Hello, World: Artificial Intelligence and its use in the Public Sector

Draft primer for public servants on the uses and considerations for AI in supporting public sector innovation and transformation.

Observatory of Public Sector Innovation (OPSI),
Reform of the Public Sector Division (RPS),
Directorate for Public Governance (GOV)
OECD

As of 1 August 2019

Note: This document represents an early version of the results of the research conducted by the OECD OPSI to help governments understand the definitions and context for AI, some technical underpinnings and approaches, how governments and their partners in industry and civil society are using AI for public good, and what implications public leaders and civil servants need to consider when exploring AI.

The document is made available to expert communities and interested individuals within and beyond OECD through an open consultation in order to surface ideas on how to make the primer even better, and to identify potential gaps or missing points. The consultation seeks to ensure that the primer it reflects an accurate representation of the current state of play of AI in the public sector.

The deadline for comments, feedback, and contributions is 1 September 2019. You may provide comments in three ways:

1. Adding comments to a collaborative Google Doc at https://bit.ly/2ST4Ujr,
2. Adding comments and edits (in tracked changes) to .doc version of the document and emailing it to opsi@oecd.org, and/or
3. Leaving comments on the public consultation announcement blog at https://oecd-opsi.org/ai-consultation.

OECD

OPSI Observatory of Public Sector Innovation

Introduction

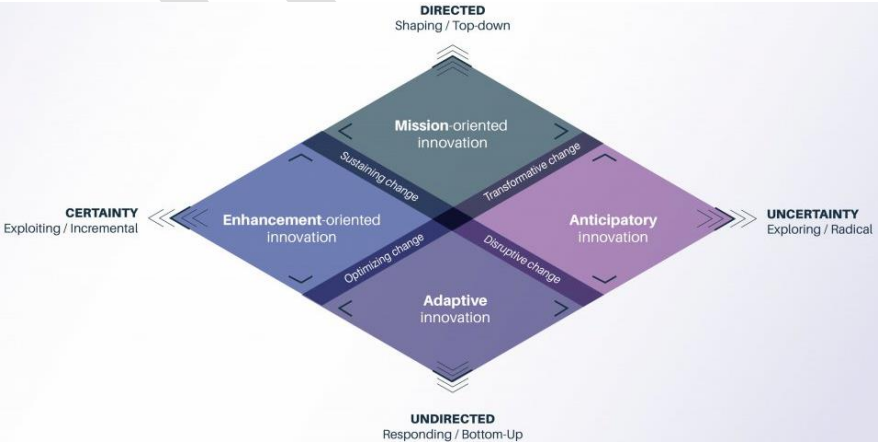
About OPSI and this document

In a time of increasing complexity, uncertainty and shifting demands, governments and public servants need to understand, test and embed new ways of doing things. The OECD Observatory of Public Sector Innovation (OPSI)[1] serves to help them in their exploration and implementation of all forms of innovative efforts, ranging from grappling with emerging technologies, to leveraging big data analytics and open data, strengthening innovation skills and capacities, promoting citizen-driven services, and fostering innovative procurement and human resource management systems, among others. OPSI's mission is to understand the dynamics of innovation in order to create and fuel systemic change in the public sector.

OPSI works to meet the needs of countries and cities around the world, and seeks to empower public servants by working with them to:

- **Uncover emerging practices and identify what is next**, by identifying new practices at the leading edge of government, connecting those engaging in new ways of thinking and acting, and considering what these new approaches mean for the public sector.
- **Explore how to turn the new into the normal**, by studying innovation in different public sector contexts and investigating potential frameworks and methods to unleash creativity and innovation and ways to connect them with the day-to-day work of public servants.
- **Provide trusted advice on how to foster innovation**, by sharing guidance and resources about the ways in which governments can support innovation to obtain better outcomes for their people.

Through its work with countries all over the world, OPSI has learned that innovation is not just one thing; it takes different forms all of which should be considered and appreciated in the public sector. OPSI has identified four primary facets to public sector innovation.[2]



- **Mission-oriented innovation** sets a clear outcome and overarching objective for achieving a specific mission.

[1] https://oecd-opsi.org.
[2] https://oecd-opsi.org/projects/innovation-facets.

- **Enhancement-oriented innovation** upgrades practices, achieves efficiencies and better results, and builds on existing structures.
- **Adaptive innovation** tests and tries new approaches in order to respond to a changing operating environment.
- **Anticipatory innovation** explores and engages with emergent issues that might shape future priorities and future commitments.

Through this work with governments, OPSI has found that a portfolio approach to innovation, which takes into account a combination of facets, is the optimal approach. In a complex world, relying on any one single approach is highly risky. Multiple options should therefore be available to offset the risk and ensure viable alternatives.

AI is a general purpose technology with the potential to have a significant effect on public policies and services. It can be used in ways that cut across and touch on multiple facets of innovation. For instance, global leaders already have strategies in place to build AI capacity as a national priority (mission-oriented). AI can be used to make existing processes more efficient and accurate (enhancement-oriented). It can be used to consume unstructured information, such as tweets, to understand citizen opinions (adaptive). Finally, in looking to the future, it will be important to consider and prepare for the implications of AI on society, work, and human purpose (anticipatory).

In order to seize its innovative potential, mitigate negative consequences, and help governments achieve a portfolio approach, public leaders and servants will need to understand AI and how it can be used, and be aware of the key considerations when doing so. To help them achieve this, OPSI has developed this primer, which draws on the work of the Working Party of Senior Digital Government Officials (E-Leaders)[3] and the OECD AI Policy Observatory.[4] It is the second in a series of overviews on topics of interest for the public sector innovation community, following on from *Blockchains Unchained*, published by the OECD as a working paper in June 2018.[5]

This document is an early version of the results of research conducted by the OECD to help governments understand how AI works and its implications for the public sector. The document is made available to expert communities within and beyond OECD through an open consultation in order to identify gaps or missing points, and to ensure it reflects an accurate representation of the current state of play of public sector AI.

What OPSI seeks through the consultation

OPSI is open to all types of feedback through the consultation, including:

- Does the report strike the right balance between technically sound yet accessible for civil servants?
- Are there any gaps, inaccurate statements, or missed opportunities for improvement? For instance, there is some debate on whether rules-based approaches should be considered AI. Did we describe this appropriately?
- Are there additional examples, tools, resources, or guidance that civil servants should be aware of?
- In what ways can AI support the primary Facets of public sector innovation?[6]

[3] http://oecd.org/governance/eleaders.
[4] http://oecd.ai.
[5] https://oe.cd/blockchain.
[6] https://oecd-opsi.org/projects/innovation-facets

Table of Contents

Summary of Initial Observations

Artificial Intelligence (AI) holds great promise for the public sector and places governments in a unique position. They are charged with setting national priorities, investments and regulations for AI, but are also in a position to leverage its immense power to innovate and transform the public sector, redefining the ways in which it designs and implements policies and services. Hype around emerging technologies often overstates or obscures their practical applications. An understanding of AI is therefore critical to helping policy makers and civil servants determine whether this technology can help them advance their missions.

Individuals and businesses interact with AI every day. Although the technology has been researched and discussed for over 70 years, there is still no uniformly accepted definition. AI means different things to different people. According to the OECD Recommendation on Artificial Intelligence, AI consists of machine-based systems that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. This guide helps to determine what this mean for public sector innovation, and aims to help public servants understand AI and navigate its implications for government policies and services.

At a technical level, while there are a variety of forms of AI, all AI today can be classified as "narrow AI"; in other words, it can be leveraged for specific tasks for which computers are well suited, for example reading and understanding text through natural language processing, identifying and classifying objects through computer vision, and understanding spoken language through speech recognition. Approaches such as "unsupervised learning", "supervised learning", "reinforcement learning", and "deep learning", which sit under the umbrella of "machine learning", hold significant potential for a variety of tasks, yet each has its own strengths and limitations, as discussed in this guide. It is important to note, however, that every AI project starts from the same point: data. Governments must take significant steps to ensure they have sufficient, quality data before they can fully take advantage of these techniques.

In adopting AI, the public sector has trailed the private sector; however, governments are seeking to rapidly catch up. To catalyse AI-driven innovation, an initial mapping conducted by the OECD has identified 38 countries (including the European Union) that have launched, or have known plans to launch, AI strategies. While at different stages of development, these include a number of common themes: economic development, trust and ethics, security and enhancing the talent pipeline. Of these 38 countries, 28 have (or plan to have) either separate strategies in place for public sector AI, or a dedicated focus embedded within a broader strategy. This is critical, as it allows AI to be integrated into the entire policy-making and service design process. These public sector components often promote a number of common themes:

- experimentation with, and funding for, government AI to automate processes, guide decision-making and develop anticipatory citizen-facing services.
- cross-government and cross-sector collaboration through councils, networks, communities and partnerships
- strategic management and use of government data, including open data, to fuel AI in all sectors
- the establishment of conditions and guidelines for transparent, ethical and trustworthy use of AI in government.
- enhancement of civil service capacity through training, tools and recruitment.

In addition to developing strategies, governments have launched real-world projects that use AI to improve efficiency and decision-making, foster positive relationships with citizens and businesses, help achieve the Sustainable Development Goals (SDGs), and solve problems in critical fields such as health, transportation and security. For instance, Canada's "bomb-in-a-box" initiative uses AI to help identify high-risk air cargo, and Latvia has developed a 24/7 virtual assistant named UNA to address customer questions. These and other projects are detailed as examples and case studies in this guide.

While AI *can* help promote innovation in government policies and services, it is not the solution for every problem. Governments must determine whether AI is the best solution for a given problem, as opposed to seeking out problems that AI can solve. Governments must also take into account many considerations when seeking to further explore and experiment with AI. They need to:

- Provide support and a clear direction but leave space for flexibility and experimentation, for example by establishing systems-wide strategies and guiding principles, communicating senior political support for AI experimentation, and developing structures inside government to incubate new approaches.
- Develop a trustworthy fair, and accountable approach to using AI, for example through establishing legal and ethical frameworks, clarifying the role of humans in AI-driven processes, pursuing the explainability of AI outcomes and developing open accountability structures.
- Secure ethical access to, and use of, quality data, for example by putting in place data strategies for managing data as an asset through its life cycle in ways that promote privacy and security while mitigating bias.
- Ensure government has access to internal and external capability and capacity to use AI through training and recruitment, collaborating and partnering externally, and designing procurement mechanisms that work for AI.

The volume of considerations that public leaders and civil servants must take into account may seem overwhelming. However, governments around the world have devised approaches to addressing each of these in their own context. This guide discusses dozens of these approaches, many of which have the potential to be adapted for use in other countries and contexts.

2. Artificial Intelligence: Definitions and context

Sometimes it can seem that Artificial Intelligence (AI) burst onto the scene only recently, but the topic has in fact been discussed and researched for over 70 years. Today, AI can be found in myriad technologies: the algorithms that mapping apps use to avoid traffic, that Netflix and Spotify use to provide recommendations for movies and songs, and that e-mail providers use to automatically filter for spam are all based on AI. Artificial Intelligence is widely debated and has become a critical asset in all sectors, including government. Dozens of countries have developed national strategies for AI, and many have pledged millions of euros (or equivalent) to fund research and development, including into the use of AI to make government operations more efficient and responsive for citizens and businesses. Governments and their partners in industry and civil society are already using AI to drive public sector innovation in areas such as healthcare, transportation and the Sustainable Development Goals (SDGs), as discussed in Chapter 3.

But what *is* AI?

Portrayals of AI in cinema or television lend themselves to visions of powerful, human-like super machines such as the cyborg in *The Terminator* or sentient computer systems with programming of various levels of quality, such as HAL 9000 from *2001: A Space Odyssey*. In the real world, people's expectations of AI range from excitement to ambivalence, and from optimism to fear. This, in part, is because AI means different things to different people. There is no uniformly accepted definition for AI, and there is not likely to be one any time soon. This guide is not intended to solve this issue, but instead provide some basics on the nature of AI to help public officials navigate this complex terrain, distinguish hype from reality and be better informed about what AI may mean in their own context. However, for the time being, AI-based technologies are still closer to Siri and chatbots than to the androids on *Westworld*.

Defining Artificial Intelligence

When talking about AI, an exact definition of "Artificial Intelligence" can prove elusive. Many definitions have been given over time, and related terms such as Machine Learning and Deep Learning **(see Chapter 2)** have recently gained traction and been associated with AI, contributing to further confusion.

The *artificial* aspect of AI is quite straightforward: it refers to anything non-natural and, in this case, man-made. It can also be represented through the use of terms such as *machines*, *computers* or *systems*. *Intelligence* is a much more widely disputed concept, explaining why there is as yet no consensus on how to define AI, even among experts (Miaihle and Hodes, 2017).

One influential approach, based on an experiment devised by Alan Turing, considers the similarities between machines and humans in displaying intelligence (see Box 1.1).

Box 1.1: The "Turing test"

In 1950, English mathematician Alan Turing developed a test, which was later named after him, that was designed to determine whether a machine (computer) could be considered intelligent. The test involved three participants: a human evaluator would ask questions, and a human and a machine would type answers. The test defines an intelligent machine as a machine that produces answers which the evaluator cannot distinguish from those of the human respondent.

Source: https://searchenterpriseai.techtarget.com/definition/Turing-test.

Many general definitions, including that used by the OECD, reflect this approach (Box 1.2). Other than being machines imitating humans, AI can also be understood as the field of knowledge associated with the design of these machines or "the discipline of creating algorithms that can learn and reason" (OECD, 2018a).

Box 1.2: OECD Definition of AI

A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy. In addition, AI are "machines performing human-like cognitive functions".

Source: OECD (2019), *Artificial Intelligence in Society*, www.oecd.org/going-digital/artificial-intelligence-in-society-eedfee77-en.htm; OECD (2019), *Recommendation of the Council on Artificial Intelligence*, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

Many other organisations across all sectors have created their own definition of AI. For instance, in their book *Prediction Machines*, Agarwal, Gans and Goldfarb (2018) adopt a business perspective asserting that "AI is a predictive technology". In the financial sector, the Luxembourg regulation authority defines AI solutions as those "focusing on a limited number of intelligent tasks and used to support humans in the decision-making process".

Multiple definitions of AI exist in large part because notions of what constitutes intelligence are subjective. For instance, Howard Gardner (1983), in his book *Frames of Mind: The Theory of Multiple Intelligences*, proposes a theory which recognises eight different abilities that make up different parts of human intelligence, including: musical-rhythmic, verbal-linguistic, logical-mathematical, and interpersonal and intrapersonal abilities. From this perspective, AI could refer to a machine that is able to write music or solve maths equations as much as a machine that can express sympathy or kindness.

Another factor that further complicates conceptions of "intelligence", and what AI entails, is *time*. What may be considered intelligent can evolve over with the passage of years. Many remarkable built-in applications available on computers or smartphones were initially considered to be a form of AI (e.g. map applications such as Google Maps which review hundreds of thousands of data points in order to provide optimal routes), but are now considered common features.

Another example is provided by chess, a game traditionally associated with AI. To some extent, a computer that can play chess against a human could be considered AI. Once computers had been taught to play the game, the objective became to see if an intelligent computer could beat a human player, and thereafter beat the top human chess players. Many tasks that computers conduct that at one point may have been considered AI are

now often seen simply as automation. A recent article made the following distinction:[7] "What's the difference between AI and automation? Well, automation is what we can do with computers, and AI is what we wish we could do. As soon as we figure out how to do something, it stops being AI and starts being automation."

While the debate about the definition of AI is fascinating and could fill many reports of its own, this guide seeks to focus more on providing an overview of the technology, and discussing the potential applications and considerations for public sector innovation and its transformation. The first step in understanding and determining the potential impact of AI for the public sector is to explore how intelligent machines can really be.

General AI vs. Narrow AI

Despite the lack of consensus in defining AI, experts on the topic generally recognise two broad perspectives that help to set expectations on how "intelligent" AI can be. The first of these is **General AI**, also known as "strong AI" or the "Artificial General Intelligence" (AGI) perspective. The second is **Narrow AI**, also known as "weak AI", "applied AI", or "Artificial Narrow Intelligence" (ANI).

General AI: A unified super intelligence

General AI refers to the idea that general human intelligence could be matched or even surpassed by using machines. In other terms, it embodies the concept that humans will someday be able to create artificial brains with the same abilities as human brains (or better). A challenging aspect of General AI, as often played out in movies, is the thought that such an AI could ultimately challenge humans and seek to replace them.

Recent developments from research in the neurosciences provide a better understanding of how our brains work and seem to suggest that General AI may be possible. Indeed, there are reasons to believe that different brain functions could be linked together and more complex cognitive functions achieved by the joining together of simpler ones (Goodfellow, Bengio and Courville, 2016). Yet, despite the broad extent of current knowledge, there is still much that is not understood about human brains, and by extension, artificial ones. While science leaves the door open, General AI remains in the realm of science fiction. Indeed, some experts are quite sceptical about the ability to manufacture such an AI at all. In the meantime, a more realistic and modest view of AI already provides many potential benefits, as well as limitations and challenges to consider.

Narrow AI: A more granular view of AI

The Narrow AI perspective, contrary to General AI, is less concerned about the creation of a unified super intelligence but rather accepts and leverages the thought that humans and computers have different relative strengths and competencies. Narrow AI takes advantage of the fact that computers are good at processing large amounts of data (see more on data in Chapter 2) and executing tasks that involve formal, explicit rules, whereas humans are still more efficient when dealing with ambiguous situations or those requiring intuition, creativity, emotion and empathy.

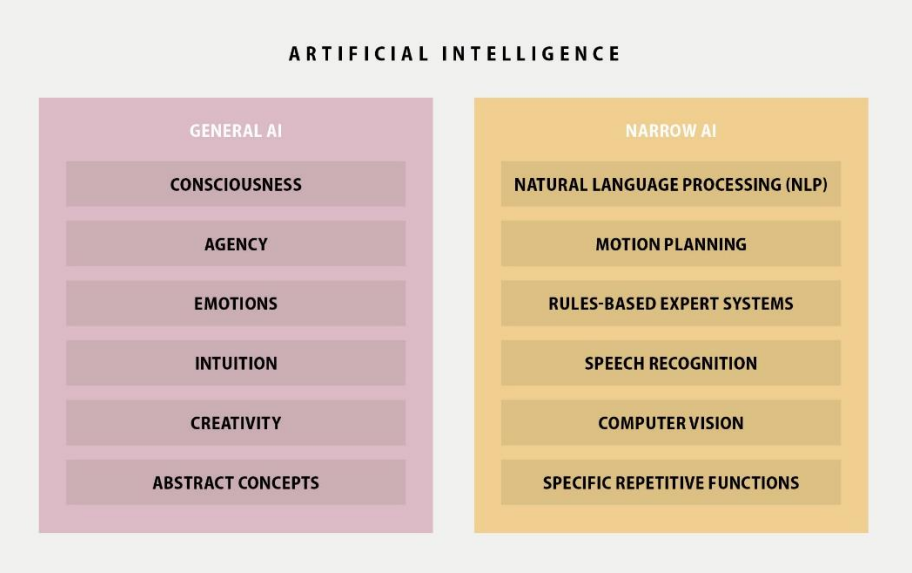
Narrow AI also reflects current machine capabilities that enable computers to be intelligent in specific areas, but do not allow these areas to be united to produce a more comprehensive intelligence. In this regard, different communities of researchers have supported the development of various AI subfields over the years, with each focused on

[7] https://arstechnica.com/features/2019/04/from-ml-to-gan-to-hal-a-peak-behind-the-modern-artificial-intelligence-curtain.

a specific set of tasks that are generally aligned with different human abilities, including the following:

- **Natural language processing (NLP)** refers to computers' ability to *read* and understand human language and perform various tasks such as translation, text generation or text analysis (see Box 1.3).
- **Computer vision** relates to the ability of AI to identify and classify objects based on images or videos.
- **Speech recognition** denotes computers' ability to analyse audio-based files in order to recognise voices and language patterns.

Figure 1.1: General vs. Narrow AI



These different approaches are discussed in more depth in Chapter 2. Box 1.3 illustrates the real-world applications of one of these tasks, NLP.

Box 1.3: OpenAI GPT-2: An example of an NLP system

In February 2019, the research organisation OpenAI published a paper presenting the GPT-2 model, a system trained to automatically predict the next word in a text. The organisation used about 40 GB of text drawn from Internet pages to train the model. For comparison, the entire works of Shakespeare take up 500 MB, 80 times less than the training dataset for GPT-2.

The model can be used to generate complete coherent sentences and paragraphs based on short human-written indications. Similarly, if the model is given a text to analyse, it can be asked related questions and will supply answers.

However, the system is not without its limits or risks. The developers observed instances of repetitive text, logic failures (e.g. writing about *fires happening under water*) and unnatural topic switching. Furthermore, such systems could be used to create content for news reports, among others, making the automation of fake news a genuine concern.

Given the potential for ill-intentioned uses of GPT-2, the creators decided to publish only parts of the source code for their model online. This decision has raised some important issues about the moral challenges of making this software fully open–source. While transparency and collaborative work should be encouraged, legal responsibilities aside, there are questions concerning which course of action to adopt in cases where there is a high risk of misuse and "weaponisation" of technology.

Applications

Next-word prediction is now a ubiquitous feature used commonly to make corrections or word suggestions when typing a text message or writing an email. In fact, email services now offer the possibility to generate answers to emails based on automatic processing of their content.

Other potential applications for this kind of system include processing a large volume of texts and producing almost instant summaries. It could also allow users to make queries based on a corpus of texts. Both these applications could be especially useful for policy makers trying to make informed decisions but also raise questions about the reliability of these systems which, as noted above, are not without their flaws.

Source: https://towardsdatascience.com/openais-gpt-2-the-model-the-hype-and-the-controversy-1109f4bfd5e8; https://openai.com/blog/better-language-models.

The above list is by no means exhaustive as there exist many more AI subfields and different task categorisations. It is important to note that these various categories, and their associated communities, are not mutually exclusive but rather evolve in ways that sometimes link up or overlap. For instance, it is evident that advances in speech recognition and NLP will influence and benefit each other as they both relate, to some extent, to understanding languages.

At present, all AI is Narrow AI. No AI algorithm, machine or computer is able to outperform humans on a *wide* range of tasks and thus fully replace humans. Instead of depicting a scenario of humans versus machines, a Narrow AI perspective demonstrates that there is sufficient space for humans and computers to collaborate and complement each other's strengths and weaknesses. By interacting with each other, humans and machines can potentially solve problems and achieve better outcomes than each could on their own. Such an approach, which emphasises interactions between humans and machines, can be referred to as Artificial Intelligence Augmentation or simply Intelligence Augmentation, as discussed later in this chapter.

Indeed, there are many examples in which teams of humans and computer working together have been able to beat not only teams of humans alone, but also teams of computers alone. One notable example of this approach is the concept of evolutionary algorithms (see Box 1.4).

Box 1.4: Intelligence augmentation: Evolutionary algorithms for human-machine collaboration

Evolutionary algorithms are a set of algorithms which can be considered part of a broader AI field that emphasises human-machine collaboration. The AI system generates a number of possible solutions to a defined problem and the human is tasked to select the instances that are most suitable. These selected instances can then be fed back into the AI system to *evolve* the model in order to further refine and perfect the proposed solutions.

Computer Science Professor Sung-Bae Cho, for example, created an AI tool that generates different designs for dresses. The user selects which designs to keep, and these choices are looped back into the AI to generate new designs. The process is repeated a number of times until it produces results the user deems satisfactory.

Similar systems have been developed in fields as diverse as industrial engineering, medicine and video games. In the case of industrial design, engineers can set constraints and generate blueprints for buildings or mechanical systems and select the best fits. In medicine, researchers can produce new drugs by using evolutionary algorithms to generate combinations of molecules and then eliminating the ones that do not produce health benefits. In the video-gaming industry, the same principles enable designers to quickly generate objects such as buildings, streets and cars that would otherwise take several hours.

Such cases highlight possibilities for humans and AI to coexist and complement each other's strengths. Decision-making in these situations is not fully automated and human intervention is essential to produce satisfying outputs.

Source: Cho (2002) "Towards creative evolutionary systems with interactive genetic algorithm", *Applied Intelligence*, *16*(2): 129-138, https://link.springer.com/article/10.1023%2FA%3A1013614519179; https://jods.mitpress.mit.edu/pub/issue3-case; www.youtube.com/watch?v=fQlQQSsC47g.

For Intelligence Augmentation approaches, it is important to evolve an appropriate design for human-computer interaction. This factor can be as important as raw computing power when considering overall performance (Case, 2018). A human operator working with a badly designed interface connected to two supercomputers can be less effective than a human operator working with a single well-designed regular computer.

With regard to the use of Intelligence Augmentation in the public sector, investments could be made to develop the use of AI to augment the abilities of civil servants in order to better do their work. Civil servants could potentially process large amounts of data faster and receive suggestions on alternative options from which to choose when providing services to citizens.

In conclusion, the distinction between General AI and Narrow AI is significant and reflects different outlooks on the potential of the technology. The current state of technology – centred on Narrow AI – already presents opportunities and pressing challenges which society and governments need to address. However, regardless of outlook, the achievements of AI may be overestimated or underestimated. The next section take a look back at the history of AI's development and tries to understand the

dynamic underpinning current enthusiasm. While the long-term evolution of AI is fundamentally uncertain, governments should seek to explore the immediate opportunities that AI presents today, while also putting in place frameworks to prepare for longer-term technological shifts.

Renewed enthusiasm for AI

Seasons of AI

As discussed, AI is not a new concept. The topic has gone in and out of fashion so frequently that dedicated communities talk metaphorically about the "winters" of AI, when people turn away from AI after it has failed to live up to its hype.

Initial scientific breakthroughs generated enthusiasm around the potential achievements of AI. People set high expectations and overestimated what could be accomplished at the time. For example, in the 1950s, at the very beginning of AI research, the introduction of the "perceptron" (a model based on the biology of human brains) suggested that humans were on the verge of creating machines that could self-teach and be used for the automation of various tasks including translation and decision making. In the 1980s, the introduction of rules-based AI approaches based on formal "if-then" rules (discussed in Chapter 2) also provoked excitement, but passion quickly slowed once rules-based systems proved difficult to deploy on a large scale.

If the bar is set too high and AI is unable to deliver on its promises, people shift their attention to other technologies. Research funding and investments are redirected towards other areas promising better returns in the short term. The 1973 Lighthill Report commissioned by the UK Parliament is frequently cited as a marker of the first AI winter (Lighthill, 1973). In this report, Professor Sir James Lighthill criticised AI for failing to meet the high expectations placed on it:

> *claims and predictions regarding the potential results of AI research had been publicised which went even farther than the expectations of the majority of workers in the field, whose embarrassments have been added to by the lamentable failure of such inflated predictions. [...] Work in the pattern-recognition field has not yet proved competitive with conventional methods. [...] Speech recognition has been successful only within the confines of a very limited vocabulary [...] The most notorious disappointments, however, have appeared in the area of machine translation, where enormous sums have been spent with very little useful result.*

As shown in Table 1.1 distinguishes three great phases in the history of AI development: the early days, knowledge-based expert systems, and the current data-driven and machine learning era.

Table 1.1: Three seasonal cycles of AI

	A	B	C	D
First spring	**1956-74** First introduction of AI Excitement around the "perceptron", a form of neural network	**1960s** Belief in machines who could think **1973** Turning point with the Lighthill Report in the UK Start of AI decline	**1950s to early 1960s** Logic-based approaches	**1956 to mid-1970s** Dartmouth Conference and first coining of the term *Artificial Intelligence* Computers able to solve mathematical proofs; use of neural networks to filter noise in telephone lines
First winter	**1974-80** Too much optimism, high expectations and inability to meet them lead to loss in research funding	**End of the 1980s** Termination of research for decades because of failure to deliver Generally low interest in AI research or no explicit mention of "AI"		Heavy investment in Automatic Language Processing (10 years, USD 20 million) results in disappointing prospects for machine translation
Second spring	**1980-87** Development of expert systems or symbolic reasoning systems based on "if-then" rules		**1970s to 1980s** Knowledge-based expert systems	**Late 1970s to 1980s** Symbolic reasoning and expert systems
Second winter	**1987-93** Difficulty to scale expert systems Too cumbersome to manually write rules			**Late 1980s** Difficulty to maintain, low return on investment Progressive integration of expert systems into standard systems
Third spring	**1993 to present (especially from early 2010s)** Machine learning and deep learning wave Instead of "if-then" rules, AI systems acquire the ability to learn through observation of input/outcome from data fed into them	**From mid-2000s onwards** Interesting performance from AI systems in categorisation and identification tasks	**2000s to present** Data-driven	**2000s to present** Deep learning boom caused by the availability of more and better quality data, but also greater computer processing power
Next step?	Sustained development or new AI winter?			

Source: OECD analysis based on: European Commission (2018a), *Artificial Intelligence: A European Perspective*, https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/artificial-intelligence-european-perspective; https://hackernoon.com/history-waves-and-winters-in-ai-dd5feb558e45;; https://towardsdatascience.com/is-deep-learning-already-hitting-its-limitations-c81826082ac3; https://warontherocks.com/2018/05/its-either-a-panda-or-a-gibbon-ai-winters-and-the-limits-of-deep-learning.

Drivers behind the current wave

Considering the recent amount of coverage and interest surrounding AI in both the private and public sectors, another spring may be underway. Many factors can be observed that explain the ongoing and growing optimism, and the following section examines these different drivers and explores what might be coming next.

Maturity of the field

As seen above, the history of AI has matured and evolved over the years. A significant body of knowledge has been accumulated with many different projects launched over the last few decades. Old algorithms and models have been refined and new ones have emerged. Programming languages have been developed and refined and many new applications created as more people become familiar with the technology.

Better technology

This knowledge creation process would not have been possible without the development and availability of high-processing computers.

Computers and even supercomputers today are cheaper, have more computing power and take up much less physical space. Moore's law encapsulates this trend observing that computers roughly double their processing capacity every two years, with current smartphones now faster than computers that occupied entire office rooms decades ago. This increase in processing power allows devices to run larger and more complex programmes and process more data and information faster.

