robot Decide My Medical Treatment? – The Markup,” March 3, 2020, https://themarkup.org/the-breakdown/2020/03/03/healthcare-algorithms-robot-medicine.

[17] Lecher, “A Healthcare Algorithm.”; McCormick, “What Happened When a ‘Wildly Irrational’ Algorithm.”

[18] Lecher, “A Healthcare Algorithm.”

[19] Supreme Court of Arkansas, “Appeal From The Pulaski County Circuit Court [No. 60cv-17-442],” 2017, https://law.justia.com/cases/arkansas/supreme-court/2017/cv-17-183.html.

[20] Jay Stanley, “Pitfalls of Artificial Intelligence Decisionmaking Highlighted In Idaho ACLU Case | News & Commentary,” *American Civil Liberties Union* (blog), June 1, 2017, https://www.aclu.org/news/privacy-technology/pitfalls-artificial-intelligence-decisionmaking-highlighted-idaho-aclu-case.

[21] Lecher, “A Healthcare Algorithm.”

[22] Ibid.

[23] Ibid.

[24] McCormick, “What Happened When a ‘Wildly Irrational’ Algorithm.”

[25] ACLU, “Federal Court Rules Against Idaho Department Of Health And Welfare In Medicaid Class Action,” American Civil Liberties Union, March 30, 2016, https://www.aclu.org/press-releases/federal-court-rules-against-idaho-department-health-and-welfare-medicaid-class-action.

[26] Stanley, “Pitfalls of Artificial Intelligence”.

[27] Lydia Brown, “What Happens When Computer Programs Automatically Cut Benefits That Disabled People Rely on to Survive,” *Center for Democracy and Technology*, October 21, 2020, https://cdt.org/insights/what-happens-when-computer-programs-automatically-cut-benefits-that-disabled-people-rely-on-to-survive/.

[28] Ibid.

[29] Ibid.

[30] Ofqual, "Awarding GCSE, AS & A levels in summer 2020: interim report," August 13, 2020, last updated August 13, 2020, https://www.gov.uk/government/publications/awarding-gcse-as-a-levels-in-summer-2020-interim-report.

[31] Anthony Kelly, "A Tale of Two Algorithms: The Appeal and Repeal of Calculated Grades Systems in the UK and Ireland in 2020," *British Educational Research Journal* 47, no. 3 (2021): 725–41, https://doi.org/10.1002/berj.3705.

[32] Jeni Tennison, "How Does Ofqual's Grading Algorithm Work?," August 16, 2020, https://rpubs.com/JeniT/ofqual-algorithm.

[33] Department of Education and Skills, "Gov.Ie - A Guide to Calculated Grades for Leaving Certificate Students 2020," May 8, 2020, https://web.archive.org/web/20200508184911/https://www.gov.ie/en/publication/1afce4-a-guide-to-calculated-grades-for-leaving-certificate-students-2020/.

[34] Kelly, "A Tale of Two Algorithms"

[35] Richard Adams, Sally Weale, and Caelainn Barr, "A-Level Results: Almost 40% of Teacher Assessments in the UK Downgraded," *The Guardian*, August 13, 2020, sec. Education, https://www.theguardian.com/education/2020/aug/13/almost-40-of-english-students-have-a-level-results-downgraded.

[36] Kelly, "A Tale of Two Algorithms."

[37] Department of Education and Skills, "Gov.ie - A Guide to Calculated Grades."

[38] Ibid

[39] Gráinne Ní Aodha, "Harris Estimates 1,000 Additional College Places May Be Required to Deal with Calculated Grade Errors," TheJournal.ie, September 30, 2020, https://www.thejournal.ie/teachers-students-calculated-grades-5218926-Sep2020/.

[40] Lucia Nalbandian, "Increasing the Accountability of Automated Decision-Making Systems: An Assessment of the Automated Decision-Making System Introduced in Canada's Temporary Resident Visa Immigration Stream," *Journal of Responsible Technology* 10 (July 1, 2022): 100023, https://doi.org/10.1016/j.jrt.2021.100023.

[41] Petra Molnar and Lex Gill, "Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada's Immigration and Refugee System." (International Human Right Project (Faculty of Law, The University of Toronto), 2018).