Democratisation of computers and programming

While the technology and power behind computers has improved significantly, it has also become available to a growing number of people around the world. New users today are also more connected and better equipped to learn and exchange about AI. Many collaborative platforms and tools supported by vibrant communities are making programming and coding possible not only for academic elites, experts and mega-corporations, but also for beginners from all backgrounds. Collaborative platforms such as GitHub and Kaggle allow people to come together to collaborate on digital solutions (see Boxes 1.5 and 1.6). Such collaborative work not only helps people develop their skills and acquire new competencies, it has also become an industry-standard workflow in software engineering and has influenced other sectors. For example, the GitHub model highlights the value of collaboration for product development, monitoring changes and catching errors.

Box 1.5: GitHub platform and the case of Estonia

GitHub is a popular web-platform that serves as both a code repository and a social network. Users can host and share computer code publicly in code repositories. The platform also allows for collaborative development where multiple users can make changes to improve the code, with those contributions being tracked through Git, a version control system. For instance, the "source code" or computer code for Estonia's data exchange infrastructure (X-Road) has been published on GitHub for several years.

The ability to exchange data and ensure components are interoperable are foundational aspects for AI success, as discussed in Chapter 2. Providing others with a means to do so and collaborate on improvements can help make more rapid and open progress with AI.

Source: OPSI; https://github.com/nordic-institute/X-Road.

The platform Kaggle mixes competitive spirit with collaboration to quickly produce solutions to targeted problems.

Box 1.6: Using Kaggle to tackle challenges with AI

Kaggle is an online community for data science competitions. The platform itself is owned by Google and allows users to host and publish datasets. Publishers can then create challenges based on these datasets by providing a description of the problem they seek to solve. Data scientists can enter the competition as an individual or as a team by proposing different models which are openly available and ranked based on evaluation criteria fixed by the competition host. After a certain deadline, monetary prizes are offered to the best solutions by the host and the solution is licensed as open source software for anyone to use. The platform offers community functions to discuss the problem and exchange ideas about challenges relative to the datasets in a spirit of collaborative problem-solving.

Pneumonia detection through Machine Learning

In August 2018, the Radiological Society of North America (RSNA) partnered with organisations including the US National Institutes of Health to organise a competition through Kaggle to develop a system for the automatic detection of pneumonia cases using machine learning based on medical images (chest X-rays). A total of USD 30,000 in prize money was offered. The competition ran up to the end of October 2018 and over 1,400 teams participated. Ten teams were eventually recognised by the RSNA during their annual meeting in November 2018. In particular, the top placed team of Ian Pan, a medical student at the time, and Alexandre Cadrin-Chênevert, a radiologist and computer engineer, developed a combination of deep learning models that produced the best results for detecting cases of pneumonia and could have major effects for the treatment of this disease. The complete code for this solution is available on GitHub.

Source: www.github.com/i-pan/kaggle-rsna18; www.kaggle.com/c/rsna-pneumonia-detection-challenge#Prizes; www.kaggle.com/c/rsna-pneumonia-detection-challenge/discussion/70421; www.rsna.org/en/education/ai-resources-and-training/ai-image-challenge/RSNA-Pneumonia-Detection-Challenge-2018.

Massive Online Open Courses (MOOCs), tutorials and various websites also participate in the democratisation of knowledge in general and coding in particular, often at no cost (see Box 1.7).

Box 1.7: Useful free courses on AI

There are many available courses on AI. Here is a selection of free courses with relevance for a public sector audience:

Elements of AI is a six-part online free course on AI developed jointly by the University of Helsinki and Reaktor, a consultancy and agency services organisation. This course serves as an introduction to AI for non-experts. http://course.elementsofai.com

Introduction to Artificial Intelligence is another online free course provided by Udacity. https://eu.udacity.com/course/intro-to-artificial-intelligence--cs271.

Coursera and edX provide free access to online courses catered to a more advanced audience. Some courses offer certifications, which may come with a cost. www.coursera.org/courses?query=artificial%20intelligence.

www.edx.org/course?search_query=artificial+intelligence.

Governments in some countries also provide training to citizens of all ages and civil servants on the basics of how to use a computer, navigate the Internet and look up information. Increasingly, some also provide specialised knowledge on AI and Machine Learning (see Box 1.8). These initiatives increase computer literacy and support people's ability to leverage existing work on AI for their own personal and professional use over the long term.

Box 1.8: Government-led learning initiatives

To build up general computer literacy among citizens, French local governments have created "Éspaces publiques numériques" (Public spaces for digital technology) – centres open to all citizens that provide in-person or online introductory computer classes. More personalized, targeted trainings and other resources are also available. The first such public space was opened by the city of Strasbourg in 1996, and a network of over 5 000 facilities now operate with additional support from the national government. "Maisons de la Réussite" (Houses for success) are another locally operated structure aimed at socio-professional insertion which can offer computer classes at little to no cost.

Governments around the world are also starting to take action to increase the skills of their public workforce. For instance, the French Institute for Public Management and Economic Development (IGPDE), which forms part of the French Ministry of Economy and Finance, offers many different training courses including short, one-day sessions such as "*Digital transformation of the state and data*" and "*Artificial intelligence, data science: New economic challenges"*. These aim to equip public servants with basic knowledge about AI and its opportunities and challenges.

The Government of Singapore provides a Machine Learning and AI workshop open to public officers and, in particular, middle and senior managers. Its aim is to increase digital literacy and provide foundational knowledge about the potential of AI for public work and public organisations. (See Box 4.16 in Chapter 4 for details about the Canada School of Public Service's Digital Academy and how they provide resources for civil servants on AI.)

Source:
www.cscollege.gov.sg/programmes/pages/display%20programme.aspx?epid=cn5g9p9ecwsdnnrg2eu7pshwu1; www11.minefi.gouv.fr/catalogue-igpde/2019/co/7783.html, www11.minefi.gouv.fr/catalogue-igpde/2019/co/8618.html; www.leparisien.fr/yvelines-78/davantage-d-informatique-a-la-maison-de-la-reussite-08-07-1998-2000150304.php; www.netpublic.fr/net-public/espaces-publics-numeriques/presentation.

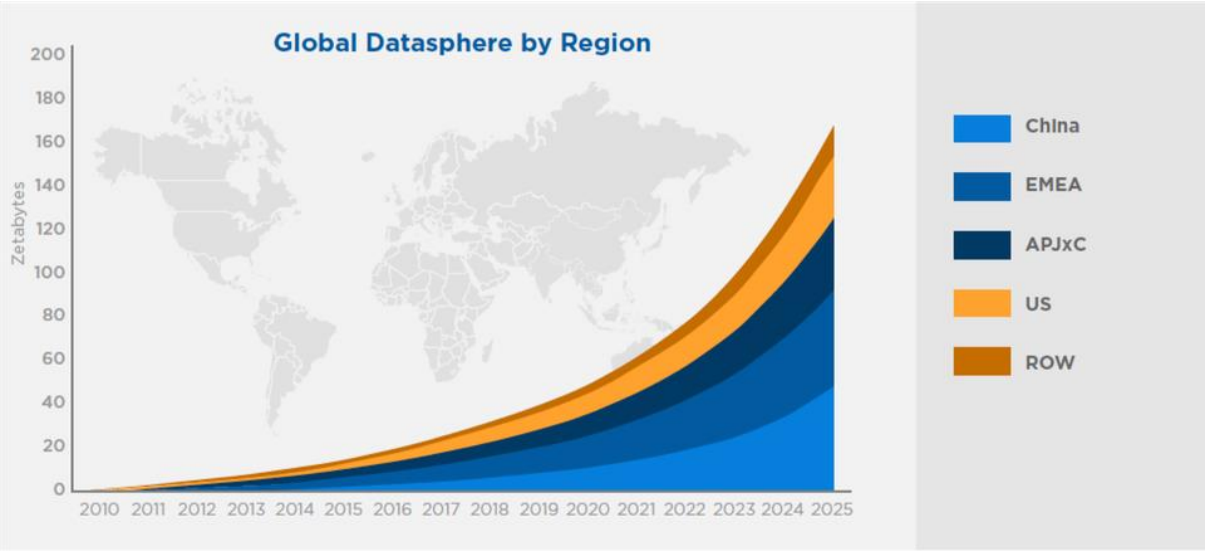
Machine learning and data availability

In addition to the availability of high-performance technology and accessible learning material, which have contributed to the development of AI as a field, the abundance of data is often-cited as the major driver behind the current wave of AI.

Countries now produce an incredible amount of data. It is estimated that 90% of the world's data was created in the last two years alone, with data generation rates still accelerating (Marr, 2018; see Figure 1.2). This phenomenon is often referred to as "Big Data" (cf. OECD, 2015a) and is characterised by:

- *Velocity*. Data are generated and processed at a faster speed than at any point in history.
- *Volume*. There is an immense quantity of data generated and stored today.
- *Variety*. Data come in different shapes and forms including text, images, video and audio.

Figure 1.2: Global data rates by region



Source: www.seagate.com/fr/fr/our-story/data-age-2025.

Public sector organisations worldwide occupy an interesting position when it comes to data generation and storage. A significant proportion of government operations relate to the maintenance of civil registries that record births, marriages and deaths, and other life events of citizens and residents. Governments also maintain tremendous amounts of other data including geospatial and weather data from satellites, property records, health and safety records, and stock exchanges, among many others. In recent years, governments have increasingly pursued the publication of government data in machine-readable format through open government data (OGD) policies. This contributes to the availability of data for AI systems to leverage. In addition, as public sector organisations increasingly offer digital services, they will be able to gather even more data enabling them to deliver more efficient and personalised services.

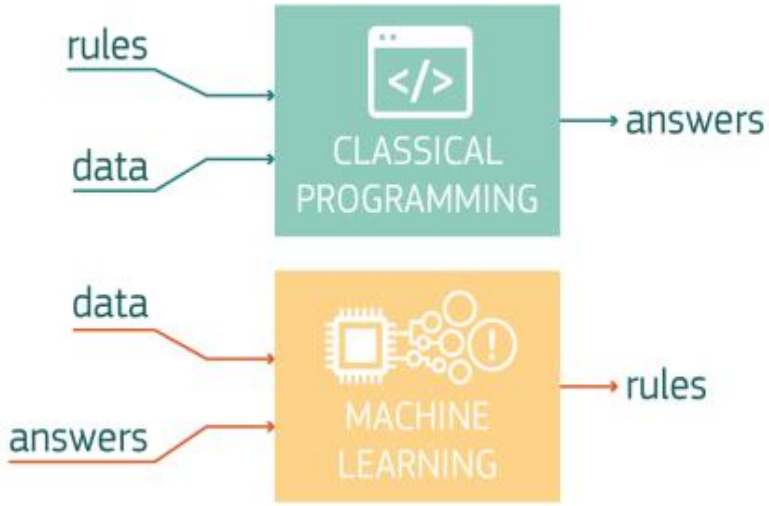
In the private sector, companies also collect data about their customers, employees and suppliers in order to run their business more efficiently. In an increasing number of cases, the collection and use of this data to present tailored advertisements constitutes the core purpose of the business. All users of web-based services generate various kinds of data: the use of mobile phones and computers, navigation of the Web, purchasing books online or dating via an app all result in the creation and sharing of large amounts of personal data. The growing use of social networks and location-based services such as personal mobility apps and travel planning tools all participate in this Big Data phenomenon.

With the advent of the Internet of Things (IoT)[8] and the development of Industry 4.0, additional important sources of data are becoming available, including access to data generated by all kinds of electronic devices, sensors, appliances, machines and vehicles.

The large and growing amount of data available, coupled with the necessary processing technologies, have made the development of Machine Learning a viable and exciting approach to AI. As illustrated in Figure 1.3, Machine Learning is a subset of AI that represents a paradigm shift in how computers can be used.

[8] IoT includes all devices and objects whose state can be altered via the Internet, with or without the active involvement of individuals. This includes laptops, routers, sensors, servers, tablets and smartphones (OECD, 2018b).

Figure 1.3: Difference between classical programming and Machine Learning



Source: European Commission (2018a), *Artificial Intelligence: A European Perspective*, https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/artificial-intelligence-european-perspective.

Instead of relying on manually written rules and having humans teaching machines how to think, the huge amount of data available can be fed to computers so that they can learn the rules themselves and produce new insights. More details on Machine Learning and other technological approaches to AI are given in Chapter 2.

The next section explores the implications of AI for the private sector, and examines how governments themselves can use the technology to improve how they function and meet their missions to serve their citizens, residents and businesses.

AI and the public sector

The previous sections have discussed AI in general and looked at the excitement surrounding AI. But what does AI mean for the public sector? How is this technology relevant for public sector leaders, public managers or civil servants working in a line agency?

A key purpose of the public sector is to elaborate and improve laws and policies, provide public goods and services to citizens and residents, and to deliver and maintain the tools, resources and structures needed for civil servants to perform their duties. Pursuant to this purpose, governments have a number of roles with regard to AI (see Box 1.9). While these roles are broad, this guide explores how they relate to innovating and transforming the processes, practices, policies and services provided by or for the government itself, rather than how they interconnect with the broader economy.

Box 1.9: Government AI roles

OECD research has identified three roles governments can play in regard to AI, often simultaneously:

- **Government as a financier or direct investor.** Governments can provide funding to support the development and adoption of emerging technologies. Some are actively pursuing a number of different funding schemes related to call for projects or pilot tenders. Such schemes include projects within the public sector, as well as private sector R&D projects whose outcomes may apply to the entire economy.
- **Government as a smart buyer and co-developer.** Governments can act as a smart buyer of existing solutions through innovative procurement practices, or as a co-developer through public-private partnerships (PPP) and other forms of collaboration to build new or tailored solutions. Governments can drive innovation from the demand side by steering the development of new solutions directly towards its needs.
- **Government as a regulator or rules maker.** Accelerated innovation cycles of emerging digital technologies call for rethinking the types of policy and regulatory instruments used and their implementation. As both an enabler and a user of emerging digital technologies, governments are facing the challenge of how to regulate them to maximise their innovative potential while minimising the risks for end users.

Source: OECD (forthcoming), "State of the art in the use of emerging technologies in the public sector".

As its role differs from that of other sectors, certain use cases and considerations for the application of AI in the public sector may be more relevant than others. For example, AI can be seen as a useful predictive machine with the potential to help policy makers make decisions on the effective and efficient allocation of resources (Agrawal, Gans and Goldfarb, 2018). It is probably less relevant for aspects such as determining which advertisements are likely to lead to sales.

In this day and age, fulfilling all these tasks can be quite challenging with governments and public organisations operating in a fast-changing environment, facing increasingly complex problems and confronting higher expectations from citizens.

Within the public sector, AI could have a positive impact in several different ways. In particular, it could be used to:

- help design better policies and make better decisions
- improve the delivery of public goods and services to citizens
- improve the internal operations of governments and public organisations in general.

While AI has a tremendous potential to positively impact the public sector, attaining these benefits will not be an easy task. Government use of AI generally trails the private sector, the technology is complex and has a steep learning curve, and the purpose and mission of – and context within – government are unique and face a number of challenges and other implications. Chapter 2 looks at the technical underpinnings that public sector stakeholders need to be aware of when trying to use AI, a baseline of which is important to know even if implementation is outsourced. Chapter 3 looks at how governments are developing AI strategies and projects. Chapter 4 discusses in detail the factors that governments and public organisations need to bear in mind when looking to

apply AI in the public sector and highlights the actions they can take to help achieve a positive outcome.

AI may not always be easy to understand and not all of its relevance to the public sector is yet clear. What is clear, however, is that governments have a significant role to play in its future. This primer aims to equip public servants with the necessary knowledge to understand this role and the basics of AI, and how it may impact public policies and services and the public sector workforce. Some of these impacts will require long-term profound changes in public sector organisations, while others may be low-hanging fruit that may be easier for governments to harvest right away. In both cases, there is a need to understand how AI can be implemented and the challenges it raises, in order for AI to have a positive impact on the public sector and ultimately on citizens' lives.

What is next for AI?

> *There is almost as much BS being written about a purported impending AI winter as there is around a purported impending [General AI] explosion.*
>
> *—Yann LeCun, recipient of ACM 2018 Turing Award*[9]

AI is developing faster than can be predicted. As noted earlier, the current development of AI is supported by changing technological conditions which are, in turn, creating enabling factors. The rapid rate of change is already impacting large parts of society including social interactions, work, education systems, the organisation of the press and how people consume media, the environment and so on.

Some of these changes incur known risks, but others may emerge through human interaction with AI systems over time. There are many unanswered questions and unknowns regarding what is next for AI. As noted earlier in this chapter, the world is now in the third cycle of AI, and there is much debate and questioning around the potential future. Will there be another winter season or is the world finally heading towards an AI summer? The truth is that no one knows for sure whether the current excitement and optimism around AI will last, and whether reality will live up to the expectations and hype that have increased in recent years.

There are many reasons to remain optimistic. A number of real-world AI systems have proven themselves with significant results, in turn receiving significant positive exposure in global media outlets. In the field of medicine, some AI algorithms have achieved greater accuracy at identifying tumours than experienced oncologists, which could potentially help reduce mortality rates (Nelson, 2019). Self-driving cars using AI are also said to be safer and cleaner than human-driven cars given the different information they can process and their resistance to distraction (Meyer, 2019).

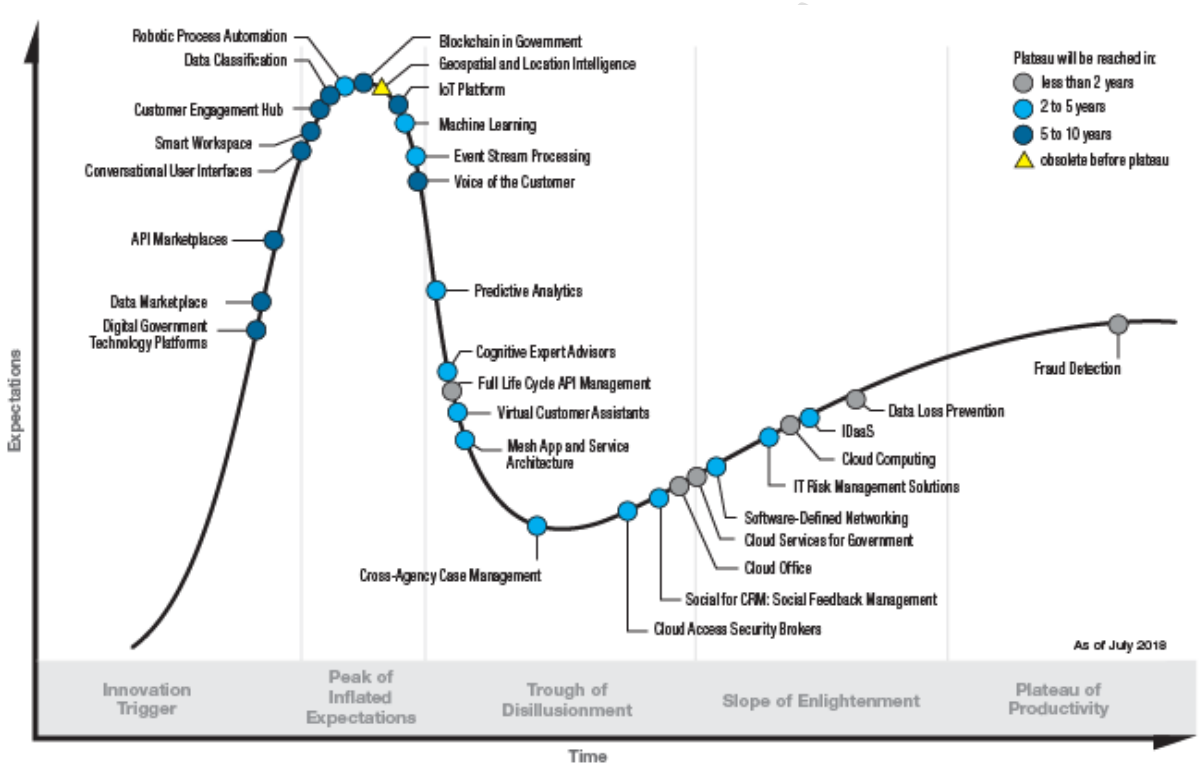
Still, as history suggests, caution should be advised. Various voices (Marcus, 2018) within the AI world are already calling for moderation and raising awareness about the current limitations and challenges of deep learning, one of the promising areas of modern AI development (see more in Chapter 2). Companies trying to profit from the AI trend and mislead customers about the level of technology being developed pose another threat and create the risk of inflated expectations with underwhelming real outcomes. A recent report by London venture capital firm MMC finds that 40% of European start-ups identified as AI companies are not actually using AI in a way significant for their business (Vincent, 2019).

When it comes to the future of AI in the public sector, the situation is slightly less clear. When compared to the broader economy and the private sector, less thought and

[9] https://twitter.com/ylecun/status/1007099197336760320.

research goes into considering how emerging technologies such as AI may impact and be used by governments. Some organisations, though, have offered theories regarding where AI may be headed in the public sector. Notably, global research firm Gartner predicts in their latest Hype Cycle for Digital Government Technology (Figure 1.4) that Machine Learning, the dominant form of modern AI, is at the "peak of inflated expectations" and is slipping down the "trough of disillusionment". Yet, it also predicts that it may take only two to five years to reach a fully productive state, which is not long when the transformational possibilities of Machine Learning are considered.

Figure 1.4: Gartner Hype Cycle for Digital Government Technology, 2018



Source: www.gartner.com/smarterwithgartner/top-trends-from-gartner-hype-cycle-for-digital-government-technology-2018.

Only time will tell if Gartner's projections are true. However, if the last 70 years of history are any indication, AI will continue to advance and grow across all sectors, even if there are periods of disillusionment along the way.

In terms of action by governments and public servants, it may not matter as much whether the trend is towards an AI winter or an AI spring. An important role for democratic governments in relation with any evolving technology is the ability to legitimately represent the voice of its citizens and to steer the development of technology towards the betterment of society as a whole. In other words, AI has a lot of potential (both positive and negative), and it is the role of governments to make sure that all people are in a position to reap the full benefits, and to mitigate risks and negative consequences on their behalf.

In order to fulfil their role, regardless of when AI approaches such as Machine Learning reach a fully productive state, governments and public servants need to understand the technology and how it may affect and impact the public sector. As discussed in Chapter 3, some governments are already deriving benefits from AI and see great

potential for the future. At a time when many governments are running to stay in place, there is little room for spectators or a "wait and see" approach. In order to remain effective decision makers, governments must acquire experiential knowledge of innovation. Public organisations need to engage with the technology and consider the implications for its institutions and the interaction between technology and citizens. Thus, OPSI advocates for experimenting with emerging technologies in informed ways that take appropriate steps to manage risks.

The next chapter of this guide is designed to help public officials (and anyone else who is interested) understand the different technical approaches and concepts behind AI. Subsequent chapters seek to explore the context of using AI in the public sector and discuss the considerations and implications that may be most relevant for government.

3. Understanding different AI approaches

When you're fundraising, it's AI. When you're hiring, it's ML. When you're implementing, it's logistic regression.

– everyone on Twitter ever[10]

Machine Learning, Neural Networks, Deep Learning and other topics related to Artificial Intelligence are now common currency. Though a high level of importance is ascribed to AI, many of these conversations are essentially symbolic. AI often remains a "black box": give it data and something innovative or futuristic will happen. While it is true that AI is a disruptive technology influencing society as a whole and driving innovation across sectors and industries, there is a general lack of understanding about how it works. For policy makers and other public servants, this presents challenges concerning how to regulate, support, use and maximise possible value from this technology while minimising negatively associated externalities. For businesses, there is a constant debate regarding when and how to use (or not) to use AI-based technologies. Decision makers in all sectors need to better understand AI, and to recognise that better, simpler solutions may exist that have a proven history or better address the problem at hand.

As discussed in Chapter 1, AI means many things to many people. This is partially because AI is an umbrella term for different types of Artificial Intelligence and different technological approaches. This chapter seeks to illuminate some of the different types, approaches and applications that fall under the umbrella of AI, and explore how they may be relevant to achieving benefits for the public sector. While the following discussion is at times technical in nature, knowing some of the specifics behind AI can assist government leaders and public servants in many ways. For example, understanding some of the technical underpinnings can help officials decide which of these tools can best be applied to address specific problems, and become more empowered in negotiations with vendors selling AI solutions.

However, before starting it is important to recognise a key precondition of almost every AI project: quality data. The importance of quality data (or lack thereof) is commonly raised by experts as the most important factor contributing to success or failure in an AI initiative. Some even state that most governments simply are not ready for AI and should focus first on getting their data in order. But what does this entail?

Data as fuel for AI

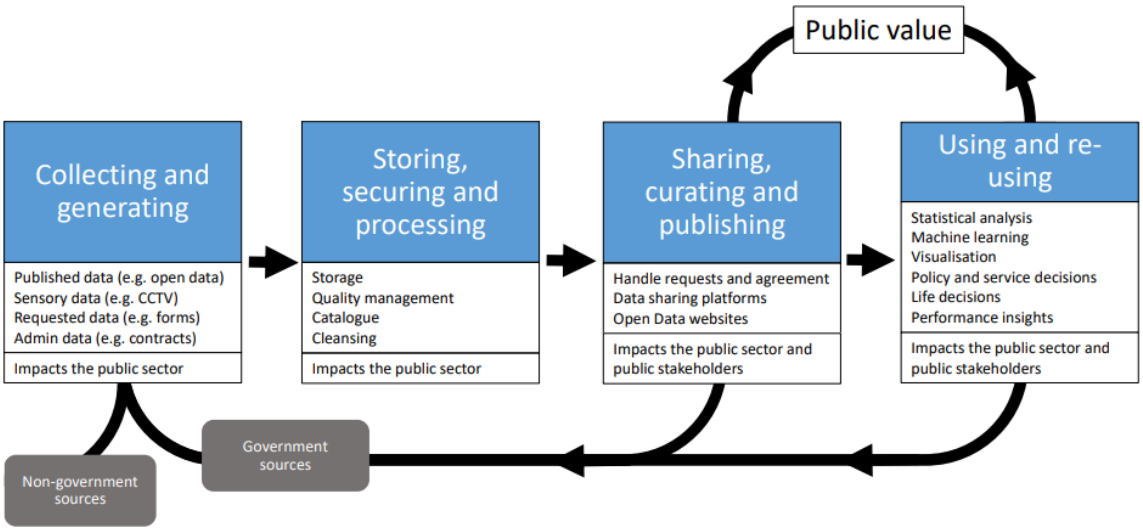
Every AI project starts from the same point: data. This is especially true of Machine Learning projects where the objective is to learn from the data. However, not all data are equal and steps must be taken to ensure that the data used for an AI project is accurate, reliable and appropriate for the task at hand. Public servants who are interested in engaging with AI need to know what data are, what types of data can be used, what sort of data AI needs and how to check if their data are ready for AI.

In order to better understand how data can – and is needed to – build the foundation for the implementation of AI, the Government Data Value Cycle (Figure 2.1) illustrates the life cycle of data and how governments can use it to generate public value, including through AI techniques such as Machine Learning. The OECD working paper *A Data-Driven Public sector: Enabling the Strategic Use of Data for Productive, Inclusive and*

[10] https://towardsdatascience.com/no-machine-learning-is-not-just-glorified-statistics-26d3952234e3.

Trustworthy Governance (van Ooijen, Ubaldi and Welby, 2019) discusses these topics in depth. Those interested in exploring AI are encouraged to read this paper, as this primer provides a higher-level discussion of this area.

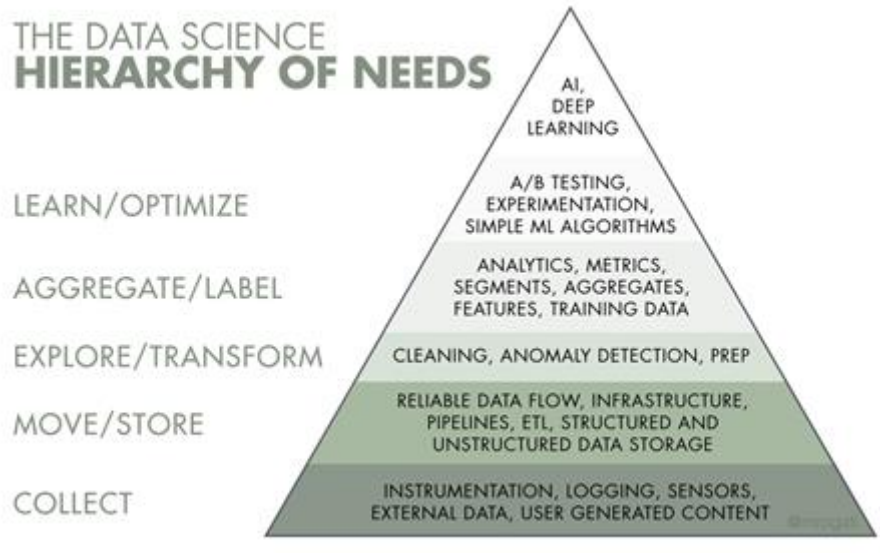
Figure 2.1: Government Data Value Cycle



Source: van Ooijen, Ubaldi and Welby (2019), *A Data-Driven Public Sector: Enabling the Strategic Use of Data for Productive, Inclusive and Trustworthy Governance*, www.oecd-ilibrary.org/governance/a-data-driven-public-sector_09ab162c-en.

Another visualisation well-considered among data scientists and AI specialists is the "Data Science Hierarchy of Needs" produced by data scientist Monica Rogati.

Figure 2.2: Data Science Hierarchy of Needs



Source: https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007.

The Data Science Hierarchy of Needs has five main levels and is read from bottom to top in order of importance. Although presented as a hierarchy, the development of an

AI project is not a purely linear process. Rather, the pyramid should serve to emphasise the importance of making foundational decisions early on in the process.

With both the Government Data Value Cycle and Data Science Hierarchy of Needs, it may be necessary to revisit and consolidate the foundations or renew them to adapt to new circumstances, as new pieces of information become available. Accordingly, they should be viewed as a dynamic and iterative process, rather than a rigid and fixed one. They both have some common focus areas that governments need to consider for building solid data management and infrastructure. The discussion below touches on aspects generally needed to build the foundations necessary to enable AI experimentation and implementation.

Collection: What data are needed and do they exist?

When considering the implementation of an AI system, the first question to ask is what kind of data is being used: is it the right data to solve the problem, or is additional data needed? An important consideration in this respect is that data always result from a specific collection process with a defined purpose. *How* data are collected ultimately determines the data that *are* collected.

Many different ways exist for collecting or generating data. Qualitative data may be obtained through methods such as individual interviews, focus groups, field observations or action research, while quantitative data may be collected through various experiments or by conducting surveys and administering questionnaires. The quantitative or qualitative nature of data will have an impact on the type of analysis that can be performed afterwards (Bryman, 2016). Governments often have vast troves of data at their disposal that have been collected for many different purposes. They also collect data from external sources, such as social media platforms, devices and sensors, and data vendors.

Governments may also share data. Owing to the development of Open Government Data (OGD) policies, national open data portals now exist for many OECD member countries. Because of the potential for data to serve as fuel for AI in all sectors, governments are expanding their OGD policies and initiatives. In fact, opening up data is one of the main components of many national AI strategies (see Chapter 3).

For parts of this section, this guide uses an example dataset to provide examples. Table 2.1 is an extract from the *Iris flower* dataset introduced by biologist Ronald Fisher in 1936. The Iris flower dataset is an example commonly used in the field of AI and especially in Machine Learning.

Table 2.1: An extract of the *Iris flower dataset*

Record	Petal length	Petal width	Sepal length	Sepal width	Species
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
...					
150	5.9	3.0	5.1	1.8	*Iris virginica*

The dataset is a record of various iris flowers observed in nature with their morphologic dimensions (length and width of petal and sepal) and the particular species of iris they belong to. In Machine Learning, it is as an example of an effort to predict a flower's species based on its dimensions.
Source: Fisher, R.A. (1936), "The use of multiple measurements in taxonomic problems", *Annals of Eugenics*, Vol. 7/2, pp. 179-188.

Storing, securing and enabling the flow of data

Being able to leverage different sources of data to fuel AI means rethinking how data are stored, how they flow and how organisations are structured, how work is managed, and how people are connected and networked. Governments need to have strong data management strategies in place to achieve this and to ensure that these items are discoverable and available from this wide array of sources inside and outside of the public sector, in an accessible and timely way (OECD, 2015b).

Traditional hierarchical bureaucracies often have limited horizontal flows of data due to rigid regulations and incompatible information management practices or good old-fashioned inter-organisational rivalries and competition (OECD, 2017). Public sector AI progress will be limited unless data can flow and be used to feed algorithms and address problems. Removing barriers and building mechanisms to enable this flow can assist with this process. In Europe, the Semantic Interoperability Community (SEMIC) is developing a common understanding about data to help facilitate exchanges across European public administrations and enable the provision of cross-border, cross-domain digital services.[11] Still, there are a number of cultural, technical and procedural challenges, as discussed in the *Data-driven Public Sector* paper.

Security is also an important consideration. As discussed in an OECD (2017) working paper, when data are being processed and stored, the databases and the arrangements around how they are managed should be transparent. Data "at rest" in this stage are often the most susceptible to digital security threats. Data security is an entire field in itself and is beyond the scope of this primer. However, public sector organisations should invest time and resources to ensure they have the proper security measures in place.

Transformation: Getting the data ready for use

> *You can't compare apples and oranges. Well actually, you can. It just needs preparation.*

Once in possession of data, it is important to ensure that they are in adequate shape to perform meaningful analyses. This step is simultaneously one of the most important, most underrated, most overlooked and least enjoyed. It is an intensive and time-consuming process. According to a survey of data scientists reported by *Forbes*,[12] it is estimated that data scientists spend up to 60% of their time on tasks related to data transformation, 20% on data collection and only 20% on the actual analysis.

To address the "apples and oranges" metaphor, both can be compared. These two types of fruits have similar shapes and different colours. They have a different texture, a different taste and may be harvested from the same country. A price tag can be set for each type. In other words, analysis is possible when the parameters are set for the analysis.

This transformation involves *data cleaning* (also known as "data wrangling" or "munging"), a "process of iterative data exploration and transformation that enables analysis".[13] For instance, data may have been collected about both apples and oranges, but the colour might be described using text in one case (e.g. red, green) and a numerical value in the other (e.g. 1 = red, 2 = green). Price information may be available for apples but not oranges. By mistake, someone may have included data about bananas, which are

[11] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/about.

[12] www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#5dd8ff3a6f63.

[13] https://web.stanford.edu/class/cs442/lectures_unrestricted/cs442-visualization.pdf.

outside the scope of this specific analysis. Box 2.1 below gives a summary of common problems encountered during data cleaning and their technical names.

Box 2.1: Common problems witnessed in data cleaning

The example of the Iris flower dataset is used to illustrate common problems encountered in data cleaning.

Record	Petal length	Petal width	Sepal length	Sepal width	Species
1	N/A	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	Long	0.2	*Rosa dumalis*
...					
150	5.9	200.3	5.1	1.8	*Iris virginica*

Missing values (in red): Sometimes certain values in datasets may be unavailable (N/A, non-available or NULL). Incompleteness in a dataset may be due to a problem in the data collection process, a subsequent bad manipulation or the result of hostile action. Missing values can be detrimental to algorithms and can result in wrong predictions, no prediction or even harm the AI system. To avoid negative outcomes from missing values, measures may be taken to understand the cause of incompleteness. Easy fixes may involve replacing the absent value with a default value or extrapolating from the existing values. More sophisticated techniques exist but are not further detailed here.

Outliers (in orange): An outlier in a dataset is a data point that stands out from the others. Causes of outliers may be the same as those for missing values. They may also indicate an issue with the modelling of the problem such as a failure to consider an important parameter or phenomenon. As with missing values, outliers can cause difficulties and negatively impact analysis. In the example above, one flower has a petal width of over 200 cm while all other irises' petal width are between 0 to 5 cm wide. When computing the average petal width, the inclusion or exclusion of an outlier has an important impact and can be subject to many interpretations during further analysis. Outlier and anomaly detection can also be a type of analysis in itself, with applications in banking and finance (see *Unsupervised learning, Clustering*).

Unexpected values (in yellow): In the iris dataset, the value *long* for sepal length is in a different format than expected (a text instead of a number), while the value *Rosa dumalis* for species is outside the range of available options (a rose instead of an iris). Here again, many causes could explain the occurrence of unexpected values. These values should be addressed by removing the record, investigating the collection process or replacing the value.

Source: OPSI.

The format in which the data have been collected can also differ: data about apples may have been typed nicely into a table in an Excel or CSV file (*structured data*), while data for oranges may include pictures and handwritten notes (*unstructured data*). In practice,

there are file formats for data such as CSV, XML and JSON, and different systems for managing databases can be used (e.g. SQL for *structured database* and NoSQL for *unstructured database*).

Figuring out these discrepancies in data and making decisions as to how to address them are steps that are sometimes overlooked and may be seen as an easy or secondary task. In fact, they are critical to producing sound analyses. Once observations and decisions have been made about the available data, data cleaning[14] per se can be carried out to address problems. This involves making the appropriate modifications to the data, a process that can be sped up through the use of rules-based AI technologies such as Robotic Process Automation (discussed later in this chapter).

Aggregating and understanding the data

The penultimate step before beginning to experiment with AI in a data project is aggregating and analysing the data to better understand them and further prepare them for AI. At this stage of the process, it is likely that some preliminary conclusions will already have been reached about the data, but work is still needed to generate working hypotheses for testing.

For example, a significant amount of information may have been gathered from interviews, but the notes made need to be structured. Alternatively, a mobile phone company may have granted access to their data on roaming and helped set up a data exchange protocol, but this may now need to be combined with geographical data to help improve traffic flow management or public transportation services. In both cases, a thorough analysis of the problem at hand and how specific variables in the data could help is necessary.

Key steps in the process are *labelling* and *feature engineering*. One way to approach this analysis is to think about what success looks like for the problem at hand, then, thinking backwards from success, determine the metrics or key performance indicators (KPIs) that need to be on a dashboard to ascertain whether the point of success has been reached. A key action is to identify a single measure of success for the project to track over time (*label*) and other things that could influence it (*features*). In some cases, these items are not immediately apparent. The use of *unsupervised Machine Learning* (a form of AI discussed later in this chapter) can help explore the data and their structure in order to better select features, as discussed in the dedicated section below.

Box 2.2 provides more details on some of the key terms mentioned in this section.

[14] https://hci.stanford.edu/courses/cs448g/lectures/CS448G-20110411-DataCleaning.pdf.

Box 2.2: Key terms for analysing and labelling data for AI

Record	Petal length	Petal width	Sepal length	Sepal width	Species
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
…					
150	5.9	3.0	5.1	1.8	*Iris virginica*

The **label** is the response variable. This is the variable whose value the process is trying to predict. It may also be referred to as the *answer*, *result* or *output*. In this case, *species* is the label and the aim is to predict the value (*setosa*, *virginica* or *versicolor*). The *label* is especially relevant in the context of *supervised learning*, a type of Machine Learning (see dedicated section below).

A **feature** is an explanatory variable in a dataset. This is one of the variables used to make predictions. In this case, the aim is to predict the iris species based on its dimensions and the descriptive factors in the table about lengths and widths for petals and sepals.

Feature engineering, according to Google's Machine Learning glossary, is "the process of determining which features might be useful in training a model". In other words, feature engineering is concerned with selecting the right data to keep as *features* for making predictions and identifying the *label*.

Feature selection is only one aspect of feature engineering. Sometimes data may exist but not in the preferred format for use by an algorithm. For instance, the model may require a person's age to allow the computation to work, but only the date of birth is available. In this case, the data needs to be **normalised**, which in this case is understood as the conversion of values from their *raw* state into a standard range of values.

Computers usually work best when they deal with numbers and even better if it is written in a binary format (a succession of zeroes and ones). The flower's *colour feature* for the iris dataset may have been added using a colour code instead of text (e.g. red is 1, blue is 2, yellow is 3). Similarly, a person's age may be of less relevance and therefore indicated using a threshold ("person is 30 years old or under" = 0, "person is over 30 years old" = 1).

Source:
http://archive.ics.uci.edu/ml/datasets/Iris, https://developers.google.com/machine-learning/glossary, https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7.

Once features and labels have been identified, it is possible to begin developing training data that allow for building Machine Learning algorithms (Box 2.3).

Box 2.3: Training and test data

With Machine Learning, data are often split into two parts: training data (usually 80% of the entire dataset) and test data (usually 20% of the entire dataset).

This section focuses mostly on reaching a necessary state to work with the data. More information on the actual processes for training and testing algorithms is given later in the chapter.

Implementing

Once the data have been collected, transformed into an adequate format and questions are flowing, the final stage of the Data Science Hierarchy of Needs is the application of AI to help produce some answers. As further detailed in the following sections, simple AI systems are based on combination of many IF-THEN rules, while modern approaches tend to involve Machine Learning and deep learning techniques.

Regardless of the approach adopted or the specific algorithm, the important thing here is to experiment with different techniques starting with more simple ones, compare their results with regard to the initial problem to solve or the hypotheses to be tested, and gradually improve on the techniques or move towards more complex models if the results are unsatisfactory.

In software development and especially website development, *A/B testing* is a common approach used to determine the best version of a website or software to keep, based on interaction with the final user. For instance, a company may want to change the template for their website, but be uncertain which design to choose. If an A/B testing approach is used, the company would show design A to 50% of new visitors and design B to the other 50%. As with any experiment, proper metrics would need to be established to measure success and determine if design A or B is the best option.

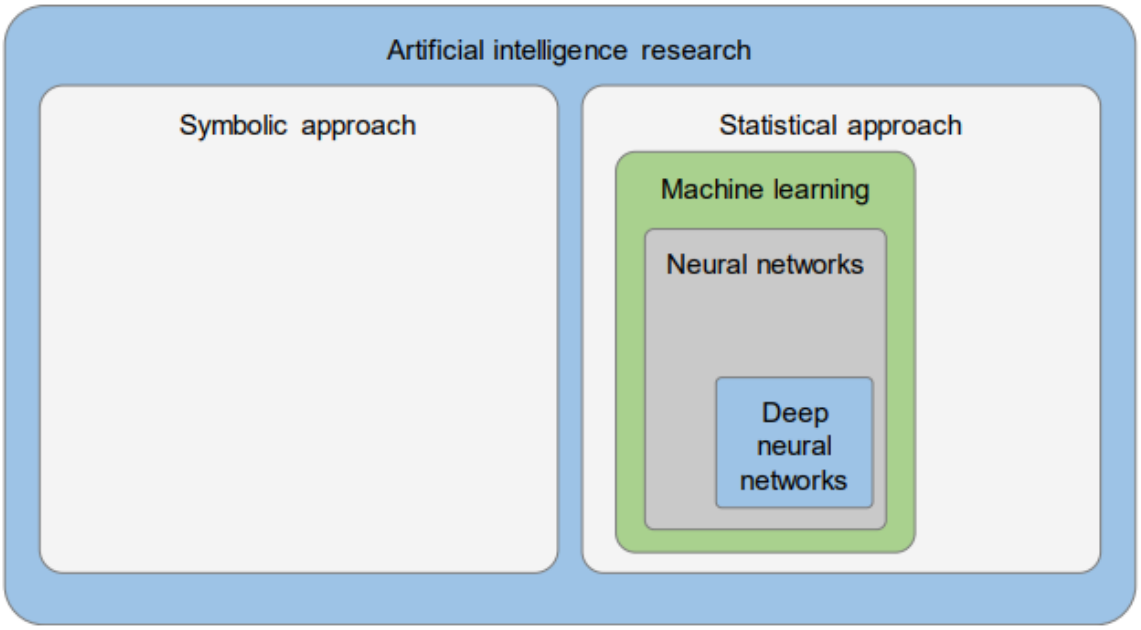
In order to better understand which kind of AI is appropriate to the problem a hand, the following sections provide some basics on how different types of AI work. They also present examples of AI used by businesses and public organisations on the basis that similar situations may be encountered by different organisations.

Evolution of AI: Rules-based AI versus Machine Learning

The previous section discussed data as a critical precondition for AI. This section explores the different types of AI and the various technological approaches that can be used for developing AI projects.

For the purposes of this report, AI can be separated into two different types: 1) rules-based AI, and 2) AI that employs Machine Learning. Over time, Machine Learning has become the dominant model, and the question of whether AI has advanced to a point where only the Machine Learning approach should be considered "intelligent" is debated by AI experts, practitioners and companies. As both may be relevant to the public sector and offered by vendors, they are both covered here.

Figure 2.3: The relationship between Artificial Intelligence and Machine Learning



Source: OECD (2019), *Artificial Intelligence in Society*, provided by the Massachusetts Institute of Technology (MIT)'s Internet Policy Research Initiative (IPRI).

This chapter provides a brief description of the characteristics of each type of AI including how they differ from one another, and their relative strengths and limitations. Box 2.4 highlights the differences between the rules-based and Machine Learning systems using the example of a computer learning to play chess.

Box 2.4: Teaching a computer to play chess: Rules-based AI and Machine Learning

Throughout the history of AI, chess has been used to illustrate how different approaches work and to judge their abilities. A chessboard consists of 64 squares, arranged in a 8x8 grid with a black and white chequered pattern. There are two possible outcomes of a chess game: one of the players wins or both players draw.

Rules-based

Each piece in chess has its own specific rules which specify the moves it can make on the board. In pseudo-code, this rule could be written as:

- "**IF** piece is pawn **THEN** it can only move one position forward straight"
- "**IF** piece is queen **THEN** it can move in all directions and for as many free squares as possible"

Players can only make one move during their turn:

- "**IF** player 1's turn **THEN** allow piece to move"

Players can capture an opponent's piece by moving into an occupied space and so on.

All the rules of chess can be formulated using "IF-THEN" (**IF** a certain condition **THEN** a certain action) statements which, taken together, describe the entire game. This collection of rules is known as the "knowledge base" of the system.

Based on this, many sets of IF-THEN rules can be developed to calculate all permitted moves in a given situation, as well as the most optimal move to make given a number of pre-programed situations to win the game.

Machine Learning

A Machine Learning approach to chess takes a different starting point. Instead of trying to list explicitly all the rules of the game, data are collected about a significant number of previously played chess games. The data could include moves made by players, the outcome of a move (is a piece captured or not) and the overall outcome of the game (win, lose or draw). Each game differs due to the possible combinations of movements made by the opposing players.

Once these data are fed into the AI system, it can be trained to infer the rules of the game on its own (as opposed to being pre-programed with rules and optimal moves) and then used to play new games.

Note: More information on building simple chess AI can be found at: https://medium.freecodecamp.org/simple-chess-ai-step-by-step-1d55a9266977 and www.j-paine.org/students/lectures/lect3/node5.html.

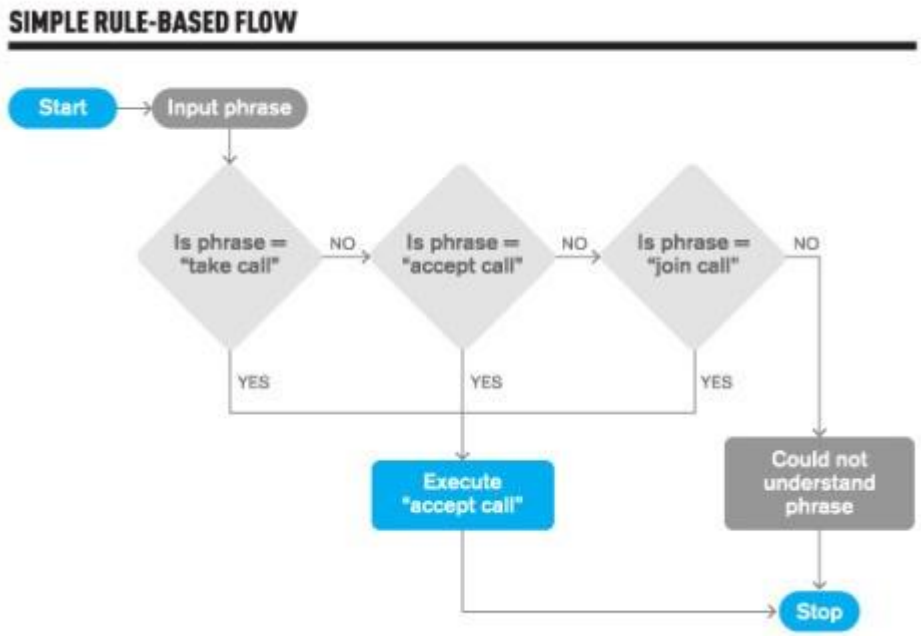
Rules-based AI

Rules-based AI, if it can be considered AI at all, is also referred to as "classic programming", "symbolic reasoning" or "symbolic AI", "expert systems" or even "Good Old-Fashioned Artificial Intelligence" (Haugeland, 1985). As discussed, these types of systems are made up of a succession of IF-THEN rules written by humans to describe a business logic or a workflow. If accepted as AI, they are the simplest type of AI. Because of their conceptual simplicity, rules-based systems generally have a high

level of *interpretability* and *explainability*.[15] Indeed, the logic behind IF-THEN statements is explicit enough that, without prior knowledge, anyone can read the different rules separately and understand or "interpret" what the system is doing.

With only a handful of adequate rules, it is possible to create a fairly elaborate system that is applicable to many different situations. For example, the case of chess provided above shows that a limited set of rules can be used to create a complex game describing many different possible situations. Using a similar programming approach, a computer could be taught the rules for processing a visa application or issuing another type of document. Figure 2.4 presents a decision flowchart for a rules-based phone menu tree similar to those used by the private and public sector to triage phone calls. Similar approaches can be used for some types of chatbots.[16]

Figure 2.4: An example of rules-based phone tree



Source: https://medium.com/botsupply/rule-based-bots-vs-ai-bots-b60cdb786ffa.

In some cases, rules-based systems can be effective and respond sufficiently to the needs of users – both citizens and civil servants – providing a fast response to straightforward requests without having to train the system with extensive data. This approach to AI was especially popular during the second wave of AI (see Chapter 1).

Nevertheless, there are several limitations to the approach. First, it requires a significant amount of knowledge about the organisation, department, team or process for which such AI systems are developed, as well as their particular context, in order to be able to explicitly articulate all the rules necessary to program the application. Another name for rules-based AI is "expert systems", because experts in these areas are needed to determine what IF-THEN rules are needed. Second, even if the knowledge is present, it may be difficult to make it explicit by verbalisation and codifying. Experts in the field need to collaborate with data engineers and data scientists to convert high-level and

[15] In this context, interpretability asks "why did the system do that" while explainability asks "how did the system do that" from https://louisabraham.github.io/ai-trust/slides.pdf.
[16] https://chatbotsmagazine.com/which-is-best-for-you-rule-based-bots-or-ai-bots-298b9106c81d.

conceptual business rules into logical programmes that machines can read. For example, policy and data specialists can model an organisation and its processes while programmers write the actual code.

Rules-based systems may be easy to implement when the processes described are relatively small and simple, and involve a limited number of actions and people to perform them. However, writing rules when the number of stakeholders grows and more and more factors have to be taken into account, rapidly becomes tedious or even impossible. When many unusual cases and exceptions have to be taken into consideration, the rule-writing process becomes complex. Another limiting aspect of rules-based systems is that the core rules are typically defined once and often do not evolve over time to reflect changes, new conditions and constraints. This is because any change or enhancement to the rules must be done manually through human intervention (e.g. programming modifications). This makes maintenance of rules-based systems a real challenge. As the number of IF-THEN statements becomes larger and more complex, it can become difficult to add new rules and statements without triggering contradictions with existing rules. Over time, this can lead to an unwieldy knowledge base with returns that diminish under its own weight. As a result, rules-based AI is most suitable for simple problems that are classified under one subject area and are unlikely to change frequently.[17] This lack of adaptability and autonomy is another reason why many believe that rules-based systems should no longer be considered "intelligent".

While these limitations are often discussed as weaknesses, they can be a strength in certain situations. Rules are easy to write, and Machine Learning could be considered an overly complex solution for some simple problems.[18] Importantly, many organisations including governments also use rules-based approaches as an initial stepping stone to the world of AI. By using rules-based approaches, these governments learn some fundamental principles and building blocks of AI. Once they have solved some "low-hanging fruit" problems, they are more likely to reach an understanding of the limitations of these approaches and encounter challenges that cannot be solved with pre-programmed IF-THEN rules. This may lead them to explore more sophisticated techniques such as Machine Learning.

An example of such a gradual approach would be governments that focus on Robotic Process Automation (RPA) to automate tasks that may be manually intensive but lend themselves well to repeated actions (Box 2.5).

[17] www.tricentis.com/artificial-intelligence-software-testing/ai-approaches-rule-based-testing-vs-learning.
[18] https://deparkes.co.uk/2017/11/24/machine-learning-vs-rules-systems.

Box 2.5: Robotic Process Automation

The distinction between automation and intelligence is often human understanding of the processes involved. In the words of British author Arthur C. Clarke, "any sufficiently advanced technology is indistinguishable from magic".

The emergence of Robotic Process Automation (RPA) as a separate technology could yet prove to be yet another case of well-understood rules-based systems being rebranded as automation. According to the IEEE Standard 2755, RPA refers to:

> *the use of a "preconfigured software instance that uses business rules and predefined activity choreography to complete the autonomous execution of a combination of processes, activities, transactions, and tasks in one or more unrelated software systems to deliver a result or service with human exception management".*

How does RPA relate to Machine Learning and other types of AI?

While Machine Learning focuses on *learning*, the core aspect of RPA is *doing*. From this perspective, RPA and Machine Learning can be complementary. The increasing automation, through RPA of tasks traditionally performed manually generates data which can then be used to train Machine Learning algorithms and obtain new insights from the work processes. In conclusion, RPA can be seen as a first step in addressing low-hanging fruit though the implementation of older-mindset AI, on top of which can be built more sophisticated and complex Machine Learning-based AI.

Source: https://standards.ieee.org/standard/2755-2017.html, www.capgemini.com/consulting-de/wp-content/uploads/sites/32/2017/08/robotic-process-automation-study.pdf, https://medium.com/@cfb_bots/the-difference-between-robotic-process-automation-and-artificial-intelligence-4a71b4834788.

Rules-based approaches such as RPA are still quite relevant today and governments often use them as a component in broader agendas concerned with efficiency, digital government and AI. This does not imply that these governments are not also pursuing techniques such as Machine Learning. Rather, they may be assembling a portfolio of actions that each leverage the comparative advantages of different technologies and techniques.

Box 2.6: RPA in the United States

In August 2018, the White House issued a new policy on Shifting from Low-Value to High-Value Work. Among other things, this policy charges US government agencies with "introducing new technologies, such as [RPA], to reduce repetitive administrative tasks". The policy guidance noted that RPA can help public administrations save time and money by automating manual and routine tasks and improving accuracy by reducing the risk of human error.

The US administration Various actions has taken various actions to support the development of RPA including the establishment of a government-wide RPA community of practice, demonstrations during public tech fairs, and the organisation of briefings, town halls and open discussions among agencies to discuss RPA and offer voluntary training sessions.

Among identified factors that could contribute positively to the development of RPA are the inclusion of employees in the change process, and an emphasis on how RPA can make civil service jobs more interesting and impactful, rather than seeing RPA as a tool for replacing jobs.

A key point highlighted by the Federal CIO, Suzette Kent, is the importance of reinvesting the savings made from RPA into other forms of IT investment, as well as helping employees move towards jobs that involve more strategic activities. Reskilling programmes have been proposed to address this challenge. Some agencies are also working on ways to track gains in government spending due to IT modernisation.

Sources: www.whitehouse.gov/wp-content/uploads/2018/08/M-18-23.pdf, www.fedscoop.com/rpa-savings-federal-agencies-reinvest-suzette-kent.

While rules-based approaches to AI were the main focus for many years, they have generally been supplanted by more sophisticated techniques that are more suitable for the growing complexity, uncertainty and interconnectedness of modern problems. The most notable of these is Machine Learning.

Machine Learning: How is it different?

> *An agent is learning if it improves its performance on future tasks after making observations about the world.*
>
> *Russell and Norvig (2016)*

In contrast to rules-based systems, Machine Learning is considered a big leap forward in the evolution and the intelligence of AI. For some, AI is not Artificial Intelligence at all unless it learns. For these people, hard coding of IF-THEN rules for machines to follow is not intelligent enough to qualify for the term "AI", even if the end results could pass the Turing Test. This is an intellectual debate beyond the scope of this guide. Regardless, Machine Learning has become the dominant form of AI, largely due to the rapid and exponential growth in the availability of data and computing power over the last few years.

Box 2.7: Key concept: Machine Learning

Machine Learning is an approach where machines learn to make predictions in new situations based on historical data. Machine Learning consists of a set of techniques to allow machines to learn in an automated manner, without explicit instructions from a human, by relying on patterns and inferences. Machine Learning approaches often teach machines to reach an outcome by showing them many examples of correct outcomes – called "training". Another approach is for humans to define a set of broad rules and generally let the machine learn on its own by trial and error.

Source: OECD (2019), *Artificial Intelligence in Society*.

What sets Machine Learning systems apart from their earlier rules-based counterparts is their ability to *learn* through experience much like humans. As already seen with rules-based systems, telling a computer about how the world works through IF-THEN rules can prove very complex for various reasons. For one, a computer has no prior knowledge of the world. In this situation, programming a computer is complex because it requires building a lot of basic blocks of information to describe the interactions. Imagine having to explain how the banking system works to children: only a limited vocabulary would be allowed and it would take time to build an understanding of commonly encountered concepts that are difficult to describe, such as money or the economy. Instead, consider the many instances where humans learn without the use of explicit knowledge, such as learning to use a new mobile phone without reading the instruction manual. Additionally, some manual activities are difficult to teach or learn using only written instructions, for example, *riding a bike*. Cognitive activities such as early language learning are another example. Various theories emphasise the role of observation, repetition and positive reinforcement or negative feedback in helping young children acquire the ability to speak before knowledge is codified and formalised.[19]

The Machine Learning approach to AI conforms to this understanding of learning. Instead of explicitly instructing computers to follow human-defined rules, computers are fed with experiences in the form of data, and allowed to extract the knowledge and rules themselves. Computers are not directly taught new knowledge, they are taught how to learn.

The remainder of the guide focuses to a greater extent on Machine Learning and its own subsets, as this approach corresponds to the current wave of interest and the dominant approach in AI. While rules-based AI may have relative strengths for certain types of applications and situations, the rise of Machine Learning fuelled by the data explosion offers new avenues and opportunities. New insights can be gleaned from Big Data and new situations can be dealt with through the use of computers even in cases where it is not possible to explicitly define rules or describe a problem.

As noted earlier, Machine Learning-based AI demands a higher level of autonomy on the part of computers, which then evolve over time through learning. However, the relative agency of AI creates ethical challenges concerning questions of ownership, responsibility and accountability, which are becoming an issue of greater concern. A number of these issues are discussed in Chapter 4.

[19] www.khanacademy.org/test-prep/mcat/processing-the-environment/language/a/theories-of-the-early-stages-of-language-acquisition.

Applying Machine Learning

To advance with Machine Learning and achieve the desired impact it is necessary to explore certain key questions:

- How do Machine Learning systems work in general terms?
- What are the different ways that machines use to learn and how can these can be used to address various problems?
- What subsets of AI exist and benefit from Machine Learning?
- What are the risks and trade-offs of Machine Learning?

Machine Learning 101: The basics

Machine Learning is an umbrella term, just like AI. Before breaking down Machine Learning further, it is important to understand what common thread shared among these different techniques justifies the idea of *learning machines*.

Training, testing and generalising

Generally speaking, the learning process in Machine Learning can be broken down into three important steps: training, testing and generalisation.