[42] Government of Canada, "Overview of the Analytics-Based Triage of Temporary Resident Visa Applications.," 2020, 4; Molnar and Gill, "Bots at the Gate," 10.

[43] Government of Canada, “Overview of the Analytics-Based Triage of Temporary Resident Visa Applications,” 1. Accessed via https://meurrensonimmigration.com/artificial-intelligence-and-canadian-immigration/.

[44] Steven Meurrens, “Artificial Intelligence and Canadian Immigration” November 4, 2022, https://meurrensonimmigration.com/artificial-intelligence-and-canadian-immigration/; Nalbandian, “Increasing the Accountability.”

[45] Government of Canada, “Overview of the Analytics-Based Triage,” 4.

[46] Ibid., 5.

[47] Ibid.

[48] Nalbandian, “Increasing the Accountability,” 5.

[49] Ibid.

[50] Nalbandian, “Increasing the Accountability”.

[51] Maciej Kuziemski and Gianluca Misuraca, “AI Governance in the Public Sector: Three Tales from the Frontiers of Automated Decision-Making in Democratic Settings,” *Telecommunications Policy*, Artificial intelligence, economy and society, 44, no. 6 (July 1, 2020): 7, https://doi.org/10.1016/j.telpol.2020.101976.

[52] Jędrzej Niklas, Karolina Sztandar, and Katarzyna Szymielewicz, “Profiling the Unemployed in Poland: Social and Political Implications of Algorithmic Decision Making” (Warsaw: Fundacja Panoptykon, 2015), 12, https://panoptykon.org/sites/default/files/leadimage-biblioteka/panoptykon_profiling_report_final.pdf.

[53] Ibid.

[54] Gianluca Misuraca and Colin van Noordt, “AI Watch - Artificial Intelligence in Public Services: Overview of the Use and Impact of AI in Public Services in the EU,” *JRC Research Reports*, JRC Research Reports, July 2020, 8, https://ideas.repec.org//p/ipt/iptwpa/jrc120399.html.

[55] Niklas, Sztandar, and Szymielewicz, “Profiling the Unemployed in Poland,” 16.

[56] Ibid., 31.

[57] Jędrzej Niklas and Seeta Peña Gangadharan, “Written Submission to Special Rapporteur on Extreme Poverty and Human Rights” (London School of Economics and Political Science, May 17, 2019), 4, https://www.ohchr.org/sites/default/files/Documents/Issues/Poverty/DigitalTechnology/LSE.pdf.

[58] Ibid., 3.

[59] Niklas, Sztandar, and Szymielewicz, “Profiling the Unemployed in Poland,” 26.

[60] Ibid., 13.

[61] Kuziemski and Misuraca, “AI Governance in the Public Sector,” 8.

[62] Naomi Appelman, Ronan Ó Fathaigh, and Joris van Hoboken, “Social Welfare, Risk Profiling and Fundamental Rights: The Case of SyRI in the Netherlands,” *Social Welfare*, 2021, 263, https://www.ivir.nl/publicaties/download/jipitec_2021_4.pdf

[63] Ibid.

[64] Marvin van Bekkum and Frederik Zuiderveen Borgesius, “Digital Welfare Fraud Detection and the Dutch SyRI Judgment,” *European Journal of Social Security* 23, no. 4 (December 1, 2021): 325, https://doi.org/10.1177/13882627211031257.

[65] Valery Gantchev, “Data Protection in the Age of Welfare Conditionality: Respect for Basic Rights or a Race to the Bottom?,” *European Journal of Social Security* 21, no. 1 (March 1, 2019): 3–22, https://doi.org/10.1177/1388262719838109.

[66] Ilja Braun, “High-Risk Citizens,” AlgorithmWatch, July 4, 2018, https://algorithmwatch.org/en/high-risk-citizens/.

[67] Ibid.

[68] Melissa Heikkilä, “Dutch Scandal Serves as a Warning for Europe over Risks of Using Algorithms,” *POLITICO* (blog), March 29, 2022, https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/.

[69] Ibid. The subsequent examples covered in this paragraph are taken directly from Heikkkilä.

[70] Appelman, Fathaigh, and van Hoboken, “Social Welfare, Risk Profiling,” 264.

[71] Gantchev, “Data Protection in the Age of Welfare Conditionality,” 17.

[72] Ibid.

[73] Braun, “High-Risk Citizens.”

[74] van Bekkum and Borgesius, “Digital Welfare Fraud Detection.”

[75] Appelman, Fathaigh, and van Hoboken, “Social Welfare, Risk Profiling,” 263.

[76] Braun, “High-Risk Citizens.”

[77] ECLI:NL:RBDHA:2020:1878, Rechtbank Den Haag, C-09-550982-HA ZA 18-388 (English), (Rb. Den Haag February 5, 2020), https://deeplink.rechtspraak.nl/uitspraak?id=ECLI:NL:RBDHA:2020:1878. See sections 6.91-6.94 for the Court's take on SyRI's insufficient transparency and verifiability mechanisms.

[78] Misuraca and van Noordt, "AI Watch," 46.

[79] Braun, "High-Risk Citizens."

[80] Misuraca and van Noordt, "AI Watch," 46.

[81] Doaa Abu Elyounes, Berkman Klein Center, Harvard University: https://medium.com/berkman-klein-center/why-the-resignation-of-the-dutch-government-is-a-good-reminder-of-how-important-it-is-to-monitor-2c599c1e0100

[82] The content of this case study is derived almost entirely from "Annex A: Case Studies" in "Hello, World: Artificial Intelligence and Its Use in the Public Sector," and TC's submission to the Observatory of Public Sector Innovation Case Study Platform, which can be accessed here: https://oecd-opsi.org/innovations/artificial-intelligence-and-the-bomb-in-a-box-scenario-risk-based-oversight-by-disruptive-technology/.

[83] Wendy Nixon, "Pre-Load Air Cargo Targeting and Artificial Intelligence (Transport Canada)," https://s3.ca-central-1.amazonaws.com/cfr2018/ENG/EN+-+Emerging+Tech+in+Regs+-+Wendy+Nixon+.pdf.

[84] Ibid.

[85] Jamie Berryhill et al., "Hello, World: Artificial Intelligence and Its Use in the Public Sector" (Paris: OECD, November 21, 2019), 150, https://doi.org/10.1787/726fd39d-en.

[86] Nixon, "Pre-Load Air Cargo Targeting."

[87] "Artificial Intelligence and the 'Bomb-in-a-Box' Scenario: Risk-Based Oversight by Disruptive Technology," *Observatory of Public Sector Innovation* (blog), accessed February 14, 2023, https://oecd-opsi.org/innovations/artificial-intelligence-and-the-bomb-in-a-box-scenario-risk-based-oversight-by-disruptive-technology/.

[88] Nixon, "Pre-Load Air Cargo Targeting."

[89] Berryhill et al., "Hello, World," 150.

[90] Tony Bitzionis, "Canada's Border Agency Quietly Tested Facial Recognition at Toronto's Pearson Airport: Report," July 20, 2021, https://findbiometrics.com/canadas-border-agency-quietly-tested-facial-recognition-torontos-pearson-airport-report-072005/.

[91] Tom Cardoso and Colin Freeze, “Ottawa Tested Facial Recognition on Millions of Travellers at Toronto’s Pearson Airport in 2016,” *The Globe and Mail*, July 19, 2021, https://www.theglobeandmail.com/canada/article-ottawa-tested-facial-recognition-on-millions-of-travellers-at-torontos/.

[92] “Faces on the Move: Multi-Camera Screening,” Privacy Impact Assessment (Canada Border Services Agency, January 14, 2016), 23, https://www.theglobeandmail.com/files/editorial/News/0627-nw-na-facial-recognition/CBSA_FOTM_PIA.pdf.

[93] Cardoso and Freeze, “Ottawa Tested Facial Recognition.”

[94] Elizabeth Thompson, “Thanks to New Technology, Crossing the Border Is Going to Get Faster and Easier | CBC News,” CBC, January 24, 2022, https://www.cbc.ca/news/politics/border-airports-technology-biometric-1.6323855.

[95] Cardoso and Freeze.