Training: During the training phase, the AI system is exposed to data which it *learns* from by applying statistical models. The training phase is similar to humans collecting experiences and learning from them by creating relations between them. Usually, only part of the entire dataset is used during training (see Box 2.2).

Example: Predict whether a person will choose to take a car or public transportation depending on the weather. The training dataset could include information about the weather such as the outlook (sunny, overcast, rain, etc.), the temperature (hot, mild, cold or real numeric values), windiness (yes, no, or numeric values for km/h) and the actual decision an individual made on their mode of transport (car or public transport).

Table 2.2

Record	Outlook	Temperature	Windy	Mode of transport
1	Sunny	Cold	Yes	Car
2	Sunny	Hot	No	Public transport
3	Rain	Cold	Yes	Car
4	Sunny	Hot	Yes	Public transport
...	...	...	...	...
200	Rain	Mild	No	Car

Testing and validation: Once the system is trained, it can provide some knowledge about the datasets. However, it is important to make sure that the system is correctly trained to solve the problem it was initially defined to tackle. For this purpose, another subset of data is used that was set aside. The validation phase is used to fine-tune and make adjustments to the parameters of the model in order to increase its performance (more details on AI performances are discussed later in this chapter).

Example: The model has been trained based on the information contained in the training dataset. Now it has to be tested to see if it is able to correctly predict whether a person will take a car or public transportation when confronted with new data (the test subset of data). If the model is unable to make the correct predictions at a sufficient level of performance, it is a sign that parameters need to be changed or maybe a different model or approach should be considered. If the model yields positive performances on the test set, further work can be undertaken to see if any improvements can be made with the validation set.

Deployment and generalisation: Once the system has been trained and has undergone proper validation and testing, it is deployed in a real environment. Now, the system works on previously unseen data collected in the field of operations in real time in order to help make decisions.

Example: The AI system has been trained on a dataset covering transport usage by a large number of citizens and the weather forecast for the past five years. It can now be deployed to better adjust the supply of public transportation based on day-to-day weather forecasts.

Although the process may appear straightforward, it is far from being linear. Many iterations may be required between training and testing to obtain the best fit of model for the task at hand. Even then, deployment of the final algorithm can be challenging as the real-life conditions may change drastically due to unforeseen events. A fourth step in the process could be added to account for the inclusion of new information, which could lead to a complete update of the model or a new cycle of training, testing and deployment.

Different ways machines can learn

As discussed, learning is the central aspect of Machine Learning. There are three main types of learning algorithms that can be used:

- unsupervised learning algorithms
- supervised learning algorithms
- reinforcement learning algorithms.

This section provides a short description of each type of learning, a basic explanation of how they work, and then discusses examples and cases to illustrate how they can be used to help improve public policy, provide better public services or make internal operations more efficient.

Unsupervised learning: Getting more insights from your data

The basics

The purpose of unsupervised learning is to gain new insights about the available data. It is closely related with the concept of *mining data* which "refers to a set of techniques used to extract information patterns from datasets" (OECD, 2015a). In particular, unsupervised learning algorithms help determine the underlying structure that may exist in a dataset by looking at the commonalities between different data points using approaches such as clustering, association rules mining or principal component analysis (see below).

Unsupervised learning requires training through the use of data. However, it does not require the use of "labelled" data (data in which the end result or answers for past actions are explicitly stated) to train a model. In other words, there is no need for human

supervision to tell the machine specifically what to look for (see labelled data in Box 2.2).

Real world applications

Clustering

Clustering is the act of trying to find common groups, or *clusters*, that exist within a dataset that may not be immediately apparent to a human observer. For humans, it can be difficult to find commonalities among elements and create groups out of a set because there are too many variables to take into consideration.

In the business world, clustering has been implemented in a number of areas. Some of the most promising ones are listed below:

- **Customer segmentation and profiling**: Companies can use clustering algorithms to segment their customers based on data about their purchase history or data collected through other means such as a membership or fidelity programme. Understanding these different segments and being able to identify those clusters of customers can produce key insights for business decisions, such as designing targeted marketing campaigns and communication plans, deciding which location to choose for a new outlet or picking the best time to launch a new product.[20]
- **Fraud and anomaly detection**: In finance and banking, clustering techniques can be used to group transactions or customers together to check for outliers that do not fit into any group. Such results may indicate fraudulent activity.[21] Other fields such as policing and security can also benefit from analysing anomalies.

Clustering applications involve the following types of analysis:

- **Intra-cluster analysis** is used to understand what is similar within a cluster.
- **Inter-cluster analysis** is used to understand what is different between clusters.
- **Outlier analysis** is used to understand why a point does not form part of a cluster.

For an example of clustering in the public sector, see the case study on using AI to crowdsource public decision making in Belgium in Annex A.

Association rules

Another application for Unsupervised Learning is to find rules and relationships between different variables in large datasets. This is known as "association rules mining". The algorithms involved work by trying to identify relationships between different transactions. This technical approach is encountered by many people in their everyday lives. For example, a supermarket may examine their sales numbers and ask how likely is someone who buys bread to also buy milk or any pair of products. Box 2.8 provides details on Amazon's similar "frequently bought together" functionality.

[20] https://towardsdatascience.com/unsupervised-learning-a-road-to-customer-segmentation-17fa2ff09d3d.

[21] www.cssf.lu/fileadmin/files/Publications/Rapports_ponctuels/CSSF_White_Paper_Artificial_Intelligence_201218.pdf.

Box 2.8: Amazon's "frequently bought together" functionality

Anyone shopping on Amazon is likely to encounter the company's "frequently bought together" functionality. Found on each item page beneath the product details, this section describes other products that customers have also bought. This functionality is a result of Amazon's version of association rules mining which they term "item-to-item collaborative filtering".

The general idea behind association rules mining is to create a list with all pairs of items and see how often customers buy them together. Doing this for every product in stock is not be efficient, as some pairs of products are never bought together by any customer. Instead, Amazon narrows the list of products to pairs by looking at customers' current shopping carts. It can also leverage its recommendation page in which customers "can filter their recommendations by product line and subject area, rate the recommended products, rate their previous purchases, and see why items are recommended" (Linden et al., 2003).

Source: www.cs.umd.edu/~samir/498/Amazon-Recommendations.pdf.

Apart from its use as a marketing tool, the mining of association rules could provide interesting insights into the health sector for medical diagnoses, by linking factors and symptoms with the probability of illness occurrence, or helping to create new medicines by checking sequences of proteins and their effects.[22] It is also frequently used to understand people's patterns of action when visiting a website, by looking at the different pages visited, the order in which they are visited or the different links that are clicked on.

Principal component analysis

Principal component analysis (PCA) is another useful aspect of unsupervised learning. The objective is to reduce the complexity of a problem by identifying the main factors that influence it. In finance and other areas, PCA can be used for risk management as it helps identify the most serious risks for prioritisation.[23]

Why is it useful?

Although increasing volumes of data are generated and collected every day, the OECD (2015a) report on data-driven innovation found that "unstructured data are by far the most frequent type of data, and thus provide the greatest potential for data analytics today".

In this context, Machine Learning systems that use unsupervised learning could provide significant benefits for governments and public organisations. By helping to make sense of large amounts of data that are available but not used effectively, unsupervised learning can convert them into practical information for making data-driven decisions.

PCA could also help public sector leaders better understand the needs of citizens based on their interactions with public services or citizens' reactions on social media, and help identify groups with common behaviours for targeted programmes. Furthermore, the aggregation of location-based or time-stamped data could help uncover new insights into topics such as emergency response, environmental monitoring and crime prevention.

[22] www.upgrad.com/blog/association-rule-mining-an-overview-and-its-applications.
[23] https://ijpam.eu/contents/2017-115-1/12/12.pdf.

With regard to other forms of learning, one positive aspect is that unsupervised algorithms do not require as much human intervention to guide the production of results. Unsupervised learning can also be seen as a preliminary step in further analysis. For example, it can be used in conjunction with supervised learning to build robust prediction systems. Unsupervised learning is *first* used to identify interesting features of a dataset, *then* supervised learning is used to correctly classify groups of data according to previously known information (see the next section for a discussion of supervised learning).

Supervised learning: The art of making predictions

The basics

Supervised learning is particularly useful when a problem has been clearly identified and there is sufficient information about the structure and content of the data. Supervised learning is generally associated with two kinds of problems: *regression* and *classification*. In both cases, the user objective is to easily generate predictions about new data points based on past observations. Regression helps to predict the numerical value of a target variable, and classification (also called categorisation) helps to predict the category to which the new data point will pertain.

- **Example of regression**: a real estate agent wants to predict the price of a house use housing market data. To do so, the agent must indicate which data the system should use. Different factors such as size, location and number of rooms act as *data features* and the price is the *target variable* to be predicted.
- **Example of classification**: a bank collects historical data on its clients and uses it to establish risk levels. The bank can then use the data to assess whether a new loan applicant is likely to repay a loan or not (e.g. high risk, medium risk and low risk).

Supervised learning differs from unsupervised learning in that it usually requires "labelled data" – data where an end outcome or answer are known for previous decisions **(**for more details see Box 2.2**)**. The term "supervised" refers to the human intervention necessary to select the output variable (i.e. outcome, answer or *label*) from the input data to guide the outcome of the AI.

Table 2.3: Predicting the use of transport modes

Record	Outlook	Temperature	Windy	Mode of transport
1	Sunny	Cold	Yes	Car
2	Sunny	Hot	No	Public transport
...	...	...	...	...
200	Rain	Mild	No	Public transport

The row coloured in green represents the target information that the user is interested in predicting (in this case, which mode of transport will be used). The row coloured in orange represents the *features* of the datasets (supporting information that can be drawn upon to make predictions). In this case, predictions are based on weather conditions using three features (outlook, temperature and windiness).

Real world applications

Regression

- Many applications of Machine Learning for regression involve the finance sector and are aimed at predicting prices, whether for stock markets, housing or any other type of assets.[24] These types of models can also be useful in the energy sector to predict demand, price and power output to optimise energy load management.[25]
- Machine Learning regression can also be leveraged in the field of emergency services to forecast demand for intervention. In 2017, the Queensland Fire and Emergency services in collaboration with the Department of Housing and Public Works, several universities and data science companies launched a project that employed regression techniques to predict the daily probabilities for different types of hazards such as floods, cyclones, fires and road crashes. This analysis then fed into a study of potential service demand scenarios and a proposal for an AI response system.

Classification

- In the telecommunication sector, classification is frequently used to understand why a customer may decide to terminate their subscription or remain. This type of analysis is referred to as *churn prediction* and, in this case, has two categories: 1) customers who stayed and 2) customers who left.[26]
- Similarly, companies can use Human Resource data to make a number of predictions, such as whether an employee is going to quit or not (*employee attrition*), and to understand the factors influencing this decision. For instance, IBM shared some human resources datasets on Kaggle and requested help with designing models to provide insights based on variables such as age, gender, job level, years at the company, years in current role and years with current manager.[27]
- Another fairly common use for is detecting spam. Aside from the inconvenience, email spam is an important issue that can directly threaten businesses and consume unnecessary resources. Supervised learning can be used to train a model on emails previously tagged as spam then classify new incoming emails. Machine Learning spam classification is the subject of much research.[28]
- Classification can also be used when there are more than two categories, such as in the case of *sentiment analysis*. For example, companies can use Machine Learning to classify tweets according to their generally positive or negative tone,[29] as well as also into more refined categories (e.g. happy, sad or angry).[30]

---

[24] https://towardsdatascience.com/predicting-house-prices-with-linear-regression-machine-learning-from-scratch-part-ii-47a0238aeac1.
[25] www.mdpi.com/1996-1073/12/7/1301/pdf.
[26] https://towardsdatascience.com/churn-prediction-770d6cb582a5.
[27] https://hackernoon.com/a-machine-learning-approach-to-ibm-employee-attrition-and-performance-b5d87c5e2415, https://www.kaggle.com/janiobachmann/attrition-in-an-organization-why-workers-quit.
[28] https://ieeexplore.ieee.org/abstract/document/5979035.
[29] www.businessinsider.fr/us/twitter-facebook-monitoring-2012-11.
[30] www.microsoft.com/developerblog/2015/11/29/emotion-detection-and-recognition-from-text-using-deep-learning.

Why is it useful?

While regression and classification problems can appear quite basic in theory (predicting a number based on various data or predicting an answer to a yes/no question), the practical use cases above indicate that many business problems can be reframed as either *regression* or *classification* situations. Similarly, in the public sector, many problems can be expressed as supervised learning problems.

Using these supervised learning methods can help make faster decisions in various public organisations as well as decisions that take into consideration more data than a single human case-handler could process. For example, since 2007, the Government of Hong Kong has been developing a system to quicken the processing of millions of application forms received at the Immigration Department, including the approval or rejection of visas.[31] In the United Kingdom, the Behavioural Insights Team worked on a decision support system to help detect children in need of specialised social care. The system combined referral information, child information and case notes to classify the children into different categories of risk.[32] Both these cases have the potential to reduce civil servants' workload, thereby improving work conditions, promoting better work performance, improving the accuracy of processing and reducing inconsistency among case handlers.

Sentiment analysis could also allow governments and public organisations to leverage social media to better capture and react faster to the needs of citizens, and understand the effects of announcing and implementing a particular policy.

Many more applications can be envisioned but require public sector stakeholders to become more familiar with the technology and understand which work problems this type of Machine Learning can help address. More use cases specific to the public sector are discussed in Chapter 3.

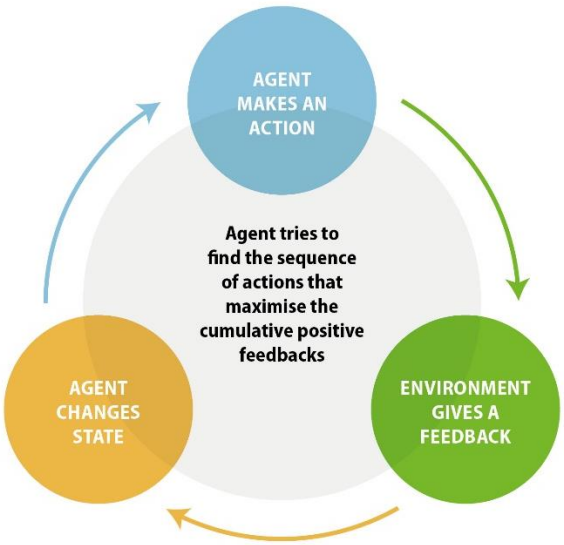
Reinforcement learning

The basics

Reinforcement learning is a type of Machine Learning that has grown in popularity recently due to advances in hardware and computing capabilities. Reinforcement learning works by having an agent (computer) complete a task by interacting with an environment. Based on these interactions, the environment will provide feedback that causes the agent to adapts its behaviour. In other terms, the agent *learns* through *trial and error*, where error is penalised by the environment and success rewarded. It then automatically adjusts its behaviour over time producing more refined actions.

[31] For more details of the Hong Kong automated visa process, see: www.cs.cityu.edu.hk/~hwchun/AIProjects/stories/km/ebrain.
[32] For more on work on children's social care, see http://38r8om2xjhhl25mw24492dir.wpengine.netdna-cdn.com/wp-content/uploads/2017/12/BIT_DATA-SCIENCE_WEB-READY.pdf.

Figure 2.5: How reinforcement works



Source: OPSI.

As an example, imagine that a company wishes to create a self-driving car. The first step would be to create a virtual simulation that replicates the intended road environment. The second step would be to define a goal or task (e.g. the car does not crash by hitting an obstacle). Given parameters such as speed, acceleration, braking and so on, the reinforcement algorithm operating the car would run in the simulated environment and learn how these different factors affect how long it is able to drive without crashing. If the algorithm performs an action that causes the car to crash quickly, it receives negative feedback, and learns to not replicate this behaviour. If it is able to drive further without crashing, it receives a positive reward and the behaviour is encouraged. Through this process of reinforcement, the machine learns how different inputs and actions affect its ability to complete its assigned task.

Real world applications

One area where reinforcement learning holds significant promise is robotics. For instance, the world's largest robot manufacturer, Japanese company Fanuc, uses reinforcement learning to create self-trained robots. The robots are first deployed in an assembly line where they train themselves through reinforcement learning to perform different tasks such as picking up objects without being explicitly taught how to do so. The training process takes about eight hours.[33]

Reinforcement learning is also frequently associated with playing traditional board games such as chess or "Go", a highly complex game popular in many parts of East Asia. Reinforcement learning algorithms have also been used to play and compete in video games tournaments.[34] In both cases, reinforcement learning has helped to achieve superhuman performances in these games due to the unorthodox strategies the computer was able to develop.[35]

[33] www.technologyreview.com/s/601045/this-factory-robot-learns-a-new-job-overnight.
[34] https://towardsdatascience.com/the-end-of-open-ai-competitions-ff33c9c69846.
[35] www.eurekalert.org/pub_releases/2019-05/aaft-dng052819.php.

Why is it useful?

These types of algorithms do not require much human supervision: once the parameters have been set, the agent learns from its own actions and errors. The system generates its own training data by experimenting and does not rely on previously collected observations unlike supervised learning. For example, a computer can be designed to learn to play chess through reinforcement learning, by playing against itself or against a human, instead of analysing data from previous games. This approach may enable computers to locate the best possible strategy rather than simply imitating human behaviour.[36] These type of algorithms could be of benefit to public sector organisations in numerous ways.

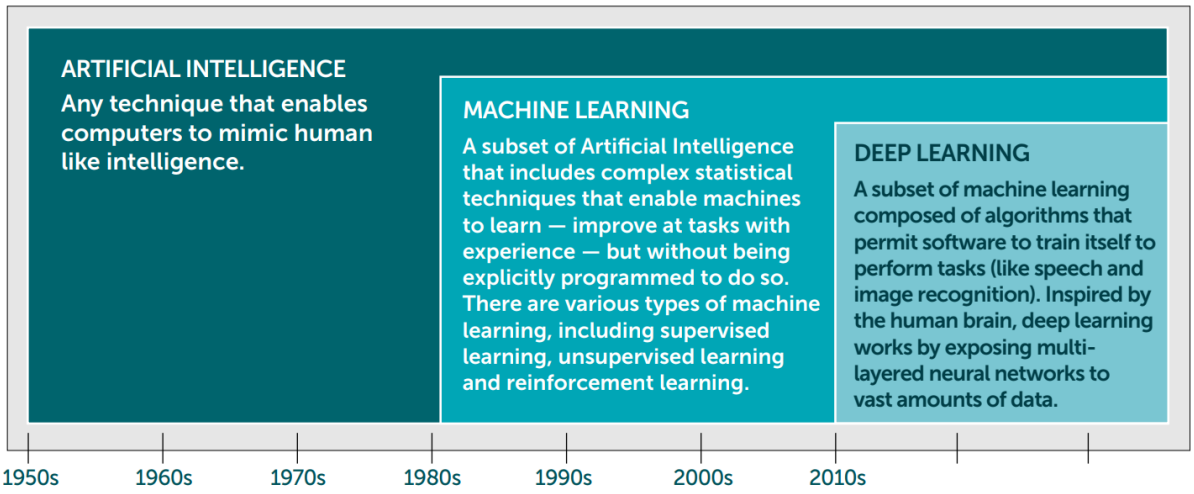
Reinforcement learning can be used during the policy design process to discover new courses of action to achieve a particular policy outcome. In 2018, a team of researchers proposed a model[37] combining reinforcement learning and deep learning to "understand the tax evasion behavior of risk-averse firms" with the objective of designing effective tax policies.

Challenges

Although reinforcement learning does not require human intervention for the agent to learn, a considerable amount of work still needs to be performed upstream to properly define the agent, the environment and the *policy*, and determine which rewards and penalties are enforced. This is not a trivial task and may require a significant level of expert knowledge. While this type of learning can be useful to determine new courses of action, it may require a lot of trial and error from the system, and therefore time and resources, before it is fully operational.

Deep learning: A biology-inspired subset of Machine Learning

Figure 2.6: Positioning deep learning within AI and Machine Learning



Source:
https://static1.squarespace.com/static/5b156e3bf2e6b10bb0788609/t/5d2c43ca74551c000190105f/1563182032127/AI+and+international+Development_FNL.pdf

[36] https://hackernoon.com/reinforcement-learning-and-supervised-learning-a-brief-comparison-1b6d68c45ffa.
[37] https://arxiv.org/pdf/1801.09466.pdf.

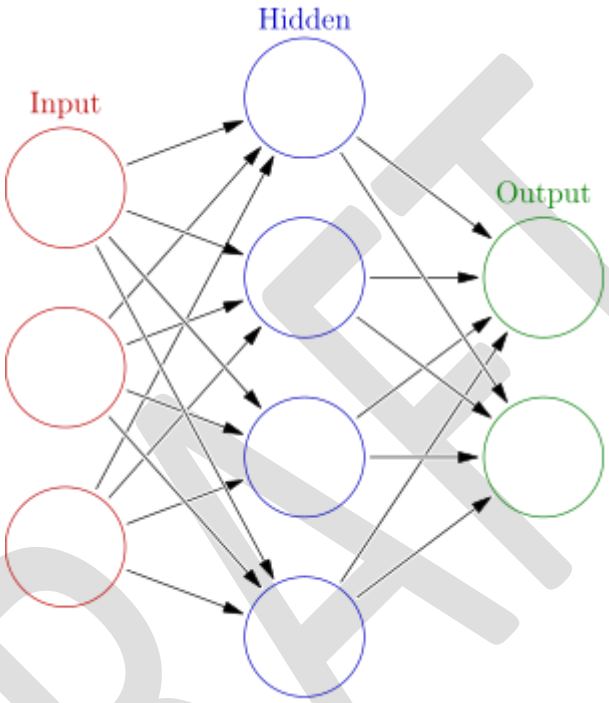
The basics

The final area of Machine Learning is **deep learning**. As with the previous Machine Learning approaches, deep learning follows the three main steps: learning, testing and generalising. The main distinction lies in the design of deep learning algorithms, which is inspired by the biology of human brains. Indeed, deep learning is often discussed in conjunction with Artificial Neural Networks (ANN), described in Box 2.9. The depth of an Artificial Neural Network relates to its number of hidden layers. Deep learning algorithms use ANNs which have two or more hidden layers. For instance, Microsoft's ResNet network is said to have 1 202 layers (Alom et al., 2018).

Box 2.9: Artificial Neural Networks (ANNs)

Scientists estimate that there are up to 100 billion *neurons* in the human brain. These are essentially nerve cells connected to each other by *synapses* that pass on information by sending electrical impulses back and forth, in the process "exciting" or "activating" the neurons.

Artificial Neural Networks try to replicate these mechanisms and behaviours using maths. ANNs algorithms are designed to have three main components: an input layer, a hidden layer and an output layer.



Each layer is made up of several *neurons* or *nodes*. Each node holds information in the form of a number. All nodes from the input layer are linked to nodes in the hidden layer which themselves are linked to nodes in the output layer. These connections are made possible through the use of various mathematical functions (e.g. the function could just be a weighted sum of the values in the previous layer's nodes). The transmission of information from one layer to the next is performed by other mathematical functions called *activation functions*.

For example, in the case of an ANN used for image classification, the nodes of the input layer may receive the colour value of each pixel from an image; the output layer is then expected to specify whether the image depicts a dog or a cat, or something different depending on the application. During the training phase, the ANN is presented with images already identified as a dog or a cat. With each new training image, the ANN learns to modify the coefficients of its activation functions to produce the expected cat/dog answer.

Source: www.ncbi.nlm.nih.gov/pmc/articles/PMC2776484.
www.youtube.com/watch?v=aircAruvnKk.
https://towardsdatascience.com/intro-to-deep-learning-c025efd92535.
https://en.wikipedia.org/wiki/Artificial_neural_network.

Real world applications

Deep learning can be used to create content that imitates human style. Figure 2.7 depicts the results of extracting style features from famous paintings and applying them to a sample image. MuseNet[38] is a deep learning network trained on hundreds of thousands of songs, which learns the style features of different composers and musicians, such as Frédéric Chopin or the Beatles, and is capable of deriving new musical pieces and even blending some of the styles.

Figure 2.7: Creating paintings in an artist's style with deep learning

Source: www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Gatys_Image_Style_Transfer_CVPR_2016_paper.pdf.

One of the most notable uses of deep learning came to light with the emergence of "deep fakes". In 2017, researchers from the University of Washington published a paper[39] in which they presented a model of deep learning that learnt about synchronisation between mouth shapes and the human voice from audio and video files of President Obama's speeches. The model was then used to create a fake video in which President Obama mouthed a re-written speech. More generally, tech giants such as Google or Baidu contribute massively to text-to-speech systems that produce AI-generated voices which read texts and are increasingly difficult to distinguish from the human voice.[40]

Aside from generating new artistic creations, deep learning can also create other deep learning algorithms and computer programs autonomously. For example, Google Brain, the team researching deep learning at Google, ran an experiment[41] in which it tasked two neural networks to exchange text communications in a protected fashion while a third network tried to decipher the messages. The deep learning algorithms managed to successfully establish secure communications using their own cryptography technique. Other tools such as Neural Complete[42] are using neural networks to facilitate the writing of new deep learning models.[43]

---

[38] https://openai.com/blog/musenet.
[39] http://grail.cs.washington.edu/projects/AudioToObama/siggraph17_obama.pdf.
[40] https://arxiv.org/pdf/1609.03499.pdf.
[41] https://arstechnica.com/information-technology/2016/10/google-ai-neural-network-cryptography.
[42] https://github.com/kootenpv/neural_complete.
[43] For more examples of deep learning applications, see www.yaronhadad.com/deep-learning-most-amazing-applications.

Why is it useful?

Deep learning is currently one of the most promising areas of AI research. It can be applied to all kinds of problems including those undertaken by standard Machine Learning in addition to more complex problems.[44] This has led some in the field to refer to it as a "universal learning approach" (Alom et al., 2018). Furthermore, deep learning algorithms can achieve impressive performance in comparison with more traditional Machine Learning techniques. For instance, deep learning networks hold the record in accuracy for algorithms used to recognise handwritten digits (see MNIST benchmarks).[45] Deep learning is also behind recent claims for best performance in cancer detection.[46] In general, deep learning algorithms also seem to be better at leveraging large amounts of data when compared with other forms of Machine Learning. More data usually translates into better performance, however Machine Learning tends to reach a cap, while deep learning performance keeps improving when more data is fed into the network.[47]

Challenges

Among the main challenges of deep learning is the current inability to fully understand what exactly happens during the training of neural networks (i.e. how exactly algorithms evolve to make their decisions). The different layers in a deep learning algorithm are believed to provide new levels of abstraction with each new layer added, and thus networks with more nodes and more layers are usually thought to be better at tackling more complex problems. In image recognition, for example, it is supposed that one layer may be able to identify the edges in a picture while another may be able to assemble those edges to recognise more complex patterns such as loops or straight lines.

Until more knowledge is produced on the inner workings of deep learning algorithms, they will suffer from the perception of being a *black box* technology. This makes their functionality and any results they generate suffer from challenges of *explainability* (see Chapter 4). Another important challenge to deep learning is the tension between the resources required to function and the performance achieved. While, deep learning techniques offer great accuracy in terms of prediction and can tackle more complex problems, they also require a lot of computational power, higher-end computers and large amounts of data to train on.

Other AI subfields benefiting from Machine Learning

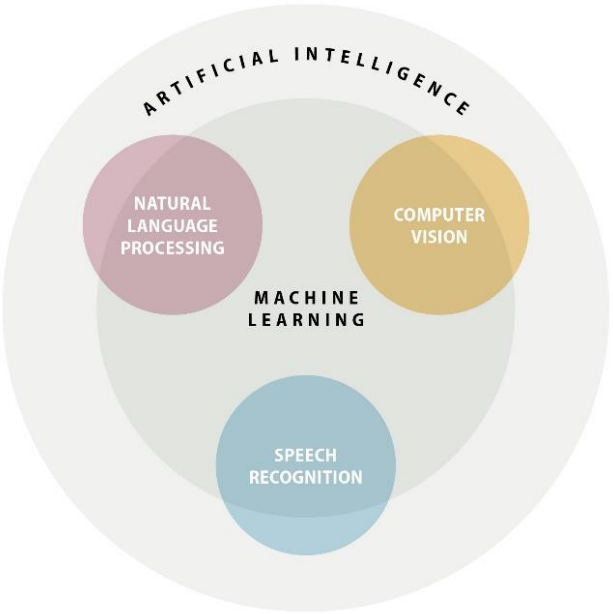
As touched on in Chapter 1, AI can be broken down into many subfields that deal with different problem areas. The rise of Machine Learning as a distinctive approach to AI has allowed more established communities to reconsider the kind of problems that can be solved, the level of performance that can be achieved and the resources required to achieve these performances.

Machine Learning can also be used as a technological approach in and of itself, as well as combined with other AI approaches (see Figure 2.8). While each of these approaches

[44] https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063.
[45] http://yann.lecun.com/exdb/mnist.
[46] https://healthitanalytics.com/news/google-deep-learning-tool-99-accurate-at-breast-cancer-detection, https://medium.com/future-today/biomind-artificial-intelligence-that-defeats-doctors-in-tumour-diagnosis-5f8ec97298b2.
[47] https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063.

could potentially be conducted in a rules-based manner, their full power may only be realised with the introduction of Machine Learning.

Figure 2.8: Approaches to AI



Source: OECD based on https://medium.com/@chethankumargn/artificial-intelligence-definition-types-examples-technologies-962ea75c7b9b.

When reflecting on these other approaches, it can be useful to think about senses such as sight, sound and touch, with learning acting as a cross-cutting dimension that links and can help co-ordinate these different senses. Box 2.10 presents the case of a robot picking up an object, an action that involves co-ordination and the ability to touch and see. These actions can represent subfields in AI. This section discusses some of the most common.

Box 2.10: Picking up an object

"Look around and pick up an object in your hand, then think about what you did: you used your eyes to scan your surroundings, figured out where are some suitable objects for picking up, chose one of them and planned a trajectory for your hand to reach that one, then moved your hand by contracting various muscles in sequence and managed to squeeze the object with just the right amount of force to keep it between your fingers."