[96] “Faces on the Move: Multi-Camera Screening,” 18.

[97] Cardoso and Freeze, “Ottawa Tested Facial Recognition.”

[98] Ibid.

[99] “Faces on the Move: Multi-Camera Screening,” 18.

[100] Alex Najibi, “Racial Discrimination in Face Recognition Technology,” *Science in the News* (blog), October 24, 2020, https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/.

[101] Blair Attard-Frost, Ana Brandusescu, Kelly Lyons, “The Governance of Artificial Intelligence in Canada: Findings & Opportunities from a Review of 84 AI Governance Initiatives,” (pre-print) 15, April 10, 2023.

[102] Nicholas Keung, “Did Canada Use Facial-Recognition Software to Strip Two Refugees of Their Status? A Court Wants Better Answers,” thestar.com, September 19, 2022, https://www.thestar.com/news/canada/2022/09/19/did-canada-use-facial-recognition-software-to-strip-two-refugees-of-their-status-a-court-wants-better-answers.html.

[103] Alex Boutillier, “RCMP Broke Privacy Laws in Using Controversial Clearview AI Facial Recognition Tools, Watchdog Says,” June 10, 2021, https://www.thestar.com/politics/federal/2021/06/10/rcmp-broke-privacy-laws-in-using-controversial-clearview-ai-facial-recognition-tools-watchdog-says.html.

[104] Ibid.

[105] Anne Kaun, “Suing the Algorithm: The Mundanization of Automated Decision-Making in Public Services through Litigation,” *Information, Communication & Society* 25, no. 14 (October 26, 2022): 2046–62, https://doi.org/10.1080/1369118X.2021.1924827.

[106] Ibid.

[107] Ibid.

[108] Agneta Ranerup and Helle Zinner Henriksen, “Digital Discretion: Unpacking Human and Technological Agency in Automated Decision Making in Sweden’s Social Services,” *Social Science Computer Review* 40, no. 2 (April 1, 2022): 445–61, https://doi.org/10.1177/0894439320980434.

[109] Kaun, “Suing the Algorithm.”

[110] Ranerup and Henriksen, “Digital Discretion.”

[111] Ranerup and Henriksen, “Digital Discretion.”

[112] Katarina Lind, “Central Authorities Slow to React as Sweden’s Cities Embrace Automation of Welfare Management,” AlgorithmWatch, 2020, https://algorithmwatch.org/en/trelleborg-sweden-algorithm/.

[113] Lena Waizenegger and Angsana A. Techatassanasoontorn, “When Robots Join Our Team: A Configuration Theory of Employees’ Perceptions of and Reactions to Robotic Process Automation,” *Australasian Journal of Information Systems* 26 (December 21, 2022), https://doi.org/10.3127/ajis.v26i0.3833.

[114] Government Technology Agency, “‘Ask Jamie’ Virtual Assistant,” February 21, 2023, https://www.tech.gov.sg/products-and-services/ask-jamie/.

[115] Treasury Board of Canada Secretariat (TBS), “Guideline on Service and Digital,” November 9, 2020, https://www.canada.ca/en/government/system/digital-government/guideline-service-digital.html. Our emphasis.

[116] TBS, “Guideline on Service and Digital,” (4.5.2: *Why is this important?*).

[117] TBS, “Directive on Automated Decision-Making,” February 5, 2019, https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592

[118] “Responsible Use of Automated Decision Systems in the Federal Government,” accessed February 1, 2023, https://www.statcan.gc.ca/en/data-science/network/automated-systems.

[119] TBS, “Algorithmic Impact Assessment Tool,” (2.2 *Impact levels*) March 22, 2021, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

[120] Ibid.

[121] "Responsible Use of Automated Decision Systems in the Federal Government."

[122] Teresa Scassa, "Comments on the Third Review of Canada's Directive on Automated Decision-Making," May 17, 2022, https://www.teresascassa.ca/index.php?option=com_k2&view=item&id=354:comments-on-the-third-review-of-canadas-directive-on-automated-decision-making&Itemid=80

[123] "Responsible Use of Automated Decision Systems in the Federal Government."