The ways that AI systems function are also broken down into similar and additional approaches, which can sometimes achieve a specific task on their own, or be combined with other approaches.

Source: Elements of AI online course (https://course.elementsofai.com/1/1), OPSI.

Computer vision

Computer vision is a subfield of AI that focuses mainly on the analysis of images and video files. Computer vision has many potential applications of great interest to the public sector. In the field of medicine, there is significant activity around the detection of diseases such as cancer (see Chapter 3). Image recognition systems also work

autonomously to scan license plates enabling cashless road tolls. However, more sophisticated computer vision systems can be developed that leverage Machine Learning techniques. Once these techniques are combined, AI can learn, recall and recognise images and identify patterns.[48] Facial recognition is one key example of this. In the public sector, the combination of Machine Learning and these techniques could be used for tasks such as person identification, policing and land-use management (see Box 3.8 in Chapter 3).

Natural Language Processing

Natural Language Processing (NLP) is another subset of Artificial Intelligence which deals with the understanding and analysis of human language by computers. Usually, NLP systems accept text-based documents as inputs. They can also be combined with other subfields such as computer vision, for example, to analyse text from scanned documents or text embedded in images and videos.

NLP has many very useful applications that span from basic spam filtering to various forms of text analysis such as document categorisation, real-time translation and sentiment analysis (see Box 2.11) in addition to text generation (see Box 1.3 on GPT-2 in Chapter 1). Personal assistants like Amazon's Alexa or Apple's Siri are among the most illustrative examples of NLP systems combined with speech recognition features. Public organisations could potentially leverage these technologies to provide more personalised services to citizens and businesses based on specific interactions (see Box 3.7 in Chapter 3 on the UNA chatbot).

[48] www.quora.com/What-is-the-relation-between-machine-learning-image-processing-and-computer-vision.

Box 2.11: Sentiment analysis of social media in Kenya

In a study published in 2019, researchers Chris Mahony, Eduardo Albrecht and Murat Sensoy attempted to use AI to elicit the relationship between online discourse and political violence in the context of Kenya.

Their starting hypothesis was that the language used by influential figures can be linked to increased tension and the risk of violence. To test this hypothesis, the researchers collected and analysed data from Twitter using a NLP programme that assessed the tweets' sentiment scores (positive, negative or violent) to track changes in the emotions of influential Kenyan political actors.

The software combined elements of NLP and Machine Learning. For instance, it used a "bag-of-words" approach to compute the sentiment scores: each tweet was attributed a score based on the frequency of words used and the intensity associated with specific words. Deep learning was then employed to convert the unstructured data (a tweet or a blog post) into structured data (a numeric score). A *random forest* algorithm was used to correlate sentiment score with daily fatalities reported through the Armed Conflict Location and Event Data Project.

Ultimately, the researchers found that their model could accurately predict increases and decreases in average casualties up to 150 days in advance. The promising result should be seen as a first step towards an AI system that could potentially anticipate political conflicts and help take measures to prevent casualties. Such a system would also help obtain a better understanding of the relationship between language and violence.

Source: www.theigc.org/wp-content/uploads/2019/02/Language-and-violence-in-Kenya_Final.pdf.

One challenge to note in the development of NLP is the lack of resources for training algorithms in languages other than English (and the general prevalence of English). This should be taken into consideration when trying to create new services based on NLP. Initiatives in different countries have started to address this particular issue, for example the Italian Natural Language Processing Lab[49] or the EUR 90 million investment made by the Spanish government to support the NLP industry.[50]

Speech recognition

Speech recognition is another area of AI closely linked with NLP. The key difference is that it focuses on the analysis of audio as input, rather than text. Paired with NLP, it has the potential to profoundly affect the ways that people interact with their electronic devices to access services and control appliances. The combination of speech recognition with Machine Learning has led to ongoing advances that are helping to create increasingly sophisticated voice interfaces more responsive to the context and which interact more and more like humans. This is important to note as it drives increased acceptance and adoption by consumers of technologies such as voice command personal and home assistants.[51] Another potential use for speech recognition would be crime-solving.[52]

[49] www.italianlp.it.
[50] https://slator.com/demand-drivers/thats-big-spain-pours-100-million-into-language-technology.
[51] https://medium.com/swlh/the-past-present-and-future-of-speech-recognition-technology-cf13c179aaf.
[52] www.globalme.net/blog/new-technology-in-speech-recognition#Voice_Technology_in_Public_Transportation.

Machine Learning performance

Training and generalising a model are essential mechanics of Machine Learning systems. However, predictions or descriptions of a dataset are only useful if they are accurate. Similarly, waiting hours to obtain this information is not beneficial. In order to assess the performance of Machine Learning systems and be able to compare various algorithms, researchers have developed indicators or *metrics*. Below is a non-exhaustive list of widely used indicators:

Metrics for all Machine Learning approaches

Time and speed

The speed at which an algorithm trains can also be an important indicator to consider when picking a particular approach to Machine Learning. Various factors can account for how fast an algorithm is trained including the computing power of the machine running the algorithm, the amount of training data to be processed, the specific algorithm used and the code used to implement it. The previous example of self-trained Japanese robots shows that the learning process can take several hours before the robots are able to be fully deployed for operations. Testing the solution and making sure that it produces meaningful results may also require time depending on the type of model and approach used.

Robustness

Robustness is another parameter that can help guide the selection of Machine Learning algorithms and is an area of intense research.[53] Broadly speaking, robustness refers to the ability of a model to cope with anomalies, outlier points or noise that may be present in a dataset, with the objective of producing consistent results. A robust algorithm would be one that can distinguish noise from interesting information.

Similarly, outliers and anomalies are observations that do not follow the general trend and may be the result of another phenomenon. For instance, in the transport mode prediction model, public transportation use may be strongly correlated with sunny weather and warm temperatures, whereas car use may be strongly linked to situations where there is heavy rain and wind. However, a user may decide to take public transportation on a rainy, windy day because the car is undergoing maintenance. Multiple people may decide similarly to take their car because of road construction. All of these cases may then appear as outliers or anomalies to the general model which only focused on weather features as predicting factors.

Another classic example is an image classification algorithm that wrongly tags pictures because of an alteration made to the training data. For instance, researchers published a paper in which they demonstrated how a top Google image recognition model incorrectly classified fire trucks as school buses with minor modifications to the pictures such as rotating the image.[54]

This latter case is especially important to bear in mind when planning to use AI, as it highlights the potential security risks that can arise when using these technologies. This issue is discussed later in the chapter.

[53] https://towardsdatascience.com/the-three-pillars-of-robust-machine-learning-specification-testing-robust-training-and-formal-51c1c6192f8.
[54] www.zdnet.com/article/googles-best-image-recognition-system-flummoxed-by-fakes.

Metrics for supervised learning

The *confusion matrix* is a useful tool within the context of binary classification (yes/no problems) to compute performance metrics.

Table 2.4: A confusion matrix

		Actual values	
		Positive	**Negative**
Predicted values	**Positive**	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

Drawing from the previous example of predicting the use of a car or public transport, the different elements of the confusion matrix can be understood as follows:

- **True positive (TP)**. The model predicted that the citizen will take the car (or public transportation), and he indeed takes the car (or public transportation).
- **True negative (TN)**. The model predicted that the citizen will not take public transportation (or the car), and he indeed does not take public transportation.
- **False positive (FP)**. The model predicted that the citizen will not take the car, and he actually takes the car.
- **False negative (FN)**. The model predicted that the citizen will take public transportation, and he actually does not take public transportation.

Among the various metrics, ***accuracy*** is the most common and functions as an expression of how often the AI system provides a correct response. It is often displayed as a percentage or ratio showing the proportion of times that the AI system made the correct call (true positive + true negative). A high percentage of accuracy suggests that the model gives correct suggestions most of the time. For instance, the model used in the mode of transportation example has an 80% accuracy rate, which means that 8 out of 10 times, the model has or will correctly assess whether an individual has taken their car or public transportation. Accuracy can come into play at two different key times for an AI system. First, it plays an important role during the testing and validation phase to help determine whether a model needs further refinement. Second, it is important for monitoring and measuring the performance of models that have been deployed in real-world situations.

Another metric related to the confusion matrix is ***sensitivity***. This measures the ratio of true positives and is especially important for health and finance applications. For instance, when predicting a disease, it is critical to correctly predict whether a patient is effectively sick (true positive) in order to be able to treat him or her in a timely fashion. On the other hand, a good or bad prediction is less significant if the patient is indeed healthy (false negative or true negative). A number of other performance metrics exist.[55]

When dealing with non-binary predictions, that is, cases where the algorithm needs to classify data points across more than two different categories (e.g. yes/no), a slightly different, more complex approach needs to be taken in order to consider all cases.

[55] https://en.wikipedia.org/wiki/Confusion_matrix.

Balancing all those different metrics is very important when planning to implement AI systems. Prioritising one metric over another will require discussions and trade-offs and will depend on the specific application for AI under consideration. While this discussion may involve various stakeholders, the final decision cannot be the sole responsibility of technical operators and needs to represent the interests and values of citizens.

Metrics for unsupervised and reinforcement learning

For unsupervised and reinforcement learning, performance evaluation of algorithms can be a more delicate and context-dependent task. Indeed, as noted, both types of learning involves some form of uncertainty about the insights produced, which can make it difficult to determine immediately if the results are good or bad.

For clustering problems, it may be useful to search for high *intra-cluster similarity* or *cluster cohesion* (elements from the same group are very similar), low *inter-cluster similarity* (elements from different groups are not similar at all) and high *cluster separation* (the cluster observed is very distinct from the other clusters) in order to compare the results of different methods of grouping. More technical details can be found in resources from Kent State University or *Introduction to Information Retrieval* (Manning, Raghavan and Schütze, 2008) on clustering evaluation and different metrics such as *purity* or the *Rand Index* (drawing on the concept of *accuracy* described earlier).

In the case of reinforcement learning, one way to compare the results produced by an algorithm is to track the amount and rate of positive feedback an agent gets over time when adopting a particular strategy.[56]

Machine Learning: Risks and challenges

As seen in the previous section, selecting an algorithm and measuring its performance is a difficult but necessary task which requires much effort and input from all stakeholders involved in the AI system life cycle. This section provides an overview of the broader technical challenges to consider when implementing an AI system.

Generalisation, underfitting and overfitting

The term *generalisation* is usually understood to mean any broad statement made about a group of people or objects – turning a fact about some cases into a fact about all cases, and making an assertion that can be true sometimes into one that is always true.

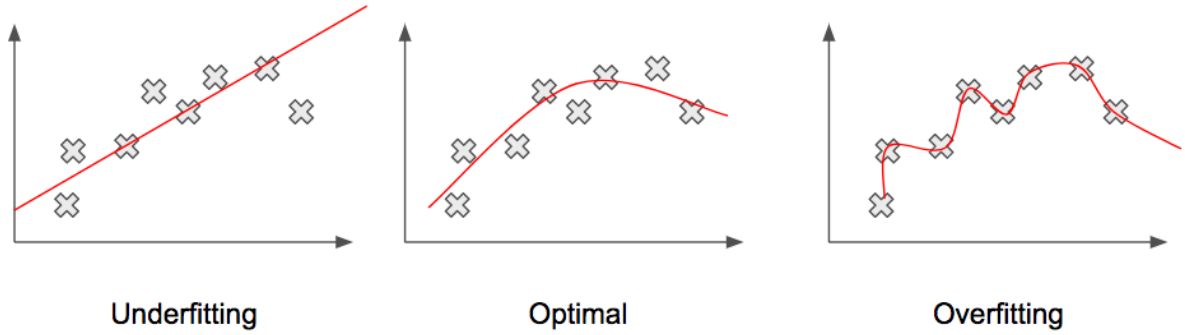
In the context of Machine Learning, the term *generalisation* refers to an AI model's "ability to make correct predictions on new, previously unseen data as opposed to the data used to train the model."[57] As seen above, in the case of supervised and unsupervised learning, new knowledge is learned based on previously collected data and these new insights are applied to make predictions. In reinforcement learning, the machine is allowed to learn from its own mistakes and the learning is applied to new situations. Thanks to their potentially super-human performances in specific areas, Machine Learning systems can convey the impression of being smarter than humans and infallible. However, it is important to emphasise that as with any form of analysis produced by humans or computers: **correlation is not causation, and prediction is not certainty**. Once trained, computers can make almost instant predictions but those predictions need to be verified. Although algorithms can uncover connections in the data, it is vital to ask whether those connections make sense.

[56] https://artint.info/html/ArtInt_267.html.
[57] https://developers.google.com/machine-learning/glossary.

The website *Spurious Correlations*,[58] run by Tyler Vigen, collects time-series data from various sources and creates graphs showing the correlation between two variables. Although some examples might be obviously humorous in nature due to the difference between the two datasets (one example links the divorce rate in Maine with the per capita consumption of margarine with over a 99% correlation computed), others might be misleading.

On a technical level, two problems arise when considering generalisation: *underfitting* and *overfitting*.

Figure 2.9: Generalisation problems: underfitting and overfitting



Source: https://pythonmachinelearning.pro/a-guide-to-improving-deep-learnings-performance.

Underfitting refers to situations in which a Machine Learning system is unable to capture the underlying information contained in the data. In such cases, the model produces poor predictions which can be observed when examining the different performance metrics, such as accuracy (see previous section). Underfitting is usually the result of the application of an inappropriate model for the problem at hand (i.e. the model is too broad or too simple with regard to the problem's complexity). In the case of deep learning, it may be useful in such situations to increase the number of nodes or add new layers in the neural network.[59] Otherwise, it may be necessary to attempt other techniques and compare the results.

Overfitting refers to cases where the algorithm is too specific to the extent that it captures and focuses overmuch on noise and anomalies. During the training phase, an overfitting model may achieve a high level of accuracy and problems may go unnoticed. However, once the trained model is exposed to new data, accuracy can drop severely. This reinforces the need for proper testing and validation before deploying an AI system and generalising the knowledge acquired.

Although these concepts may appear to be technical tinkering, they can have a huge impact on certain critical decisions. *Technology-based decision-making* may sound forward-thinking, but it can create a false sense of authority and confidence. It is incumbent on public sector stakeholders to ensure that this does not equate to *bad prediction-based decision-making*.

Bias, data and other security issues

Bias doesn't come from AI algorithms, it comes from people.

[58] www.tylervigen.com/spurious-correlations.
[59] www.mikulskibartosz.name/how-to-deal-with-underfitting-and-overfitting-in-deep-learning.

From a technical standpoint, it is important to distinguish between different types of *bias.* Much has been written about the ethics of AI, and Chapter 4 provides further details on the issue of data and ethics as well as ways for governments to tackle it.

When it comes to *AI bias* or *bias in algorithms*, it is important to distinguish *statistical bias*. This refers mostly to a model that consistently generates an error in prediction when compared with the expected outcome. For example, in the case of a house pricing AI model that predicts the value of a property based on available data but consistently over-prices by EUR 1 000, the error should be rectified before deploying the system.

A fundamental appeal of Machine Learning for decision makers is its ability to make predictions based on digital assets: data. But what should be done if the data itself – and not the model – are unfit? As seen in the earlier section on "Data as fuel for AI", crucial steps need to be taken to ensure data quality and representativity, and to allow the model to not only generate accurate predictions but also produce fair outcomes for citizens. These considerations would fall under another type of bias: *sampling bias* (i.e. bias in the process of collecting data). An example of this could be collecting data only for certain segments of the population.

In spite of efforts to curate data, real-life conditions can sometimes be unfavourable and AI systems can be subjected to malicious actions from malevolent people. In their white paper[61] on Artificial Intelligence, Luxembourg's finance regulator, the *Commission de Surveillance du Secteur Financier* (CSSF) highlights three such types of action: data poisoning, adversarial attack and model stealing. *Data poisoning* refers to the manipulation of data used for training, resulting in the AI system learning the wrong insights. This is especially relevant for AI systems that draw on data available online to continuously updating their training. For example, people may generate content on social media to interfere with the functioning of an AI system built to perform sentiment analysis in order to prevent it from making the correct predictions. As noted earlier, images can also be altered in ways not perceptible to the human eye and fed into an algorithm to make it misclassify new pictures. *Adversarial attack* is another type of security risk whereby attackers attempt to bypass detection from AI systems. For example, attackers may try to evade an AI-based spam filter by sending different kinds of email and probing for potential cracks in the system, then designing emails that can avoid the filter. Another concern for Machine Learning systems is the risk of *model stealing*. Here, the objective is for attackers to reverse engineer and duplicate the AI algorithm to obtain sensitive information. For example, attackers could try to recreate stock market prediction systems to benefit from the predictions. They might also be interested in learning which data were used to train a model and the knowledge they might acquire with this information.[62]

Interpretability, explainability

As noted earlier, deep learning is the subset of Machine Learning that currently holds the most promise for the future, producing better performance overall than any other technique. Unfortunately, such performance can be undermined by the lack of interpretability or explainability. In the case of deep learning, AI truly acts as a black

[60] https://towardsdatascience.com/what-is-ai-bias-6606a3bcb814.

[61] www.cssf.lu/fileadmin/files/Publications/Rapports_ponctuels/CSSF_White_Paper_Artificial_Intelligence_201218.pdf.

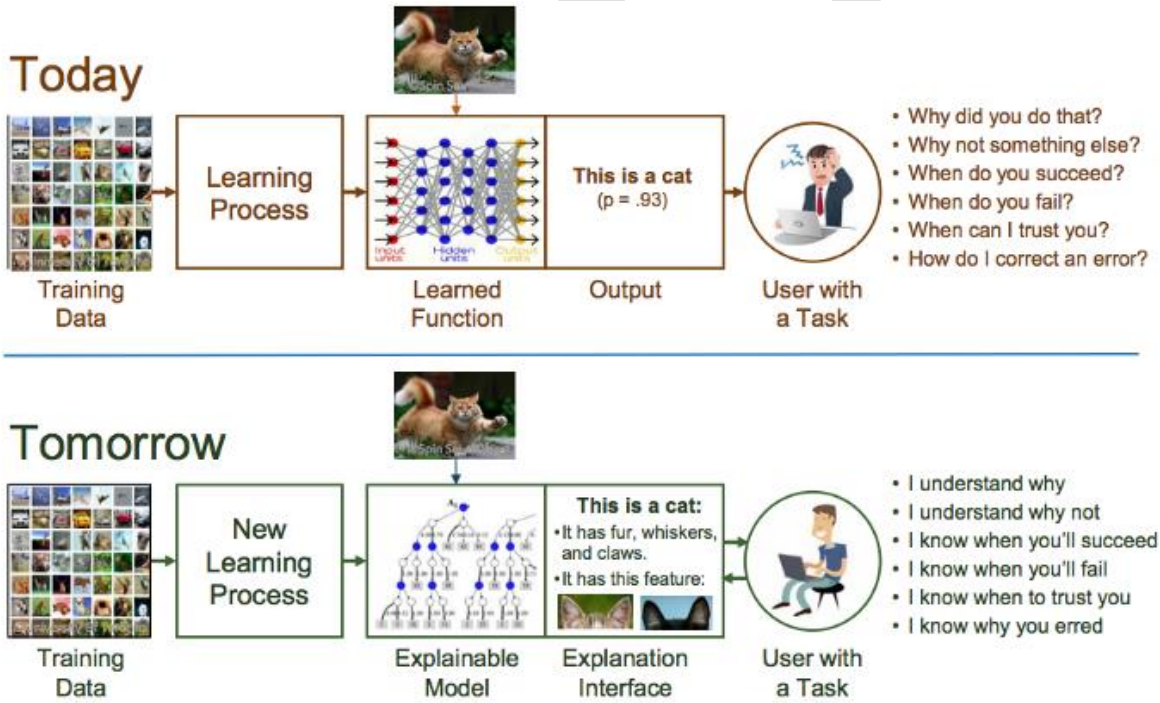
[62] More information on attacks against machine learning and defence strategies can be found at: https://elie.net/blog/ai/attacks-against-machine-learning-an-overview.

box: it is able to produce results but the process by which the results are produced and the reasons why the algorithm makes specific decisions are not understood.

In some cases, explainability may be of lesser concern, as the results themselves are more important than the process by which they were produced (e.g. correctly predicting whether a patient has a disease). However, in the case of public organisations, explainability is key as decisions made based on AI must fully be understood and explainable for reasons of accountability and transparency. Moreover, decisions made by public stakeholders may have a strong impact on the lives of citizens. This topic is discussed further in the context of public sector considerations in Chapter 4.

Figure 2.10 shows the approach taken by the US Defense Advanced Research Projects Agency (DARPA) to attempt to solve this problem by modifying the learning process to include a step that generates an easy-to-understand model. The figure suggests the use of decision trees to link the results produced with explanations.

Figure 2.10: Explainable AI: what are we trying to do?



Source: https://towardsdatascience.com/ai-policy-making-part-4-a-primer-on-fair-and-responsible-ml-and-ai-28f52b32190f.

Undermining humans

Finally, a common issue when considering Artificial Intelligence is to underrate the role played by humans in developing not just prediction machines but actual complete solutions to a given problem that may or may not include prediction machines. The case of Tesla is a good cautionary tale against over-reliance on AI and automation. In 2017, Tesla announced its plan to produce the new Model 3 car two years ahead of schedule thanks to a completely new approach to car manufacturing involving "hyper-automation" (i.e. a fully automated assembly line).[63] By 2018, Tesla had to revise its estimation and the underlying vision of a hyper-automated factory, with CEO Elon

[63] https://techcrunch.com/2019/03/05/elon-musk-wasnt-wrong-about-automating-the-model-3-assembly-line-he-was-just-ahead-of-his-time.

Musk recognising that excessive automation was a mistake.[64] This case is representative of common situations that arise when an organisation focuses overly on technical solutions and disregards the importance of knowledge possessed by field experts, especially in the performance of complex manual tasks, or the knowledge of civil servants who face specific issues in their day-to-day jobs.

[64] www.theguardian.com/technology/2018/apr/16/elon-musk-humans-robots-slow-down-tesla-model-3-production.

4. Emerging government practices and the global AI landscape

It is clear that AI is rapidly transforming many aspects of people's everyday lives, and that this transformation is accelerating at an exponential pace. The public sector is not immune, and in fact is charged with setting national priorities, investments and regulations when it comes to AI. Most relevant to this guide, governments are also in a position to leverage the immense power of AI to innovate and transform the public sector in order to redefine the ways in which it designs and implements policies and provides services to its people. Such innovation and transformation is critical for governments as they face ever-increasing complexity and demands from their citizens, residents and businesses.

AI can be integrated into the entire policy-making and service design process. As AI and Machine Learning technology evolves, more administrative and process-driven tasks will be able to be automated, boosting public sector efficiency and freeing up public servants to focus on more meaningful work. Governments will also be able to better understand and make decisions within their organisations and anticipate the needs of their people. If done well, automated processes can assist government to make decisions that are more fair and accurate than previously was the case.

This chapter discusses how governments around the world are adapting to the new possibilities and new realities presented by AI to transform government, and how they are building capacity to anticipate and prepare for where AI may take them in the future. It leverages and builds upon the forward-thinking work of the OECD Digital Government and Open Data Unit[65] and the OECD Working Party of Senior Digital Government Officials (E-Leaders),[66] as well as the work that the OECD has undertaken to develop the OECD Principles of Artificial Intelligence and the forthcoming OECD AI Policy Observatory.[67]

Government AI strategies

Government commitment to AI is reflected in several recent declarations signalling support for international collaboration. In 2018, all EU member countries signed the Declaration of Cooperation on Artificial Intelligence,[68] committing to work together to boost European AI capacity and adoption, address socio-economic challenges and ethics, and ensure an adequate legal and ethical framework. They also committed to making AI available and beneficial to public administrations, to sharing best practices in procuring and using AI in government, and to implementing open data practices. Ten governments[69] also signed the *Declaration on Artificial Intelligence in the Nordic-Baltic Region*,[70] pledging to, among others, improve skills development and access to data, and to develop ethical guidelines.

The most comprehensive and granular strategies, however, are found at the national level. Many countries worldwide have adopted national AI strategies or comparable guiding policies to set strategic visions and approaches to AI. These include AI-related

[65] www.oecd.org/governance/digital-government.
[66] www.oecd.org/governance/eleaders.
[67] http://oecd.ai.
[68] https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence
[69] Denmark, Estonia, Finland, the Faroe Islands, Iceland, Latvia, Lithuania, Norway, Sweden and the Åland Islands.
[70] www.norden.org/sv/node/5059.

priorities and goals and, in some cases, a roadmap for achieving them. Such strategies can help countries build a common foundation for success in their AI progress, as well as align the capacities, norms and structures of the relevant AI actors and ecosystems. Around the world, at least 38 countries (including the European Union) have developed, or are in the process of developing, a national AI strategy (see Figure 3.1). While this implies that a significant majority of countries are not yet planning a strategy, it does indicate that many countries now see AI as a national priority.

A number of common themes emerge when viewing these strategies as a whole. Nearly all of the countries have (or intend to have) a major focus on catalysing economic development through research and R&D funding. For instance, the European Union has called for the public and private sector to increase investments in AI by at least EUR 20 million by the end of 2020, and has sought to kick-start efforts by allocating EUR 1.5 billion in research funding. China has also pledged billions of euros (equivalent) in research funding for domestic projects. Similar funding efforts are taking place in many countries.

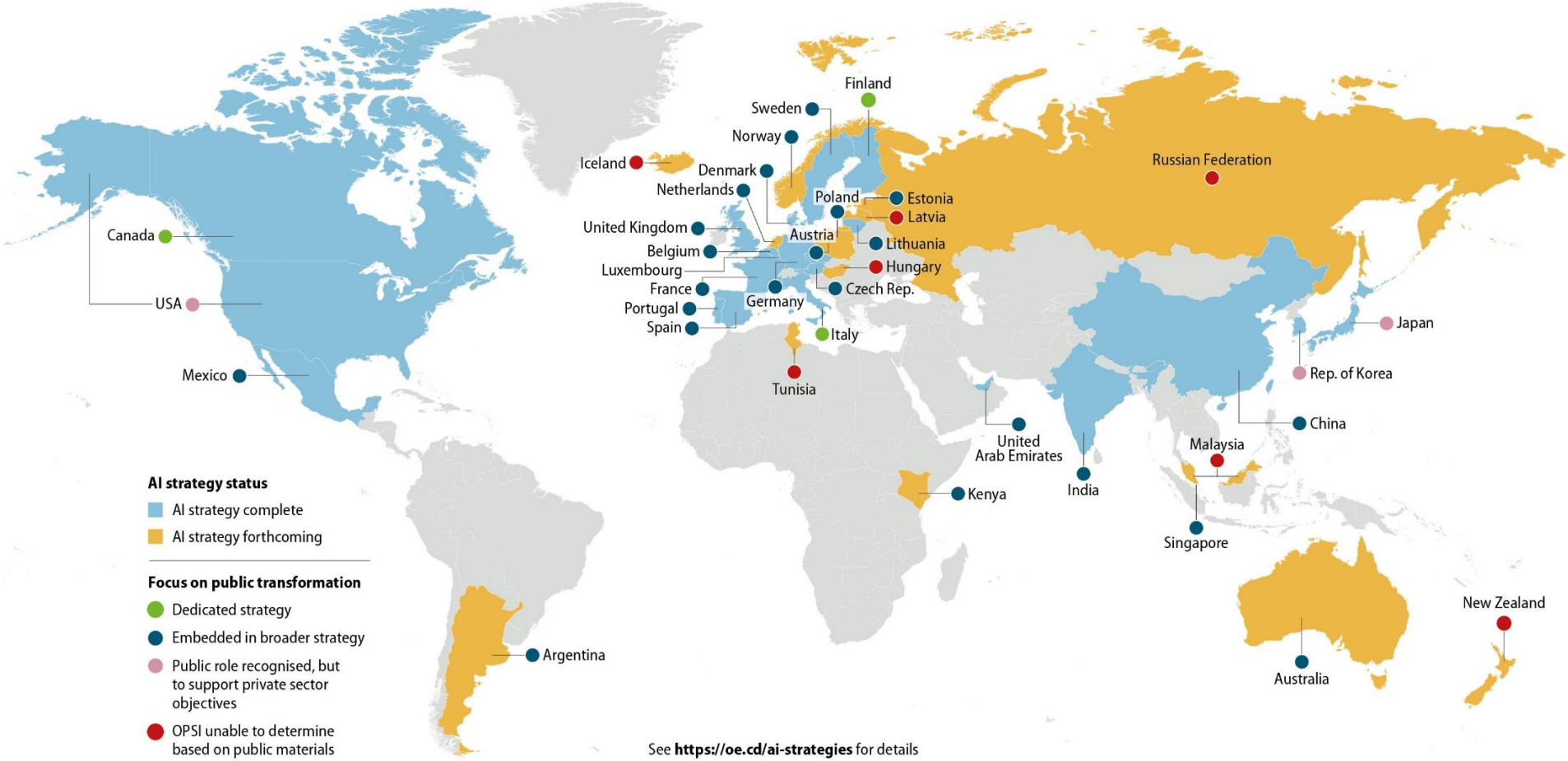
Most strategies also include provisions to help ensure that AI systems are designed and implemented in an ethical, trustworthy and secure manner. They generally also include elements to strengthen the national pipeline of AI talent, often through educational programmes and training. Most importantly for this guide, the majority include a specific focus on the use and implications of AI for innovation and transformation of the public sector.

Public sector components of national strategies

Out of 38 countries (including the European Union) with complete or forthcoming national AI strategies, 28 have either strategies in place for public sector transformation through AI or a dedicated focus embedded within a broader strategy.[71] In contrast, a number of strategies discuss the importance of government's role in AI, but only in the context of support for the broader economy. For several of the forthcoming strategies, the OECD Observatory for Public Sector Innovation (OPSI) was unable to determine whether the eventual strategy would focus on the public sector, due largely to a lack of public statements on their content. Annex A contains a case study on Finland's approach to AI, which includes a strategy that covers the broader economy, as well as a human-centric strategy specifically for public sector innovation and transformation.

[71] For forthcoming strategies, this is based on public statements regarding the expected contents of the forthcoming strategy. Details can be found at https://oecd-opsi.org/projects/ai/strategies.

Figure 3.1: AI strategies and the extent to which they include public transformation



Source: OPSI analysis of national strategies (see https://oecd-opsi.org/projects/ai/strategies). OECD (forthcoming), *State of the Art on Emerging Technologies in the Public Sector*.

As with the broader national strategies, a number of key themes emerge across the public sector-focused strategies. These include:

- experimentation with AI in government and the identification of specific AI projects currently underway or that will be developed in the near future
- collaboration across sectors, such as through public-private partnerships
- fostering of cross-government councils, networks and communities to promote systems approaches
- automation of routine government processes to enhance efficiency
- use of AI to help guide governmental decision-making (e.g. in policy evaluation, emergency management and public safety)
- strategic management, leverage and opening up of government data to develop tailored and anticipatory services, as well as to fuel AI in the private sector
- provision of guidance on the transparent and ethical use of public sector AI
- enhancement of civil service capacity through training, recruitment, tools and funding
- assurance that AI is used to augment, and not replace, human talent.

Similar to funding for broader R&D, governments and international bodies are carving out funding for projects that involve the public sector. For instance, the European Union has pledged EUR 2.5 billion for public-private partnerships,[72] and the governments of Finland, Portugal and Slovenia have each committed over EUR 10 million for public sector projects (OECD, forthcoming).

OPSI has developed a digital supplement to this report that discusses each country's complete or forthcoming national AI strategy, including the extent to which each agenda specifically addresses public sector innovation and transformation. The site also includes links to key strategy and policy documents. This resource can be accessed at https://oecd-opsi.org/projects/ai/strategies. In addition, the OECD (forthcoming) report on the *State of the Art on Emerging Technologies* provides information on the key actors involved in releasing these strategies.

AI projects with a public purpose

While governments are increasingly developing AI strategies, there is enormous potential for Artificial Intelligence to be applied across the public sector to improve how government engages with and serves its people.

Research indicates that some of the most immediate impacts of AI for the public sector will involve automating simple tasks and guiding decisions to make government more efficient and informed (Partnership for Public Service/IBM Center for the Business of Government, 2019). The OECD Digital Government working paper, the *State of the Art on Emerging Technologies in the Public Sector*, supports this finding and demonstrates how the use of AI can advance data-driven policy decisions, leading to better governance. The working paper also identified the "main applications areas" for government AI transformation – health, transportation and security (OECD, forthcoming).

OPSI's work has also found that AI is well suited to fostering positive relationships with citizens and businesses. Recent research showed that AI has significant cross-cutting

[72] http://europa.eu/rapid/press-release_IP-18-3362_en.htm.

potential to help achieve the Sustainable Development Goals (SDGs) (IDIA, 2019). This section explores these areas and provides a non-exhaustive set of examples of real-world government projects.

Improving government efficiency and decision making

In the context of government, one of the most important and most immediately achievable benefits of AI is to change the way that public servants themselves do their jobs. AI has the potential to help government shift from low-value to high-value work[73] and better focus on core responsibilities by "reducing or eliminating repetitive tasks, revealing new insights from data… and enhancing agencies' ability to achieve their missions" (Partnership for Public Service/IBM Center for the Business of Government, 2019).

The average civil servant spends up to 30% of their time on documenting information and other basic administrative tasks (Viechnicki and Eggers, 2017). By automating or otherwise avoiding even a fraction of these tasks, governments could save a tremendous amount of money, as well as re-orient civil servants' work around more valuable pursuits, resulting in more engaging jobs (Partnership for Public Service/IBM Center for the Business of Government, 2018; see Box 3.1).

Box 3.1: Eliminating tedious tasks at the United States Department of Labor (DOL)

Every year, the Bureau of Labor Statistics at the DOL is tasked with analysing hundreds of thousands of surveys related to workplace injuries and illnesses in businesses and public sector organisations across government. This analysis is important in both understanding these afflictions and in developing guidance that can help prevent them in the future. Bureau employees must learn a complicated coding system, read each report and code various characteristics, a process that is time consuming and monotonous, and takes up 25,000 employee hours each year.

Starting in 2014, the Bureau began to experiment with using AI to code surveys, starting with the easiest and most clear-cut responses. Over time, the use of AI increased and is now used on half of all surveys. The Bureau has found that AI can code as much in one day as a trained employee could do in a month, with a higher level of accuracy. Bureau leaders further found that it was important to actively communicate the benefits of AI to employees, emphasising that its purpose was not to replace them, but rather to allow them to focus on more complex and valuable tasks. The Bureau also provided training sessions for employees on Machine Learning, and how it can add value to their work.

Source: Partnership for Public Service/IBM Center for the Business of Government (2018), *The Future Has Begun.* www.businessofgovernment.org/sites/default/files/Using%20Artificial%20Intelligence%20to%20Transform%20Government.pdf.

As discussed in Chapter 1, a key factor that has resulted in growing interest in AI is the vast and increasing amount of available data. However, the large volumes of data involved can hinder governments from extracting useful knowledge, a phenomenon often referred to as "information overload". AI can help governments overcome information overload, gain new insights and generate predictions to help them make better policy decisions. Korea, for example, is using Machine Learning to identify

[73] www.whitehouse.gov/wp-content/uploads/2018/08/M-18-23.pdf.

opportunities across government ministries to catalyse innovation in the broader economy (Box 3.2).

Box 3.2: Korea's R&D Platform for Investment and Evaluation (PIE)

In Korea, government funding for R&D has grown steadily; however, this trend has not fully contributed to innovative economic outputs. The Ministry of Science and ICT has identified several key problems, including:

- R&D programmes are fragmented among 14 different ministries and agencies, and information sharing is limited.
- Basic, fundamental research is not connected to later stages of applied and commercial research and development.
- Regulatory barriers are not considered adequately at the development stage.
- The feedback cycle between evaluation and funding is often not well aligned.

To address these issues and make national R&D more sustainable and anticipate future challenges and opportunities, the Government of Korea is implementing a new innovation investment model: the "R&D PIE". This model pulls together data from multiple areas (e.g. academic research, patents, public and private tech trends, economic impact information, and other market information), and then applies Big Data analytics and Machine Learning to assess disruptive changes in the technology landscape, and identify overlaps, potential opportunities and missing links across Korean ministries, as well as stakeholders in the private sector and academia.

A separate R&D PIE platform with relevant data is provided for a number of strategic focus areas – autonomous vehicles, precision medicine, high-performance drones, air-pollution mitigation, smart farms, smart grids, intelligent robots and smart cities. Korea is also looking to expand R&D PIE into additional areas.

Through the use of R&D PIE, the government has found a way of identifying missing links in innovation initiatives, fostering collaboration among agencies, universities, and companies, and addressing social problems. By better understanding project potential, feasibility and potential future issues, the government is in a position to make more informed decisions about what to invest in, and what to avoid.

Source: https://oecd-opsi.org/innovations/rd-platform-for-investment-and-evaluation-rd-pie.

Healthcare

AI is already being used in the healthcare field in a number of ways, and its potential for future applications in the public sector is tremendous for countries that have national health services. As discussed in the *State of the Art* paper (OECD, forthcoming), AI applications, especially those involving Machine Learning, can help interpret results and suggest diagnoses, and predict risk factors to help introduce preventative measures. They can also suggest treatments and help doctors create highly individualised treatment plans. Combined with the knowledge of doctors and other medical experts, AI can lead to better accuracy, higher efficiency and more positive outcomes in the health field (see Box 3.3 and 3.4).

Box 3.3: United States Precision Medicine Initiative (PMI)

Launched in 2015, the PMI is a nationwide initiative to move away from the "one-size-fits-all" approach to health care delivery and to instead tailor treatment and prevention strategies to people's unique characteristics, including environment, lifestyle and biology.

Supported by the creation of "Next Generation DNA Sequencing" (NGS) technologies, precision medicine allows for detailed molecular characterisation of disorders and cancers via fast sequencing of patients' DNA at affordable cost. Machine Learning algorithms can accurately analyse the sequenced information and leverage the gigantic amount of data in an individual's medical records with direct benefit for the patient. This helps physicians to make better decisions and create more effective treatment plans.

Source: www.healthit.gov/topic/scientific-initiatives/precision-medicine, www.oecd.org/education/ceri/GEIS2016-MadelinReport-Full.pdf.

Box 3.4: Cancer detection through AI-enable image processing

Lung cancer is one of the leading causes of cancer-related deaths, and catching it early is crucial to treating the disease. Typical processes for diagnosing the disease have high rates of false positives and false negatives. Such errors can result in delays that prevent patients from receiving effective treatment.

Google and Northwestern Medicine, an academic medical centre in Chicago, collaborated to develop a "deep learning" AI algorithm to review image scans used to diagnose lunch cancer. The algorithm was then able to review scans independently to predict whether a scan indicated cancer. Researchers compared the predictions of the AI system with those of radiologists with significant experience in the field. In all cases, the AI system's predictions were as accurate as those of the radiologists. In some situations, the AI system outperformed the doctors.

Source: www.medicalnewstoday.com/articles/325223.php.

In another example, Mongolia is piloting a combination of AI and blockchain technologies to help identify counterfeit medicines before they make it into the hands of consumers. This case is covered in-depth in OPSI's report *Embracing Innovation in Government: Global Trends 2019*.[74]

Transportation

One of the most widely publicised usages for AI is autonomous vehicles, such as the self-driving cars being tested by Uber and a number of major motor companies. While government certainly has a role to play in regulating and understanding the implications of such vehicles, the opportunities they present for public sector innovation are less evident. Instead, governments are using AI to transform the ways in which they predict and manage traffic flows and handle potential safety issues.

[74] See https://trends.oecd-opsi.org.

Box 3.5: Government AI projects for transportation

Hangzhou, China

The city of Hangzhou, which has a metropolitan population of about 6 million, has partnered with tech firm Alibaba to launch the "City Brain" project. The initiative uses hundreds of cameras around the city to collect real-time data on road traffic conditions. These machine-readable data are then centralised and fed into to an "AI hub" which makes decisions affecting traffic lights at 128 city intersections. The system does not simply monitor and adjust traffic based on vehicle volume; it can also make more strategic decisions, such as identifying and clearing paths for ambulances on emergency calls, reducing their travel time by 50%.

Singapore

SMRT Corporation, a public transportation organisation in Singapore, has worked with private company NEC on a pilot using AI to predict the likelihood that public bus drivers would crash within the next three months. If the AI systems indicated a high chance of a crash for a driver, they are required to take a training course. The AI pilot used historical road performance data, and two data scientists observed bus driver behaviour in order to identify potential risk factors.

Source: https://trends.oecd-opsi.org; https://govinsider.asia/security/five-chinese-smart-cities-leading-way; www.theaustralian.com.au/business/technology/artificial-intelligence-to-predict-accident-risk-of-bus-drivers/news-story/4e7f8e6a4b7ac6e8715966a86284de16.

Security

Security is one of the main focus areas for governments exploring the use of AI. The term encompasses both physical security and cybersecurity, and can cover a broad swath of topics for which governments are responsible including law enforcement, disaster prevention and recovery, and military and national defence. The *State of the Art* paper (OECD, forthcoming) notes, for instance, that "in the field of surveillance, computer vision and natural language processing systems can process large amounts of images, texts and speeches, to detect possible threats to public safety and order in real time".

As an example of physical security, the Government of Canada's Transport Canada has piloted the use of AI to perform risk-based oversight by scanning pre-load, air cargo information to identify potential threats. Annex A presents a case study of their "bomb-in-a-box scenario" pilot. Another example is Queensland Fire & Emergency Services' use of Machine Learning to forecast the likelihood of major hazards (e.g. cyclones and fire) to help allocate their resources, as presented on OPSI's Case Study Platform.[75]

Law enforcement is another area where AI is growing. Facial recognition has been used in a number of cities around the world to help locate suspected criminals and counter terrorism. This practice can be highly controversial, however, as discussed in the next chapter. The International Criminal Police Organization (INTERPOL) is one entity using facial recognition and other types of AI for law enforcement, and has published *Artificial Intelligence and Robotics for Law Enforcement*,[76] which explores the potential of AI for policing and details real-world projects already underway.

[75] https://oecd-opsi.org/innovations/queensland-fire-emergency-services-futures-service-demand-forecasting-model.

[76] www.unicri.it/news/article/Artificial_Intelligence_Robotics_Report.

On the cybersecurity front, governments have been on the receiving end of crippling cybersecurity incidents in recent years. For instance, the US Office of Personnel Management (OPM) was the victim of a hack that resulted in the disclosure of critically sensitive information for over 21.5 million records, including detailed security-clearance background information and the fingerprints of 5.6 million public employees.[77] AI can assist government in monitoring network issues and detecting irregularities. Countries such as Thailand are also using AI cybersecurity tools, and others have published guidance for their use, as discussed in Box 3.6.

Box 3.6: AI for cyber security

Thailand

"Thailand is using AI to monitor network traffic and conduct big data analyses to detect suspicious user behaviour – for instance, two unusual logins with the same credentials, but hundreds of kilometres away."

United Kingdom

The UK National Cyber Security Centre has issued guidance on *Intelligent Security Tools* to help users understand considerations when using off-the-shelf AI security tools, and guide those seeking to build in-house AI security tools. It provides useful information on how to establish needs, deal with data, factor in available resources and get the most from AI. It presents a series of questions to help determine whether an AI solution is a good approach for a particular problem and set of needs.

Source: https://govinsider.asia/digital-gov/how-thailand-is-using-ai-for-cybersecurity; www.ncsc.gov.uk/collection/intelligent-security-tools.

Relationships with citizens and businesses

In addition to using Artificial Intelligence to address specific topics, governments are also utilising AI applications in a number of ways to engage with citizens, residents and businesses. One popular type of AI use in both the public and private sectors, especially in the early stages of an organisation's exploration of AI, is chatbots. Simple chatbots use a rules-based approach to interact with citizens in order to do things such as answer frequently asked questions. More sophisticated versions leverage Machine Learning to allow for more complex, less concrete interactions. The use of reinforcement learning (see Chapter 2) enables chatbots to continuously refine themselves to become more responsive to user needs (see Box 3.7).

[77] www.opm.gov/news/releases/2015/09/cyber-statement-923.

Box 3.7: UNA – Latvia's virtual assistant for the public administration

Latvia's Register of Enterprises has developed UNA, a 24/7 virtual assistant chatbot that provides answers in writing to frequently asked questions posed by current and future Latvian entrepreneurs, including status updates about submitted registration documents. UNA can be accessed through the Register of Enterprises website, as well as Facebook Messenger. It provides an alternative to an in-person visit or telephone call, and enables users to receive answers to questions at any time of the day.

UNA was developed by the government in co-operation with a private vendor. Register of Enterprises officials see UNA as a catalyst for change management, as it enables civil servants to delegate technical routine work and focus on higher value tasks. Employees are continuously teaching the AI system additional questions and answers to make it even more responsive. Since its launch in June 2018, UNA has answered over 22,000 questions from almost 4,000 users. Beyond answering client questions, Latvia is also exploring the use of UNA as a training tool for new employees.

Source: https://oecd-opsi.org/innovations/una-the-first-virtual-assistant-of-public-administration-in-latvia; www.ur.gov.lv/en/about-us/una, www.ur.gov.lv.

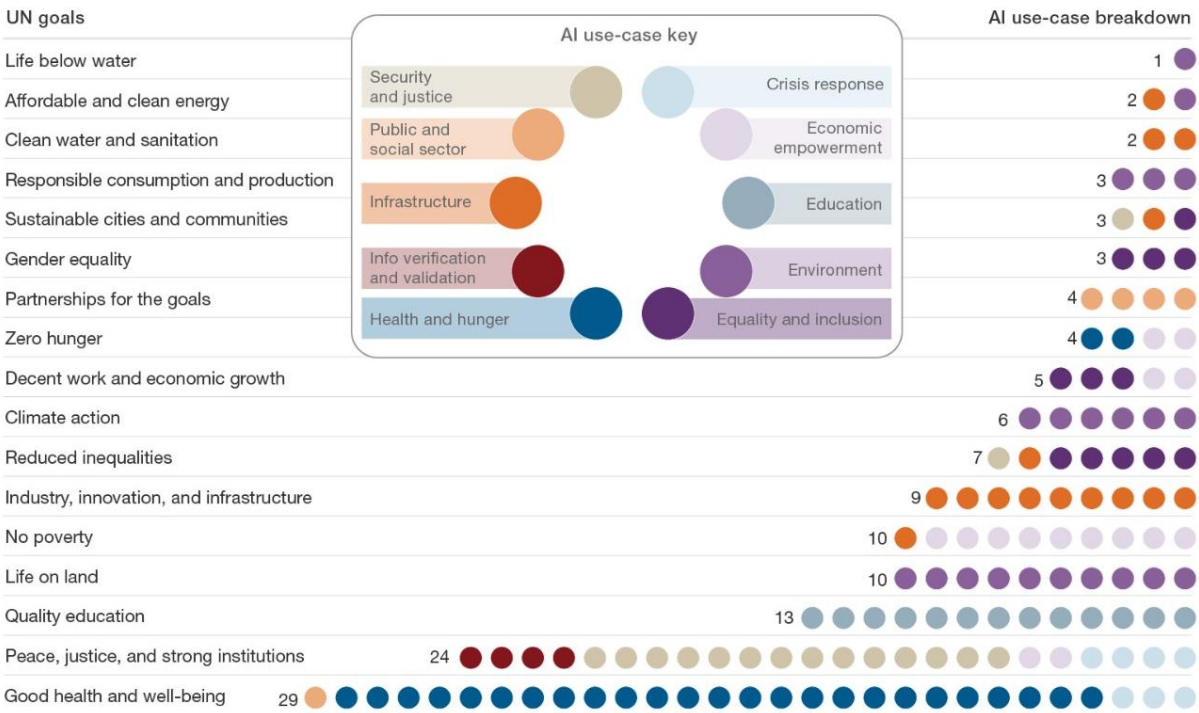
AI can also be used to help governments understand the opinions and perspectives of their citizens at scales that were previously not possible. For instance, the use of clustering Natural Language Processing and clustering techniques (see Chapter 2) enables governments to gain valuable insights into the views of their people. CitizenLab, a civil society organisation in Belgium, works with government to do just this (see the case study in Annex A).

Sustainable Development Goals (SDGs)

With the adoption of the 2030 Agenda for Sustainable Development, nations worldwide committed to a set of universal, integrated and transformational goals and targets, known as the Sustainable Development Goals (SDGs). The 17 goals and 169 targets represent a collective responsibility and a shared vision for the world. Governments are working to make progress to reach them by 2030, and many are exploring the potential of AI to help achieve this objective.

Research by McKinsey Global Institute (MGI, 2018) has identified a non-comprehensive set of about 160 cases that demonstrate how AI can be used for the "noncommercial benefit of society". Of these, 135 touch on one of the 17 SDGs (see Figure 3.2). These cases often take the form of private sector initiatives, or partnerships among the private sector, public sector and/or civil society. Interestingly, the research shows a concentration around SDGs such as "good health and well-being, and "peace, justice, and strong institutions", but little emphasis on goals such as "life below water", "affordable and clean energy" and "clean water and sanitation."

Figure 3.2: McKinsey-identified AI uses cases for SDGs



Source:
www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Applying%20artificial%20intelligence%20for%20social%20good/MGI-Applying-AI-for-social-good-Discussion-paper-Dec-2018.ashx.

In the public sector, the OECD has found that governments are pursuing uptake of AI geared towards preserving the environment, natural capital and climate resilience (OECD, forthcoing). These aims support a number of SDGs, such as: Clean Water and Sanitation (SDG 6), Affordable and Clean Energy (SDG 7), Responsible Consumption and Production (SDG 12), Climate Action (SDG 13), Life Below Water (SDG 14), and Life on Land (SDG 15). In its report *Harnessing Artificial Intelligence for the Earth*,[78] the World Economic Forum (WEF) explores the ways that AI can help address environmental challenges. Some public sector examples are discussed in Box 3.8.

[78] www3.weforum.org/docs/Harnessing_Artificial_Intelligence_for_the_Earth_report_2018.pdf.

Box 3.8: AI projects supporting environmental SDGs

Machine Learning for land mapping

In Australia, the Queensland Government Department of Environment and Science has adopted Machine Learning to automatically map and classify land use features in satellite imagery. Identifying different land uses (e.g. agriculture or housing) is crucial for conserving biodiversity, natural disaster monitoring, and biosecurity disease outbreak readiness and response. It can also be useful in providing a near real-time analysis of potential crops impacted during large disasters such as tropical cyclones and floods. The process is 97% accurate. With traditional manual methods, mapping land use for the whole state takes years, but the same process takes only six weeks with new technology.

Predicting energy consumption

Using Machine Learning clustering techniques, a research organisation in the United Kingdom leveraged data from digital electricity meters to develop an unsupervised AI model that could predict which types of appliances are likely to be used and when, thus predicting power consumption patterns. This information allows public utilities to predict future energy needs and enables residents to heat their homes in a smarter way, for example, by automatically turning off heating when they are likely to be away. Such optimised energy consumption can reduce both prices and energy waste.

Source: https://trends.oecd-opsi.org, www.gov.uk/government/case-studies/using-data-from-electricity-meters-to-predict-energy-consumption.

Governments are also broadly adopting, or planning to adopt, AI projects that support citizen-facing welfare services (OECD, forthcoming) and better lives for individuals (see Box 3.9). These aims cut across the SDGs for No Poverty (SDG 1), Zero Hunger (SDG 2), Good Health and Well-Being (SDG 3), Gender Equality (SDG 6) and Reduce Inequalities (SDG 10).

Box 3.9: AI projects supporting better lives

Welfare decisions

Denmark has plans to develop AI Machine Learning algorithms to help civil servants make decisions about whether citizens and businesses receive financial and other assistance from the government (e.g. support for older Danes, assistance for low-income families and housing assistance). The government believes that this technology can produce more accurate and objective decisions free from human bias. In addition, it can help address the challenge of an aging population, with only a limited number of civil servants available to process an increasing number of welfare requests. To make this possible, the government has focused on two specific challenges:

- How to put in place proper legislation to enable automated decisions.
- Making underlying data and decisions flows readable and understandable by machines.

Preventing slavery from space

The UK-based research laboratory Rights Lab recently launched "Slavery from Space", a project to end modern-day slavery. It uses Machine Learning algorithms that study high-resolution satellite data to estimate the number of brick kilns in South Asia's "Brick Belt" – an area where slavery is highly prevalent – thereby helping to calculate the extent of modern slavery in the region. Prior to this work, the full scale of brick kilns and, by proxy, slavery, was unknown, hindering action by the appropriate agencies. This innovation provides data to help NGOs and governments fight modern slavery. Using this technology, the Rights Lab team estimates that a third of slavery may be detectable from space.

Global Pulse projects

Global Pulse is the United Nations' flagship initiative on Big Data and consists of a network of innovation labs. Global Pulse is working to implement AI-driven speech-to-text analytics on local radio content to help understand local sentiments regarding refugee inflows. For instance, by analysing discussions on local radio, Machine Learning algorithms have uncovered valuable insights not previously gathered by other mechanisms. They have been able to identify small-scale disasters and their impact on the public, as well as surface areas of vulnerability for refugees. Global Pulse has a number of other projects underway that use AI to support SDG-related aims.

Source: https://govinsider.asia/innovation/exclusive-denmark-plans-to-use-ai-for-welfare-payments, https://oecd-opsi.org/innovations/slavery-from-space, https://rightsandjustice.nottingham.ac.uk, www.unglobalpulse.org/projects/pilot-studies-using-machine-learning-analyse-radio-content-uganda-2017, www.unglobalpulse.org/projects.

These are by no means the only ways that AI can support the SDGs and international development in general. A forthcoming report on *Artificial Intelligence in International Development* by the International Development Innovation Alliance (IDIA) provides additional discussion on this topic (IDIA, 2019).

Keeping up with advancements in public sector AI

The field of AI is advancing and growingly rapidly across all sectors, including the public sector, with new government strategies and projects being launched on a continuous basis.

While this chapter seeks to provide a current snapshot of national approaches and government trends in AI, the state of play will continue to change rapidly. To help public servants and other interested readers remain up to date, OPSI periodically updates its country-by-country overview of national AI strategies and public sector components, available at https://oecd-opsi.org/projects/ai/strategies.

While this chapter seeks to provide illustrative examples of specific AI projects, it is impossible to provide a comprehensive list, as new projects are being considered in governments on a daily basis. OPSI encourages public servants to keep up with the latest developments by accessing the following resources:

- OPSI's **Case Study Platform**[79] collects and shares hundreds of government innovations to help disseminate good ideas. Any public sector innovator may submit innovations to the platform. Of the over 300 cases currently on the platform, about 30 include an AI component.[80]
- The UN International Telecommunications Union (ITU) has developed a **Global AI Repository** of projects that promote progress towards the SDGs.[81]
- The OECD **Digital Government Toolkit** provides resources on good digital government practices by country, including many on managing data as an asset.[82]

Finally, while this chapter seeks to demonstrate that Artificial Intelligence *can* help promote innovation in government policies and services, it is important to note that AI is not the solution for every problem. Public officials and all levels must take into account numerous considerations when evaluating the use of AI. OPSI promotes experimentation with AI, as appropriate, and in an informed way. The next chapter discusses how this can be done and explores examples of current government experimentation.

[79] https://oecd-opsi.org/innovations.
[80] https://oecd-opsi.org/case_type/opsi/?_innovation_tags=artificial-intelligence-ai.
[81] www.itu.int/en/ITU-T/AI/Pages/ai-repository.aspx.
[82] https://oecd.org/governance/digital-government/toolkit/goodpractices

5. Public sector implications and considerations

As previous chapters have shown, there is a significant potential for the application of AI in the public sector. There are also many challenges and implications that government leaders and public servants need to consider when determining whether AI can help them address problems and achieve their missions. Building support will depend on setting a clear direction and narrative for use of AI in the public sector to better serve citizens and businesses. In fact, many governments are already publishing or developing AI strategies. Governments also need to ensure sufficient space for flexibility and experimentation to facilitate rapid learning.

Importantly, governments will need to develop ways to determine whether AI is the best solution for a given problem, and provide conduits for identifying and devoting attention to such problems. As many governments and international bodies have acknowledged, it is critical that they develop a trustworthy, fair and accountable approach to designing and implementing AI that identifies trade-offs, mitigates risk and bias, and ensures an appropriate role for humans.

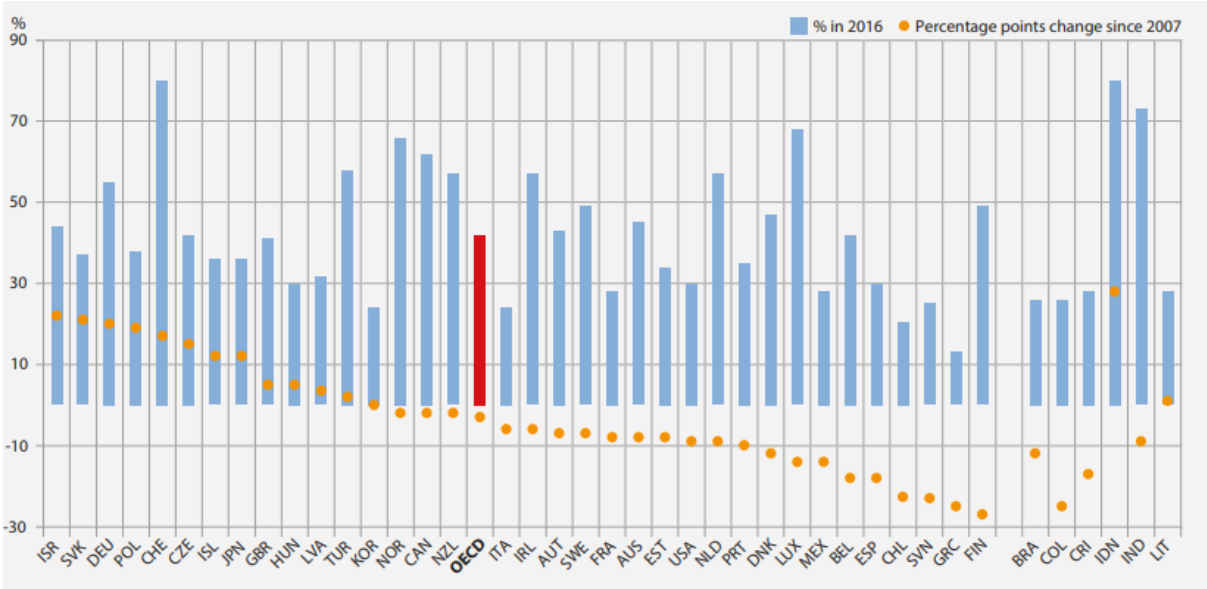
Governments must also consider the foundational elements that make AI-driven innovation possible. Data are the fundamental building blocks for AI, and a clear data strategy that enables governments to access robust, accurate data, in a manner that maintains privacy and conforms to societal and ethical norms, is necessary to effectively deploy AI. Governments will also need access to talent and essential products and services in both the public and private sector.

This section explores these issues with the aim of helping government leaders and civil servants to maximise the benefits of AI, learn from the actions of others and minimise potential risks. It concludes by setting out a framework to help them reflect on their approach to using AI for public sector innovation.

Provide support and a clear direction, but leave space for flexibility and experimentation

Faced with continued public demand and pressures on resources, Artificial Intelligence presents a significant opportunity to improve the productivity and quality of public services and government operations. However, low levels of trust in government (see Figure 4.1) emphasise the need for the public sector to set the right tone from the highest levels, and to take an approach that emphasises trustworthy, ethical and fair AI.

Figure 4.1: Trust in government has been declining, often from a low starting point



Source: Gallup World Poll.

Sustained and high-level political support will be necessary to create a stable, enabling environment for AI solutions to mature. The tone set by the highest levels of government has a crucial convening role in setting the direction of the technological development of AI and its use in wider society. This tone also sends signals to – and provides "top cover" for – public servants at all levels, enabling them to push for innovation and progress.