[124] "Responsible Use of Machine Learning at Statistics Canada," accessed January 11, 2023, https://www.statcan.gc.ca/en/data-science/network/machine-learning.

[125] Ibid.

[126] Roxana Radu, "Steering the Governance of Artificial Intelligence: National Strategies in Perspective," *Policy and Society* 40, no. 2 (April 3, 2021): 178–93, https://doi.org/10.1080/14494035.2021.1929728.

[127] Statistics Canada, Government of Canada, "The Data Literacy Training Initiative," May 3, 2021, https://www150.statcan.gc.ca/n1/pub/89-20-0006/892000062021001-eng.htm.

[128] Ibid.

[129] Statistics Canada, "The Data Literacy Training Initiative." (Our emphasis).

[130] Ibid.

[131] "Responsible Use of Machine Learning at Statistics Canada."

[132] Ibid.

[133] High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI," April 8, 2019, 11. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[134] Ibid., 12.

[135] Ibid.

[136] Ibid., 13 *fn* 31.

[137] Ibid., 11.

[138] Ibid., 15.

[139] Ibid., 19.

[140] Ibid. For the complete list of technical and non-technical methods for implementing the seven requirements, see pages 21-23.

[141] "Ethics Guidelines for Trustworthy AI | Shaping Europe's Digital Future," April 8, 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[142] "Algorithm Charter for Aotearoa New Zealand - Data.Govt.Nz,", https://data.govt.nz/toolkit/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter/.

[143] "Algorithms at MBIE: Transparency and Accountability for Our Algorithm Use," Ministry of Business, Innovation & Employment, https://www.mbie.govt.nz/science-and-technology/science-and-innovation/research-and-data/algorithms-at-mbie/.

[144] "The Algorithm Charter," Ministry of Health NZ, https://www.health.govt.nz/our-work/digital-health/digital-health-sector-architecture-standards-and-governance/algorithm-charter.

[145] "Emerging Health Technology - Advice and Guidance," Ministry of Health NZ, https://www.health.govt.nz/our-work/digital-health/vision-health-technology/emerging-health-technology-advice-and-guidance.

[146] Steven Babitch, "GSA Launches Artificial Intelligence Community of Practice," November 5, 2019, https://www.gsa.gov/blog/2019/11/05/gsa-launches-artificial-intelligence-community-of-practice.

[147] The Centers of Excellence, "Artificial Intelligence | GSA - IT Modernization Centers of Excellence," https://coe.gsa.gov/communities/ai.html.

[148] Ibid.

[149] Federal AI Community of Practice, "Artificial Intelligence Governance Toolkit," January 2022, https://coe.gsa.gov/docs/AICoP-AIGovernanceToolkit.pdf, 2.

[150] Ibid.

[151] Ibid., 11-15.

[152] USA Department of Commerce, "AI Risk Management Framework: AI RMF (1.0)," https://doi.org/10.6028/NIST.AI.100-1.

[153] Ibid., 1.

[154] Ibid., 13.

[155] Ibid., 5.

[156] Ibid., 5-6.

[157] Fasken Law, “Artificial Intelligence Risk Management Framework Published by NIST,” Technology, Media and Telecommunications Bulletin, February 9, 2023, https://www.fasken.com/en/knowledge/2023/02/artificial-intelligence-risk-management-framework-published-by-nist.

[158] Wang, X. and M. -A. Williams, (2011). "Risk, Uncertainty and Possible Worlds," 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA, USA, 2011, pp. 1278-1283, https://doi:10.1109/PASSAT/SocialCom.2011.130.

[159] Zachmann, K. (2014). Risk in Historical Perspective: Concepts, Contexts, and Conjunctions. In: Klüppelberg, C., Straub, D., Welpe, I. (eds) Risk - A Multidisciplinary Introduction. Springer, Cham. https://doi.org/10.1007/978-3-319-04486-6_1

[160] Hansson, S.O. (2012). A Panorama of the Philosophy of Risk. In: Roeser, S., Hillerbrand, R., Sandin, P., Peterson, M. (eds) Handbook of Risk Theory. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-1433-5_2

[161] International Organization for Standardization. (2009). ISO Guide 73: Risk management — Vocabulary.