There are many possible trajectories for AI. Governments must ensure that it is used in a way that promotes and protects societal goals and values (Mateos-Garcia, 2018). Plans for deployment within government should also be consistent with – and support plans for – driving innovation through R&D, and promoting AI in the wider economy through infrastructure and skills investment, the wider regulatory environment and other industrial strategy policies. While a primary objective of governments may be to use AI to improve public services, they should also consider their role in shifting the technological frontier and adopting an "entrepreneurial state" approach to driving growth and innovation, using all the tools at their disposal to shape markets and take risks to achieve their vision (Mazzucato, 2013). The US Government, for example, provides such senior support by setting out its vision for maintaining AI leadership, as set out in Box 4.1.[83]

[83] Additional details on the US strategy and AI strategies from around the world can be found in a digital supplement to this report at https://oecd-opsi.org/ai.

Box 4.1: The President of the United States Executive Order on Maintaining American Leadership in Artificial Intelligence

The President's Executive Order offers an example of a clear vision for AI and how it will benefit US economic growth, security interests and the lives of its citizens. It articulates the objectives of maintaining US global leadership and ensuring that AI evolves "in a manner consistent with our Nation's values, policies, and priorities".

It then explains how the levers of the Federal Government will be used to achieve these goals through:

- promoting and funding R&D to drive technological breakthroughs
- developing appropriate technical standards and encouraging experimentation to increase AI deployment
- creating the skills to develop and apply AI technologies, including among the Federal workforce
- fostering trust in AI by ensuring that it protects privacy and individual freedoms
- generating an international environment that creates markets for US AI firms and protects US security and strategic interests.

Source: www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence.

In addition to having senior political support, governments will need to articulate a compelling vision for how AI can transform public services and operations to benefit citizens and businesses while maintaining public trust. Chapter 3 notes that most of the countries with national AI strategies include an explicit focus on AI for public sector innovation and transformation, with a few even having an explicit strategy dedicated to government. For example, Finland's AuroraAI Strategy (see case study in Annex A) clearly articulates an ambitious goal of developing a human-centric society based on the holistic welfare of its people, businesses and society as a whole. However, most countries lack an AI strategy or public sector-focused approach. Developing strategies and prioritising practical use cases that demonstrate how AI can improve services for citizens can create the basis for public support.

Each national strategy and approach must operate within its own unique context and its own culture and norms. Governments should engage with citizens and businesses in deliberative dialogue to more clearly understand their perspectives and values (Balaram, Greenham and Leonard, 2018). In particular, users of public services may want meaningful engagement and assurances on how the use of AI will impact the services on which they depend.

Similarly, securing and maintaining support will require a clear narrative explaining how AI can assist public sector employees to better deliver services, reduce the amount of time they spend on routine tasks and allow them to focus on higher-value tasks where they can have the most impact.[84] Resistance among public sector workers will slow the deployment of AI, limit its effectiveness and damage morale. Making a case for AI based on its potential for reducing employee numbers is unlikely to garner support and is not credible, as it is unlikely that AI will replace public sector workers in the short

[84] www.digital.nsw.gov.au/digital-transformation/policy-lab/artificial-intelligence.

term. The Canadian Digital Academy (see Box 4.16) offers an example of an innovative approach to boosting public servants' knowledge of AI.

It is important to note that individual governments do not need to handle every aspect of developing robust agendas and ecosystems for AI. Instead, they can take advantage of opportunities to collaborate internationally on AI approaches and standards (Mateos-Garcia, 2018). Many governments are grappling with the same issues related to AI and major opportunities exist to work together to address them and explore common standards and collaborative approaches. The OECD Principles on AI (Box 4.2) offer the world's first set of international standards agreed by governments. Similarly, the "Ethics Guidelines for Trustworthy Artificial Intelligence" (see case study in Annex A) articulate a series of principles for fostering and securing robust and ethical AI.[85]

Box 4.2: The OECD Principles on Artificial Intelligence

The OECD Principles on Artificial Intelligence support AI that is innovative and trustworthy and that respects human rights and democratic values. OECD member countries adopted the principles on 22 May 2019 as part of the OECD Council Recommendation on Artificial Intelligence. The principles set standards for AI that are sufficiently practical and flexible to stand the test of time in a rapidly evolving field. They complement existing OECD standards in areas such as privacy, digital security risk management and responsible business conduct.

The Recommendation identifies five complementary, values-based principles for the responsible stewardship of trustworthy AI:

- AI should benefit people and the planet by driving inclusive growth, sustainable development and wellbeing.
- AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.
- There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.
- AI systems must function in a robust, secure and safe way throughout their life cycles, and potential risks should be continually assessed and managed.
- Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles.

Source: https://oecd.ai.

Experimentation and iterative learning are crucial to developing AI capabilities. If practitioners do not have the freedom to try new ways of developing and delivering services, it is unlikely that the potential for AI in public services and operations will be realised. However, adopting an experimental approach to AI use may counteract efforts to put in place robust systems and consistent processes across government. On the other hand, deployment of AI systems is likely to be slow if decision makers delay until ideal governance frameworks and standards are in place. In short, governments need to carve out time and space for experimentation, as New Zealand has done (see Box 4.3); otherwise, AI may not be prioritised over urgent day-to-day pressures. Without open

[85] https://ec.europa.eu/knowledge4policy/publication/ethics-guidelines-trustworthy-ai_en.

experimentation and learning, there is a risk that unethical or careless practices will become entrenched and normalised, leading to sub-optimal, or even dangerous, long-term trajectories.[86]

Box 4.3: The New Zealand Service Innovation Lab and the Better Rules Initiative

The Service Innovation Lab is an all-of-government neutral space that enables public sector organisations to collaborate on innovations to facilitate public access to government services. While not focused strictly on AI, it serves as a design and development lab to experiment, drive and enable systemic change in government for the benefit of society, focused on the needs of the user. The lab also works to direct public funding towards systemic improvements, horizontal efforts around shared goals, high-value reusable components and actionable innovation for all participating public sector organisations.

The Service Innovation Lab collaborates with agencies and partners across New Zealand to promote greater innovation throughout the public service. While not focused on AI, the Lab provides an example of a cross-agency working to experiment, address systemic barriers to innovation, and prototype new approaches to integrated service delivery that are designed around user needs. It therefore offers an example of how governments can adopt an agile and adaptive approach to systemic innovation.

As an example, the Lab's Better Rules project re-writes laws as machine-consumable code to help ensure proper implementation and develop real-time feedback loops between legislative design and the implementation process. By serving as a machine-readable source of truth, such code can serve as a foundation for AI models and algorithms. If laws change, such changes can be immediately and accurately reflected in the algorithm to help ensure correct implementation.

Source: https://oecd-opsi.org/innovations/the-service-innovation-lab, https://trends.oecd-opsi.org.

The rapid pace of technological change means that governments need to take an agile and adaptive approach in order to adjust to new opportunities and changing behaviours. Because "there is no upper limit to how smart AI can become", tasks that AI cannot deliver effectively today will become feasible in the future. AI strategies and frameworks must be flexible enough to evolve with changing capabilities and contexts. AI technology is dynamic and how it interacts with humans in the complex systems of public service delivery will evolve over time (Kattel, 2019). Governments should therefore avoid long-term contracts that lock the public sector into current or proprietary technologies and ways of working. For example, the UK Government's Service Manual argues that when choosing a technology, "the most important thing is to make choices that allow you to: change your mind at a later stage [and] adapt your technology as your understanding of how to meet user needs changes".[87] Similarly, strategic plans must be living documents with regular reviews to monitor implementation and assess whether planning assumptions still hold. This viewpoint is at the heart of *Artificial Intelligence: A Strategic Vision for Luxembourg*,[88] a living strategy produced by the Government of Luxembourg that will be updated regularly. In France, the government's Etalab has produced guidance on the use of algorithms for public administrations (see Box 4.10),

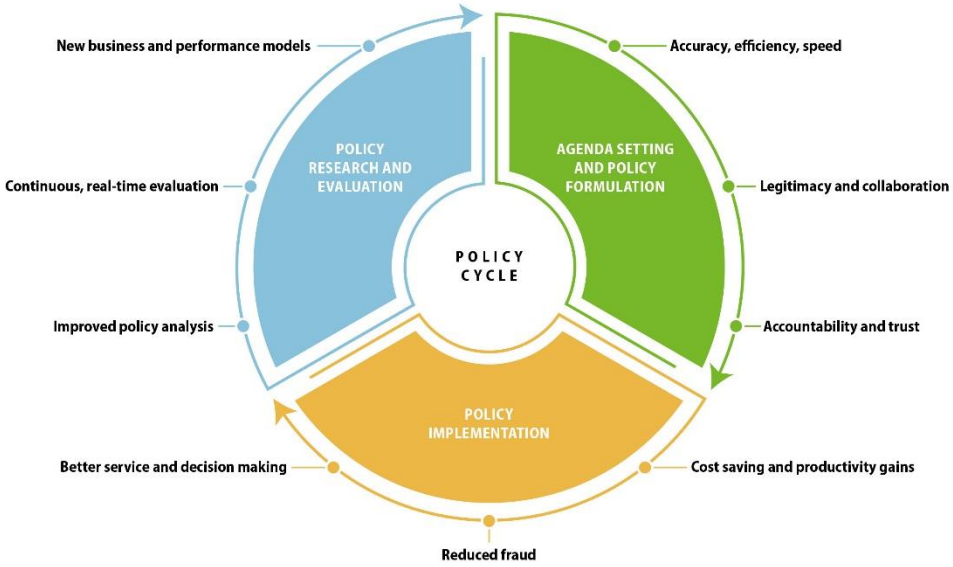
[86] www.slideshare.net/JuanMateosGarcia/d4p-complex-economicsaiv2.
[87] www.gov.uk/service-manual/technology/choosing-technology-an-introduction.
[88] https://digital-luxembourg.public.lu/initiatives/artificial-intelligence-strategic-vision-luxembourg.

an editable version of which is stored on GitHub, where contributors can make amendments to improve the content.[89]

Is AI the best solution to the problem?

A common problem with emerging technologies is the risk that people start with solutions and *then* look for problems for the technology to solve. If AI is to achieve its potential in the public sector, governments must focus first on the outcomes that governments and citizens want to achieve, and *then* identify whether AI (or something else) is the best solution to help achieve these goals (Mulgan, 2019). AI can identify patterns or irregularities in data to improve the accuracy of decision making, better allocate resources, anticipate unmet needs, spot fraud or safety risks, among many other things. These capabilities allow it to make a positive contribution to government activities throughout the policy cycle, from agenda-setting and policy formulation, to implementation and evaluation (see Figure 4.2).

Figure 4.2: Benefits of AI at each stage of the policy cycle



Source: Pencheva, Esteve and Mikhaylov (2018), *Big Data and AI – A Transformational Shift for Government: So, What Next for Research?* https://journals.sagepub.com/doi/pdf/10.1177/0952076718780537.

The first stage in effectively assessing the appropriateness of AI is diagnosis and problem definition. This process generally starts by breaking down the relevant activities or services into their constituent tasks, and identifying whether these can be more effectively delivered by AI. Automation of tasks can then be prioritised on the basis of the biggest impact on service cost-effectiveness. Well-designed AI is likely to perform prediction better than humans in cases where factoring in complex interactions between many indicators improves prediction; where there is a large volume of stable, representative data (allowing interactions to be a good predictor of future events); and where predictions are routine rather than rare (Agarwal, Gans and Goldfarb, 2018).[90]

However, AI may not be the optimal technological solution for many or even most problems. Careful analysis of the capabilities of specific AI tools is necessary to

[89] www.etalab.gouv.fr/algorithmes-publics-etalab-publie-un-guide-a-lusage-des-administrations.
[90] https://faculty.ai/products-services/ai-strategy.

determine whether they should form part or all of the solution to a specific challenge. Chapter 2 sets out the capabilities of a number of AI tools and explains what types of problems they might help address. For many public sector digital challenges, the most appropriate solutions are often simple but effective uses of existing technologies and improved interoperability, including with legacy systems. For example, UK start-up Accurx originally set out to use Machine Learning to improve the effectiveness of the prescription of antibiotics (e.g. to help prevent antibiotic resistance), but found that a more effective business model involved the use of text messaging to increase the number of patients attending doctor appointments (Lewin, 2019).

A rigorous focus on using AI only when it is likely to provide the optimal solution to a specific problem will reduce the risk of inappropriate adoption in areas where it will not add value. The UK Government has produced guidance on assessing whether AI is the right solution (see Box 1.4).

Box 4.4: UK Government guidance on how to assess whether AI is the right solution

The UK Government has created guidance for officials to help them to determine whether AI will help them meet users' needs. They recommend considering if:

- the available data contain the information required
- it is ethical and safe to use the data and consistent with the Government's Data Ethics Framework
- there is a sufficient quantity of data for the AI to learn from
- the task is too large and repetitive for a human to undertake without difficulty
- the AI will provide information a team could use to achieve outcomes in the real world.

The guidance then recommends assessing the current level of skills and the existing data stack, selecting the AI tool most appropriate to address the issue at hand, and then deciding whether to build or purchase the solution.

Source: www.gov.uk/guidance/assessing-if-artificial-intelligence-is-the-right-solution; www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework.

Successful AI strategies require the development of mechanisms that provide conduits to raise or identify specific operational problems which AI has the potential to address. Governments can adopt a number of different approaches to match resources to problems. The following are two opposing approaches:

- **Decentralised, demand-driven approaches**. Entrepreneurial managers or line staff in operational roles identify problems that AI can help address and bring in embedded experts to drive service transformation. This approach would facilitate problem-driven iterative adaptation but would not necessarily lead to effective prioritisation or consistent approaches across government (Andrews, 2018).
- **Centrally driven transformational leadership**. Potential AI applications are mapped across government and expertise and attention is oriented towards areas and problems deemed most likely to benefit from AI. This would enable consistency, prioritisation and systems approaches, but could lead to service managers adopting AI as a solution rather than focusing on problems and missing opportunities better perceived at the ground level.

Intermediate solutions that address some of the weaknesses of these two options include:

- Centrally determined missions or challenges to which experts both inside and outside government can pitch solutions.
- Promoting and allocating resources to communities of interest or networks of practitioners, enabling them to collaborate and share expertise across organisational boundaries.
- Building up central funds or teams of AI experts and then encouraging service managers to identify fruitful areas for AI exploration and bid for their time or resources.

Examples illustrating these approaches are discussed in Box 4.5.

Box 4.5: Government strategies linking key challenges to technological solutions

Missions and grand challenges

The UK Government created a GBP 20 million GovTech challenge to incentivise tech firms to deliver innovative solutions to public sector problems. Such mission-oriented approaches encourage small, emerging technology businesses to create and develop innovative solutions for public services. Once proven, the solutions can be scaled to match the market and society.

Five different challenge competitions awarded funding in the first GovTech Catalyst round: automating the identification and cataloguing of Daesh still imagery propaganda online, tracking waste through the waste chain, tackling loneliness and rural isolation, cutting traffic congestion and deploying smart sensors on council vehicles to improve services. Out of these, five companies working across the United Kingdom were awarded up to GBP 80 000 to develop innovative digital solutions to tackle the challenge of tracking waste from its source through to treatment and final disposal. The technological focus of the fund is broader than Artificial Intelligence, but one of the five successful companies employs AI in its solution.

Communities of interest and networks

The Office for AI and the Government Digital Service in the UK Government have developed guidance for service managers on how to assess, plan and manage AI in public services and administration.

The United States Emerging Citizen Technology Office (ECTO) works with public servants across government agencies, as well as businesses and civic organisations, to develop government-wide public service modernisation initiatives. These assess potential use cases and work with partners to develop "shared resources for the potential adoption of the technology".

Central funds or teams with bottom-up proposals

The Estonian Government's AI Expert Team has analysed their existing legal framework to ascertain whether it provides sufficient clarity and protection in the context of AI, and developed an action plan to promote the use of AI across government.

The US Government's Technological Modernization Fund (TMF) is a new funding model for technology modernisation projects. Government agencies can submit proposals for funding and technical expertise to a TMF Board consisting of senior government IT leaders. Proposals are assessed on:

- their impact on the agency mission (improving outcomes for users and security)
- feasibility (including agency capability)
- generation of opportunities (potential cost savings and service quality improvements)
- common solutions (replacement of insecure, outdated systems with scalable platforms that could be used by other organisations).

The Fund enables the government to focus efforts on areas where they can achieve maximum public benefit by prioritising technology solutions to improve delivery of mission-critical services and projects that can serve as common solutions and/or

inspire reuse. While its remit is broader than AI, US officials have encouraged agencies to submit proposals for modernisation projects driven by emerging tech.

Source: www.gov.uk/government/news/smart-tracking-of-waste-across-the-uk-govtech-catalyst-competition-winners-announced, www.gov.uk/government/collections/govtech-catalyst-information, https://emerging.digital.gov/TMF, https://tmf.cio.gov, https://investinestonia.com/artificial-intelligence, www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector, https://emerging.digital.gov/what-we-do.

Working out whether AI is the best solution to a policy problem is an inherently multi-disciplinary process: it requires consideration of technological, legal and ethical policy issues and constraints in a "common theoretical framework". Clearly, an AI solution needs to be technologically feasible, but equally it needs to be acceptable to a range of stakeholders (including the public) and permissible under the law. If the AI solution is deliverable and acceptable, then governments must assess whether it is the optimal means to achieve policy goals and generate public value Wingfield et al. (2016). The New South Wales Government in Australia proposes three questions that government agencies should ask themselves when considering the adoption of AI technologies (see Box 4.6).

Box 4.6: New South Wales Government's key questions on AI technology adoption

- Is it viable? You should understand the scope and the limits of the technology and then assess if the solution is viable.
- Is it valuable? Just because something can be automated does not mean that it should. How valuable would automation be? Would it deliver value to the community, and not just to your organisation's operations? What would the knock-on effects be? Can you make the outcomes fair and ethical?
- Is it vital? Is your proposed implementation unworkable without AI?

Source: www.digital.nsw.gov.au/digital-transformation/policy-lab/artificial-intelligence.

This multi-disciplinary approach should continue if AI is determined to be the optimal solution. Effective design of AI-enabled services at the operational level will require the technical expertise both of line staff and programme managers who understand the specifics of the service being delivered and how an AI will affect the overall workflow.[9192] This will allow them to maximise the transformative impact of AI by identifying tasks that are no longer required, new tasks that are needed, and the implications for service design and skills requirements in the workforce (Agarwal, Gans and Goldfarb, 2018). OPSI research has shown that multi-disciplinarity is one of the most critical factors for the success of innovation projects, especially those involving tech. It recommends that "at the outset of any innovation project, governments should convene a group consisting of the skilled individuals necessary to make the project a success. Such individuals could include policy analysts and advisors, field experts, user-experience designers, software developers and attorneys."[93]

[91] www.pwc.com/us/en/services/consulting/library/artificial-intelligence-predictions/functional-specialists.html.
[92] Most benefits are likely to be achieved at the operational policy implementation level, rather than the strategic level, though AI may have the cumulative impact of facilitating new strategic approaches to service delivery. For more information, see: https://journals.sagepub.com/doi/pdf/10.1177/0952076718780537.
[93] https://trends.oecd-opsi.org.

Develop a trustworthy, fair and accountable approach

A number of factors go into developing a trustworthy, fair and accountable approach. These are discussed in the subsections below.

Establish legal, ethical and technical frameworks at the design stage and monitor compliance with them during the implementation phase

Artificial Intelligence is a general purpose technology with the potential to have a transformative impact on how public services are delivered and administrations perform. This disruption means that AI trajectories are defined by complexity, uncertainty and risk (Mateos-Garcia, 2018). As such, the development of rigorous frameworks to shape decision-making in public sector organisations will be crucial for realising AI's potential to transform public services and administration. As discussed above, articulating clear principles for AI helps to bring about a conducive environment that is generally aligned with the societal goals and values articulated in the principles. Committing to ethical principles is likely to be a necessary but not sufficient condition for effective deployment of AI. If principles are to have maximum impact on behaviour, they will need to be actionable and embedded in the processes and institutions that shape decision making within government. Researchers at the Oxford Internet Institute[94] and the Alan Turing Institute have worked with other institutions to synthesise the ethical principles, as well as the underlying factors and corresponding best practices for AI to help actualise them (see Box 4.7).

[94] www.oii.ox.ac.uk.

Box 4.7: An ethical framework for a good AI society (AI4 People) and AI for Social Good (AI4SG)

Luciano Floridi and Josh Cowls collaborated with other researchers to develop the following synthesis of existing expressions of ethical principles for AI produced by reputable organisations:

- **Beneficence** – promoting well-being, preserving dignity and sustaining the Planet
- **Non-maleficence** – privacy, security and "capability caution" (do no harm and avoid misuse/overuse of technology)
- **Autonomy** – the power to decide or whether to decide (humans should always retain the power to decide which decisions to take, exercising the freedom to choose where necessary, and ceding it in cases where overriding reasons, such as efficacy, may outweigh the loss of control over decision-making).
- **Justice** – promoting prosperity and preserving solidarity (fairness, non-discrimination and ensuring the benefits are broadly shared).
- **Explicability** – enabling the other principles through intelligibility and accountability.

In a separate paper, they draw on a range of case studies to set out the essential factors that underpin the design of successful AI for Social Good (AI4SG) systems.

Factors	Corresponding best practices
Falsifiability and incremental deployment	Identify falsifiable requirements and test them in incremental steps from the lab to the "outside world".
Safeguards against the manipulation of predictors	Adopt safeguards which: (i) ensure that non-causal indicators do not inappropriately skew interventions; and (ii) limit, when appropriate, knowledge of how inputs affect outputs from AI4SG systems, to prevent manipulation.
Receiver-contextualised intervention	Build decision-making systems in consultation with users interacting with and impacted by these systems with understanding an of users' characteristics, the methods of co-ordination, the purposes and effects of an intervention, and with respect for users' right to ignore or modify interventions.
Receiver-contextualised explanation and transparent purposes	Choose a Level of Abstraction for AI explanation that fulfils the desired explanatory purpose and is appropriate to the system and the receivers, then deploy arguments that are rationally and suitably persuasive for the receiver to deliver the explanation, and ensure that the goal (the system's purpose) for which an AI4SG system is developed and deployed is knowable to receivers of its outputs by default.
Privacy protection and data subject consent	Respect the threshold of consent established for the processing of datasets of personal data.
Situational fairness	Remove from relevant datasets variables and proxies that are irrelevant to an outcome, except when their inclusion supports inclusivity, safety or other ethical imperatives.
Human-friendly semanticisation	Do not hinder the ability for people to semanticise (i.e. to give meaning to and make sense of) something.

Source: https://ssrn.com/abstract=3388669, www.research.ed.ac.uk/portal/files/77587861/FloridiEtalMM2018AI4PeopleAnEthicalFrameworkFor.pdf.

The Government of Canada's Directive on Automated Decision-Making seeks to operationalise a set of legal, ethical and technical principles to ensure standards and a consistent approach to risk management in AI across the public sector, both in the design

and the implementation stage. To accompany the Directive, the Government of Canada developed an Algorithmic Impact Assessment that assesses the potential impact of an algorithm on citizens. This provides granular, risk-based actions that enable officials to focus on putting in place effective mitigation where risks are highest. The Directive and the Assessment are discussed in-depth in a case study in Annex A. The case study on the European Commission's Ethical Guidelines for Trustworthy AI also provides considerations for assessing ethical issues.

Monitoring during the implementation stage will be needed to ensure that the system is operating as intended, that risks are being mitigated and that unintended consequences are identified. A differentiated approach will be required to focus attention on AI systems where the risks are highest, for instance, where they influence the distribution of resources or have other significant implications for citizens (Mateos-Garcia, 2017). The New York City approach to monitoring its use of AI is set out in Box 4.8.

Box 4.8: The New York City Automated Decision Systems Task Force

The Mayor of New York announced the creation of a task force to monitor the city's use of AI in order to ensure accountability, equity and fairness across all of its areas of responsibility. By December 2019, the task force will recommend procedures for reviewing and assessing AI tools to ensure equity and opportunity. The aim is to promote transparency and consistent adherence to common standards and values. The task force will comprise officials responsible for services, academics, legal and technology experts, civil society groups and think tanks.

Source: www1.nyc.gov/office-of-the-mayor/news/251-18/mayor-de-blasio-first-in-nation-task-force-examine-automated-decision-systems-used-by.

Clarify the appropriate role for humans in the decision-making process

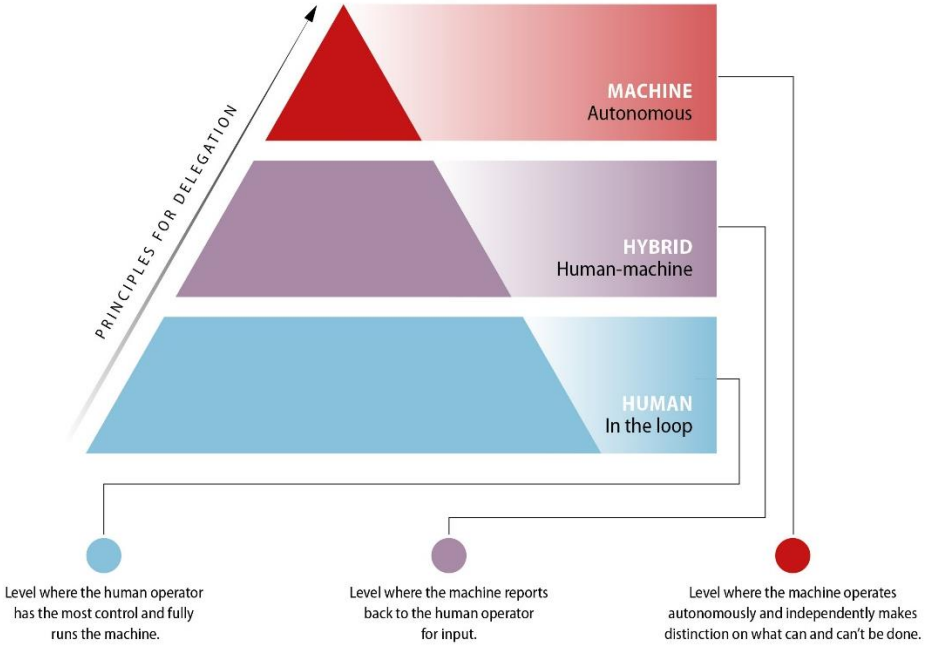
In many, if not all, cases, governments will want a human "in the loop", particularly when a new system is being deployed. In such cases, it will be crucial that the officials working alongside the AI system are clear about their precise role in the decision-making process. A framework for reflecting on the different levels of human-machine interaction is set out in Box 4.9. The officials will need to possess the necessary knowledge and skills to understand how the AI system functions, and its strengths and weaknesses, so that they can monitor it effectively and spot anomalies. They must also be certain of their own level of decision-making authority. Effective delivery will also require systems to be technically robust, safe and secure.[95]

[95] https://ec.europa.eu/knowledge4policy/publication/ethics-guidelines-trustworthy-ai_en.

Box 4.9: The AI-Human Pyramid of Interaction

As part of their AI Initiative, the Future Society (TFS) at the Harvard Kennedy School of Government (HKS) has developed a simple framework for understanding the nature of AI-human interaction in the context of armed conflict.

The AI-Human Pyramid of Interaction enables public sector leaders to assess the nature of these interactions in the AI systems they are responsible for. This constitutes a first step in determining whether they are appropriate given the associated costs, benefits and risks.



Source: http://ai-initiative.org/wp-content/uploads/2016/08/AI_MSC.-FINAL.pdf.

It is important for public officials to understand that putting in place effective controls will not reduce risk to zero. Algorithms need to be trained to provide a viable service and the imperative to experiment suggests that there is always a chance that an AI will not perform as intended. Even an unbiased algorithm is unlikely to be 100% accurate. However, it is also important to consider the counterfactual: postponing AI deployment will mean delaying the realisation of the benefits it can bring, and existing decision-making processes are unlikely to be completely accurate and unbiased. Governments will therefore need to determine the appropriate trade-off between strong controls and experimentation and risk, based on the relative costs and benefits.

Develop open and transparent accountability structures

Establishing trustworthy, fair and accountable processes and structures is likely to help governments realise the potential of AI to transform public services and administration and build public confidence in their ability to do so. If the public does not trust the government to use AI ethically, they will avoid services using AI and oppose their introduction. Therefore, addressing public concerns will be crucial and can be supported by enabling scrutiny, accountability and fair processes.

Transparency and accountability depend on the adoption of legal, ethical and technical frameworks, and systems for monitoring implementation and managing risk, as set out

above. Codes of practice and decisions/rules help establish when the use of AI is permitted in the public sector and what controls and safeguards will need to be put in place. Consistent adoption of these frameworks may help promote procedural fairness, compliance with the law and due process. However, they will only aid accountability if they are communicated to the public in a clear and simple manner. For example, the Government of Canada requires public sector organisations to publish the results of their Algorithmic Impact Assessment as open government data to aid public awareness of decisions that may affect them (see case study in Annex A).

Accountability frameworks are more likely to be effective if governments provide sufficient information on their AI activities to enable facilitate scrutiny by external stakeholders, including experts. For example, the UK Government's Centre for Data Ethics and Innovation (see Box 4.18 later in this chapter) provides oversight of public sector use of AI by conducting governance reviews to identify gaps, risks and opportunities, and recommend improvements.[96] In another example, Etalab, the Prime Minister of France's taskforce for open data and open government, has published a guide for public administrations on how algorithms should be used, with an emphasis on transparency and accountability (see Box 4.10).

[96] www.gov.uk/government/groups/centre-for-data-ethics-and-innovation-cdei.

Box 4.10: Etalab Guidance on Accountability for Public Algorithms

Etalab, the Task Force under the French Prime Minister's Office in charge of open data and open government, has produced a guide for public administrations on the responsible use of algorithms in the public sector. The guide sets out how organisations should report on their use to promote transparency and accountability.

This guidance forms part of a work programme on public algorithms that also includes the production of case studies, the identification of and technical support for AI projects in the public sector, anticipation of the impact of AI on stakeholders and reflection on ethical issues associated with AI use in the public sphere.

The guidance covers three elements:

- **Contextual elements**. These focus on the nature of algorithms, how they can be used in the public sector, and the distinction between automated decisions and cases where algorithms function as decision-supporting tools.
- **Ethics and responsibility of using algorithms to enhance transparency**. This includes public reporting on the use of algorithms, how to ensure fair and unbiased decision-making, and the importance of transparency, explainability and trustworthiness.
- **The legal framework for transparency in algorithms** including the European Union's General Data Protection Regulation (GDPR) and domestic law. This includes a set of rules to be applied to administrative decision-making processes on what specific information must be published about public algorithms.

Etalab also proposes six guiding principles for the accountability of AI in the public sector:

- **Acknowledgment**: agencies are obligated to inform interested parties when an algorithm is used.
- **General explanation**: agencies should provide a clear and understandable explanation of how an algorithm works.
- **Individual explanation**: agencies ought to provide a personalised explanation of a specific result or decision.
- **Justification**: agencies should justify why an algorithm is used and reasons for choosing a particular algorithm.
- **Publication**: agencies should publish the source code and documentation, and inform interested parties whether or not the algorithm was built by a third party.
- **Allow for contestation**: agencies should provide ways of discussing and appealing algorithmic processes.

Source: www.etalab.gouv.fr/datasciences-et-intelligence-artificielle; www.etalab.gouv.fr/how-etalab-is-working-towards-public-sector-algorithms-accountability-a-working-paper-for-rightscon-2019, https://github.com/etalab/algorithmes-publics/blob/master/20190611_WorkingPaper_PSAAccountability_Etalab.pdf; www.europeandataportal.eu/fr/news/enhancing-transparency-through-open-data; www.etalab.gouv.fr/algorithmes-publics-etalab-publie-un-guide-a-lusage-des-administrations.

Consider the explainability of AI systems and automated decision making

In order for accountability to work effectively, governments must be able to explain why an AI system made the decisions it did, particularly if the decision has the potential to impact people's lives. However, the complexity of AI algorithms can make it difficult to provide a clear narrative that explains and justifies a decision. AI seeks to make optimal predictions or inferences based on correlations; it does not depend on an overarching theory or story that explains why those correlations are important causal relationships (Anastasopoulos and Whitford, 2019). In addition, data protection laws may mandate explainability. Under GDPR, organisations are required to explain to citizens how their data are being used by AI to make automated decisions (Raja, 2018).[97]

Regardless of how decisions are made, it is important to know who precisely is empowered to make decisions about how AI is deployed, who is responsible for each decision and to whom they are accountable. Clearly, these accountability processes are crucial where decisions could have a significant impact on citizen's lives. Governance frameworks that give service users voice and oversight will be particularly important (Whittaker, 2018). In this regard, a number of approaches exist that governments could adopt to mitigate explainability issues and enable accountability:

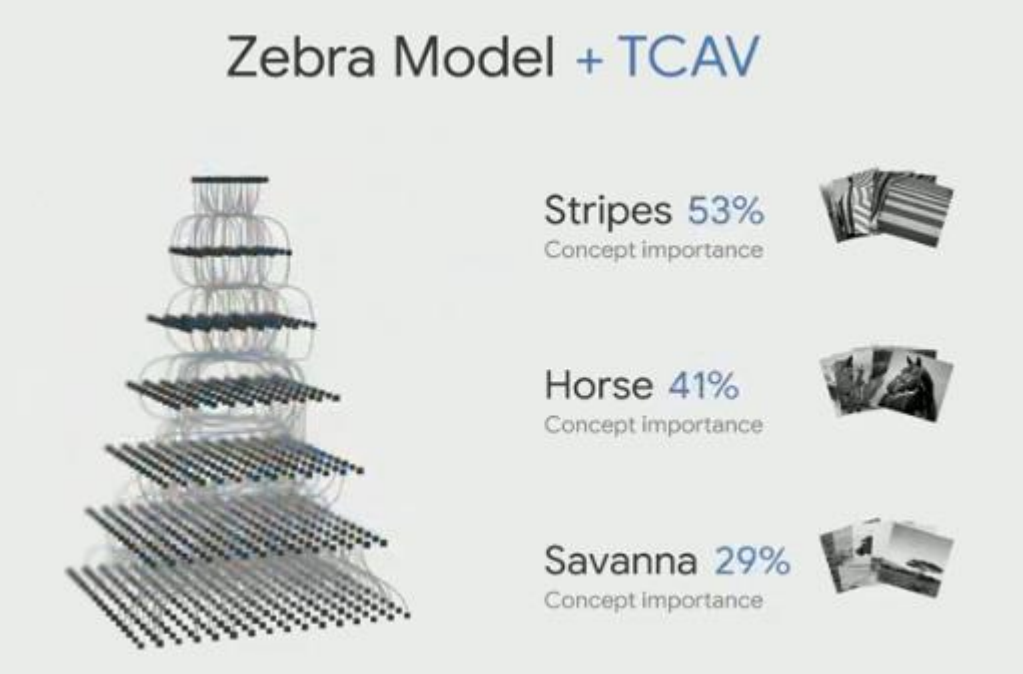
- **Creating explainable AI**. Governments may undertake efforts to create AI that is explainable by design. However, this can result in a trade-off between cost and interpretability.[98] Providing explanations such as those expected by individuals under current law "should often be technically feasible but may sometimes be practically onerous" (Kortz and Doshi-Velez, 2017; see Box 4.11).
- **"Human in the loop" approaches**. Cases where AI supports decision-making by officials may raise fewer explainability issues than full-automated decision-making, but will not fully mitigate the risk and will increase the costs of AI systems, especially when they are deployed at scale (Mateos-Garcia, 2017). If the outputs of an algorithm affect officials' decisions, then the officials and the wider public should be able to understand why the algorithm recommended that decision.

[97] Further information on the links between data protection and AI is given in the section on securing ethical access to, and use of, quality data.

[98] www.pwc.com/us/en/services/consulting/library/artificial-intelligence-predictions/explainable-ai.html.

Box 4.11: Creating explainable neural networks with Testing with Concept Activation Vectors (TCAV)

Neural networks have the potential to make very accurate predictions; however, their complexity makes them difficult to explain. Nonetheless, efforts are underway to explain even the most complex AI. For instance, Google are exploring Testing with Concept Activation Vectors (TCAV) to understand what signals neural networks are using for prediction. This will allow the most salient factors in determining a decision, including sources of bias, to be identified. The example below illustrates the concepts that might be important for an algorithm to identify an image of a zebra:



Source: www.zdnet.com/article/google-says-it-will-address-ai-machine-learning-model-bias-with-technology-called-tcav.

Establish safeguards against bias and unfairness

A key purpose of AI transparency is to mitigate and monitor for bias and distributional fairness in algorithmic decision making. If decisions are made by a "black box" system, it will be harder to monitor whether the outcomes contain bias and may lead to an unfair impact on people's lives. It is therefore necessary to create governance frameworks at the design stage that include a means of monitoring outcomes to identify and mitigate against discrimination on the basis of characteristics, such as ethnicity, gender, income, disability status and age. If there is no means to mitigate an AI's bias, it will be difficult to justify its use in the public sector. Box 4.12 sets out the issues around bias in criminal justice risk assessments in the United States.

Box 4.12: Concerns about algorithmic bias in the US criminal justice system

In some parts of the US criminal justice system, judges use risk assessments that assess the likelihood that a criminal will re-offend to inform decisions about sentences, access to rehabilitative services and to decide whether an accused individual will be held in jail pending trial.

In theory, data-driven decisions should reduce bias in judges' decisions. However, the algorithms estimate recidivism rates based on historical correlations between variables, which do not necessarily represent causal relationships. Therefore, if these correlations are themselves affected by bias, then discrimination will be embedded in the system. For example, if previous judges' decisions have been affected by bias, or there is a correlation between, for example, ethnicity or income and recidivism, then people may receive tougher outcomes because of these characteristics. Therefore, "the algorithm could amplify and perpetuate embedded biases and generate even more bias-tainted data to feed a vicious cycle".

Analysis of the outputs of a risk assessment model used in Broward County, Florida found that when predicted recidivism was compared to actual recidivism rates, black defendants were often predicted to present a higher risk of recidivism than was actually the case, while white defendants were often predicted to present a lower risk. Moreover, as the algorithms are proprietary software, it is not always possible to access the source code to understand how the decisions are made.

A Partnership for AI report identified three sets of issues with the use of these risk assessments:

- **Concerns about the accuracy, bias and validity in the tools themselves**: it should not be assumed that tools are objective and unbiased simply because they are based on data.
- **Issues with the interface between the tools and the humans who interact with them**: tools must be interpretable and explainable so that users can understand how the tools make predictions.
- **Questions of governance, transparency and accountability**: these predictions have a significant impact on citizens' lives, so the people who "specify, mandate and deploy" these tools must be held accountable.

Accordingly, they recommend that either risk assessment tools should not be used or that standards should be put in place to mitigate each of these issues.

Source: www.technologyreview.com/s/612775/algorithms-criminal-justice-ai, www.partnershiponai.org/artificial-intelligence-research-and-ethics-community-calls-for-standards-in-criminal-justice-risk-assessment-tools, www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, https://ainowinstitute.org/AI_Now_2018_Report.pdf.

However, it should not be assumed that AI bias is an inevitable barrier. Improving data inputs, building in adjustments for bias and removing variables that cause bias may make AI applications fairer and more accurate. Creating diverse teams and building in peer review will also mitigate bias (Moneycontrol News, 2019). In many cases, automated decisions may have the potential to be more fair than human decision-making, if they only consider relevant information and do so in a transparent and explicable way.[99]

[99] www.digital.nsw.gov.au/digital-transformation/policy-lab/artificial-intelligence.

In addition to bias, there are also issues of fairness in the distribution of services and social stigma related to the use of AI. "Data scores" that combine data from a variety of sources as a way to categorise citizens, allocate services and predict behaviour have become increasingly common in public services. Such scoring may be used for questionable purposes that can result in further entrenching social inequalities. The example of China's Social Credit scores (Box 4.13) is illustrative of some of the challenges related to data scores.

Box 4.13: China's Social Credit Scores

There are ongoing trials in a number of Chinese cities of a system of social credit that can influence access to services, credit, jobs and travel based on whether the citizen is deemed trustworthy. The system that determines a social credit score is powered by AI, including facial recognition technology linked to CCTV surveillance, data collection from smartphone apps to measure online behaviour, financial assets and government records, such as education, medical and state security assessments.

This gives the authorities the ability to control and shape the behaviour of citizens in what has been called "digital dictatorship". What someone says, purchases and who they associate with can influence their ability to participate in public life. This may have a chilling effect on dissent and scrutiny of the state.

It appears likely that this type of social credit system is technologically feasible in many countries but that does not mean it is desirable or inevitable. Whether such systems emerge, and what controls they are subject to, are political questions. The answers may in part depend on the balance afforded to the importance of a stable and safe society or privacy and individual freedom.

Source: www.abc.net.au/news/2018-09-18/china-social-credit-a-model-citizen-in-a-digital-dictatorship/10200278, https://datajustice.files.wordpress.com/2018/12/data-scores-as-governance-project-report2.pdf, https://time.com/collection/davos-2019/5502592/china-social-credit-score.

However, governments often use similar practices to address pressing social issues. For instance, faced with high rates of refugees seeking better conditions, Switzerland is piloting the use of data-driven refugee profiles, analysed by algorithms, to place refugees in areas where they will have the best chance of achieving positive integration outcomes, including employment. The algorithm is believed to increase employment outcomes by 40-70% on average compared to the status quo (Bansak et al., 2018).

In the United Kingdom, local governments and police forces in some cases have sought to combine a range of datasets, for example, to use AI to predict which children are at risk of abuse or neglect in order to better target services (Dencik et al., 2018) or identify patterns in criminal activity (BBC News, 2019). Well-designed AI services along the lines of Finland's AuroraAI strategy may share information and join up services around the user.[100] Nonetheless, while these applications may not lead to the same concentration of power as the Social Credit example, they still surface a number of issues that public officials should consider:

- A historical correlation between certain characteristics and a negative outcome does not prove the existence of a causal link that will hold over time. Entrenching these relationships in data scores may lead to stereotyping, discrimination and a perception that these correlations cannot be changed by

[100] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/event/semic-webinar-artificial-intelligence-and-public-administrations-09-04-2019-1000-1130-cet.

effective public policy or personal choice. It can therefore lead to stigmatisation of people with certain characteristics.

- AI can be used to identify irregularities and discipline citizens, for instance by identifying benefit fraud, and thereby reduce service costs in the context of tight public finances. There is a risk that the use of algorithms for these purposes will de-personalise public services previously delivered by caseworkers and create a punitive system that could negatively affect the most vulnerable in society. These people may also find it difficult to seek redress if they are incorrectly identified as being in breach of the rules (Shafique, 2018; Whittaker et al., 2018)). Therefore, users, particularly marginalised groups, may experience frustration as a result of "increased administrative burdens" in the form of confusing bureaucracy and complex regulations that create barriers to accessing services (Herd and Moynihan, 2018).
- Related to this, there may be a trade-off between provision of universal public services that are regarded as a citizen's right on the one hand, and tailored services based on characteristics captured in data scores on the other. While the latter may provide better targeted and more appropriate services, such services may lead to increased complexity for users. Furthermore, if services are no longer universal, then that may lead to a reduction in support for services from which not everyone benefits (Shafique, 2018).

Governments will need to consider both bias and fairness when exploring the potential for AI-driven policies and services.

Secure ethical access to, and use of, quality data

As discussed in Chapter 2, data are the foundational building blocks for AI. A clear data strategy that enables governments to access rich, accurate and useful data, maintains privacy, and conforms to societal and ethical norms will be a necessary pre-condition to effectively deploying AI (see Box 4.14 and Annex A for a case study on the US Federal Data Strategy and associated Action Plan). AI is dependent on access to quality data, however obtaining such data is costly and administratively complex. Governments should therefore have clear oversight of their existing assets and a strategic approach to building up quality datasets in areas that are ripe for AI development.

Box 4.14: Existing national government data strategies

A number of countries have established strategies to capitalise on their data assets. These can entail the creation of consistent standards, cross-government data-sharing protocols and the opening up of government data. For example, the Uruguayan Government has developed an interoperability platform to facilitate and promote government digital services and improve integration between public sector organisations. The New Zealand Government also has a set of principles for data management that incorporate open data principles.

The US Federal Data Strategy sets out consistent principles and practices "to deliver a more consistent approach to federal data stewardship, use, and access".

Furthermore, there may be a case for international alignment of digital and data standards. Many forums for sharing best practice in digital government and data already exist, including the Digital 9 Group and the OECD Working Party of Digital Government Officials (E-Leaders).

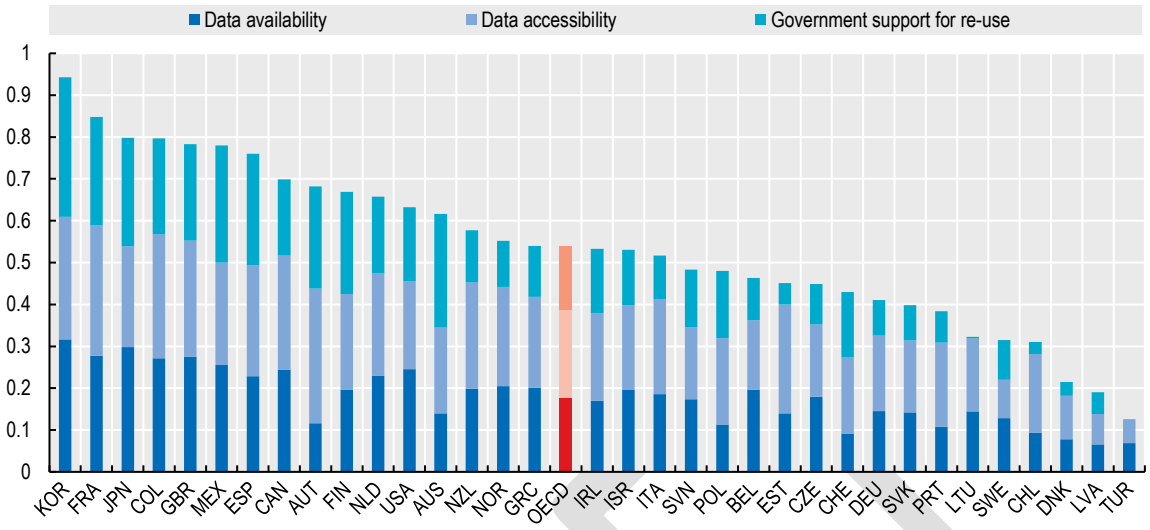
Source: www.oecd-ilibrary.org/governance/a-data-driven-public-sector_09ab162c-en, www.agesic.gub.uy/innovaportal/v/1711/9/agesic/que-es.html?idPadre=3922, www.digital.govt.nz/standards-and-guidance/data-2/data-management, https://strategy.data.gov, www.digital.govt.nz/digital-government/international-partnerships/the-digital-9.

As discussed in Chapter 3, opening up government data is one of the most prevalent themes in national AI strategies. Government offices, private companies and civil society organisations can all benefit from access to public data. They may be able to utilise it to generate AI tools that facilitate innovation, the creation of economic development and public value. AI is increasingly being used to improve service delivery and users' experience, and open data provides information on user behaviour and preferences in order to fuel citizen-driven design.[101]

The latest OECD information on open government data (OGD) indicates that issuing OGD in machine-readable formats is now one of the top priorities in national OGD strategies, and that countries are providing most of their datasets in machine-readable formats. Machine readability is a major factor in data accessibility and overall OGD efforts, as measured by the OECD OURdata Index (see Figure 4.3).

[101] https://trends.oecd-opsi.org.

Figure 4.3: OECD open, useful and reusable data (OURdata) index, 2017



Source: OECD Open Government Data Report: Enhancing Policy Maturity for Sustainable Impact, https://oe.cd/2pq.

AI enables the use of a richer variety of data as inputs to algorithms to inform public policy. Evidence-based policy making has long depended on gathering and analysing information to shape policy development and delivery. However, this information has conventionally taken the form of structured data, such as surveys. AI also enables the incorporation of unstructured data, for example images and open text from social media interactions. It can also harness information generated by digitised service delivery. It therefore creates opportunities for improved problem definition and policy framing, and allows for a quicker, deeper and more precise understanding of citizen preferences and context.

However, these data also create new challenges for policy makers. Inadequate data will lead to AI systems that recommend poor decisions. If data reflect societal inequalities, then applying AI could reinforce them, and may distort policy challenges and preferences (Pencheva, Esteve and Mikhaylov, 2018). If AI has been trained on data from a subset of the population that has different characteristics from the population as a whole, then the algorithm may yield biased or incomplete results. This could lead AI tools to reinforce existing forms of discrimination, such as racism and sexism.[102] Unstructured data may also be more difficult to anonymise and thus undermine privacy standards.

The Data Science Hierarchy of Need, set out in Chapter 2, provides a clear basis for thinking about collection, storage, transformation, analysis and implementation – the steps from taking raw data and using it to generate new insights and information. Each of these stages is likely to entail ethical and legal, as well as technical, issues.

At a minimum, use of AI by governments must conform to national data protection laws. The General Data Protection Regulation (GDPR), brought into force in May 2018, creates consistent data protection rules for all organisations operating in the European Union, in order to give people more control over their personal data and create a "level playing field" for businesses.[103] Among other provisions, it sets rules around building

[102] www.digital.nsw.gov.au/digital-transformation/policy-lab/artificial-intelligence.

[103] https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en.

data protection capabilities in organisations, promotes transparency, and gives citizens a say in how their personal data can be stored and what it can be used for.

For instance, Article 5 of the GDPR describes the concept of "purpose limitation" which restricts the terms under which organisations are able to reuse data they collected or acquired elsewhere. In particular:

> *If your company/organisation has collected the data on the basis of consent or following a legal requirement, no further processing beyond what is covered by the original consent or the provisions of the law is possible. Further processing would require obtaining new consent or a new legal basis.*[104]

Governments will need to consider these laws in the design and development of data strategies and AI initiatives. While this approach may slow the deployment of the AI in the short term, it could create the foundations for more ethical and inclusive AI over the longer term. GDPR has been voluntarily adopted by some states outside the European Union, and may come to form the basis of a global standard for data protection. This conforms with the European Commission's emphasis on developing AI in a way that builds public trust by building in strong standards and protections, as set out above.

Cultural norms will influence popular views on privacy, what data it is ethical to use and what restrictions or permissions should be required. The balance between privacy and using data to improve services and make them more personalised will be difficult to codify and will vary with context. Generating a stable consensus across society on the various trade-offs, for instance between privacy, transparency and service quality (Janssen and van den Hoven, 2015), will be challenging. Where trust in government is low, there is likely to be opposition to gathering large quantities of data and using them in ways that are not clear to the public. For example, facial recognition technology is a method of gathering data that has highlighted concerns about the appropriate balance between more effective services and privacy and potential bias (see Box 4.15).

[104] https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr/purpose-data-processing/can-we-use-data-another-purpose_en.

Box 4.15: Facial recognition technology, privacy and bias concerns

Facial recognition technology can have many transformative applications. For instance, its use to pay for subway rides is being trialled in Futian, China. However, the technology has become a lightning conductor for concerns about privacy. As facial recognition has matured, it has become increasingly capable of identifying faces in a crowd through the use of facial image data. For instance, by matching images from CCTV to police databases, it can provide real-time surveillance and improve safety and security by identifying criminal suspects or missing people, among other applications. In China, facial recognition has been complemented by "gait analysis", which identifies people by the way they walk. However, privacy advocates are concerned that it enables governments to gather a huge amount of information about citizens without their consent, which could be used for a number of purposes.

In addition, facial recognition technology that is trained on datasets which are not sufficiently diverse can reduce the accuracy of identification for some groups, leading to an increased risk of false positives. For example, police forces in the United Kingdom have come under criticism for failing to test the impact of ethnicity on prediction accuracy. An MIT study, for which the results are contested, found that multiple facial recognition tools are less accurate for black people and women, leading to potential bias on the grounds of gender and ethnicity.

There are also cases where inappropriate procedures by police forces led to the use of poor quality input data, substantially weakening the accuracy of facial recognition software. For example, police forces in the United States have sought to match drawings of suspects, poor quality CCTV stills, computer-enhanced images and even a picture of a suspect's celebrity doppelganger to image databases. These examples suggest that clearer rules are required on precisely how the software should be used and to clarify whether a match is sufficient grounds for arrest.

In a context of rapidly changing technology and low levels of trust in government, there are concerns that this technology gives too much power to the public sector. Contentious cases such as these are likely to spark societal debate about whether the use of facial recognition technology is consistent with respect for individual autonomy and, if so, what safeguards need to be put in place to protect liberal values. Citizens are likely to demand a proper consultation on whether the technology is being used in a way that might affect them. In a backlash against its use, San Francisco has become the first city in the United States to ban the municipal use of facial recognition.

Source: https://towardsdatascience.com/how-ethical-is-facial-recognition-technology-8104db2cb81b , www.scmp.com/tech/innovation/article/3001306/you-can-soon-pay-your-subway-ride-scanning-your-face-china, www.bbc.co.uk/news/technology-47117299, https://medium.com/@AINowInstitute/after-a-year-of-tech-scandals-our-10-recommendations-for-ai-95b3b2c5e5, www.flawedfacedata.com, www.americaunderwatch.com, www.bbc.co.uk/news/technology-48222017, www.vox.com/future-perfect/2019/5/16/18625137/ai-facial-recognition-ban-san-francisco-surveillance, www.sfchronicle.com/politics/article/SF-could-ban-facial-recognition-software-13842657.php.

Ensure government has access to internal and external capability and capacity

Differing initial levels of institutional maturity and capacity in government, academia, civil society and the private sector will lead to constraints on government access to talent and necessitate different strategic approaches to realising the benefits of AI. For instance, effective deployment of AI is likely to depend on technical and service transformation skills and capabilities that are unlikely to exist at present in the public sector. Such skills may be difficult to build internally, but may also be challenging to

obtain externally due to cumbersome procurement processes with the private sector or inadequate mechanisms to collaborate with academia and civil society.

The approach taken by governments to developing AI talent should focus not just on technical skills but also on multi-disciplinary capacity building of the social, ethical and legal implications of AI, and the shift in mindset and ways of working needed to collaborate with mixed teams and AI (AI Now, 2018).[105]

Governments may seek to address these technical and non-technical challenges through innovative approaches to training, recruitment, procurement and partnership. To realise the potential of AI for the public sector, governments will need to pursue a mix of approaches. However, regardless of this mix, governments will always need to maintain a unique and integral role in direction setting, standard setting and monitoring compliance with policies and laws, as it will always be the responsibility of government to ensure the appropriate design and use of AI within the public sector.

Build internal capacity

Widespread AI transformation is likely to have substantial implications for the skills required to effectively deliver public services. These changes will include the following:

- Senior leaders will need to understand how to maximise the value of AI in public services.
- Service managers will oversee delivery through effective commissioning.
- Internal technical expertise may be needed to enable government to be a thought leader and to negotiate effectively with contractors.
- Front-line staff will need the skills and capabilities to work alongside, interpret and complement AI.

Across all these levels, the development of a diverse AI workforce that reflects the make-up of the population, through recruitment practices and building an inclusive culture, will be a crucial safeguard against unethical practices, bias and group-think (du Preez, 2018). Investing in training to build levels of AI literacy across the organisation may reduce workplace anxiety about the implications for staff. The Digital Academy at Canada's School of Public Service (see Box 4.16) offers an example of different levels of training in AI and digital technologies to fit the needs of different groups of officials.

[105] www.pwc.com/us/en/services/consulting/library/artificial-intelligence-predictions/employer-impact.html.

Box 4.16: Canada School of Public Service's Digital Academy

The Digital Academy is a teaching organisation hosted at the Canada School of Public Service. It provides support to public servants to improve their operations by delivering digital services. The Academy's programme is part of a wider public sector reform agenda to create an agile, inclusive and equipped public service.

The Digital Academy offers training for officials at all levels of seniority and with differing levels of existing specialist expertise. It uses real-life challenges and problems and a mix of events, online learning and podcasts (called busrides.ca and designed to give quick introductions to topics related to government digital services). Learning opportunities follow three tiers:

1. **The Digital Foundations** tier is intended for all public servants and all levels of expertise. It aims to provide timely information on the digital world that will affect how public servants do their jobs and even live their lives.
2. The **Digital Premium** tier, or the specialised streams for practitioners, focuses on data, design, development, AI and Machine Learning, DevOps and disruptive technology.
3. The **Digital Leadership** tier aims to develop digital skills and mindsets for those who lead service design and delivery – not to mention the culture change required to "do digital" successfully in the Federal public service.

Source: www.csps-efpc.gc.ca/About_us/Business_lines/digitalacademy-eng.aspx.

Governments should explore ways to build up the expertise of a technologically literate senior leadership cadre that can champion the deployment of AI in government. Senior leaders, including at the political level, will need to possess a strategic understanding of what AI can do, and know how to identify the kinds of problem that AI can address and the key questions to ask to ensure effective oversight of delivery (Agrawal, Gans and Goldfarb, 2018). This may not require in-depth technical knowledge but will require the ability to act as an interlocutor or translator able to understand the technical, ethical and legal aspects of delivering feasible AI, and combine them with an understanding of how public services and administrations function (du Preez, 2018).[106]

Managers of AI-enabled services will require deeper technical expertise, even if the services are delivered by external contractors. Knowledge of AI, effective negotiation skills and sector expertise will help service managers design adequate contracts that ensure effective oversight and hold external contractors to account. These aptitudes can also help proposed AI solutions to be properly assessed to verify fitness for purpose and accurate pricing. Service managers will need to work closely through networks of private sector, civil society and academic actors, drawing on their knowledge and collaborating effectively. However, they will also need to avoid being unduly influenced by external stakeholders and maintain a focus on their organisation's objectives and maximising public value. Therefore, building up internal expertise and negotiation skills will help governments to commission services and collaborate effectively with external stakeholders.[107]

[106] www.pwc.com/us/en/services/consulting/library/artificial-intelligence-predictions/functional-specialists.html.

[107] https://joinup.ec.europa.eu/sites/default/files/event/attachment/2019-04/SEMIC_Webinar%20on%20AI%20and%20PA_Presentation_AI%20projects%20in%20PA%20in%20Japan_Kenji%20Hiramoto%20%28JP%29_09-04-2019.pdf.

There may be cases where governments wish to develop in-house technical expertise to help them take on a leadership role in the AI space. There is a strong case for developing AI talent within government, especially in areas that have sensitive security implications or where there are particular opportunities for sharing learning across the organisation. Data protection rules mean that it is often easier to move people than it is to move data. In these cases, approaches may include embedding external contractors in the public sector, exploring inward secondments or building capability internally (Mikhaylov, Esteve and Campion, 2018).

One method of strengthening internal capacity is to build up internal AI talent by upskilling existing staff, such as statisticians and data scientists, with relevant skills and aptitude, and exploring outward secondments to innovative organisations to build organisational knowledge. Box 2.6 in Chapter 2 on Robotic Process Automation in the United States offers an example of reinvesting the cost-savings from automation into upskilling existing staff, so they can operate in more strategic roles.

Another method is to recruit expertise into government. This can often be challenging, as governments often have strict rules on public sector recruitment. AI skills are in high demand and the public sector may struggle to recruit and retain staff if they are unable to be flexible and compete with private sector salaries. Governments may have experience (or even special hiring authorities) to draw on from their previous initiatives to source specialist skills, such as scientists, experienced project managers and economists. Options include seeking to attract staff by offering flexible working conditions, unique development opportunities and experiences that will aid their long-term career development. Some governments have also leveraged the significant civic impact of government work as a recruiting tool. The US Digital Service,[108] for example, recruits tech specialists into government to further social missions through term-limited "tours of civic service".

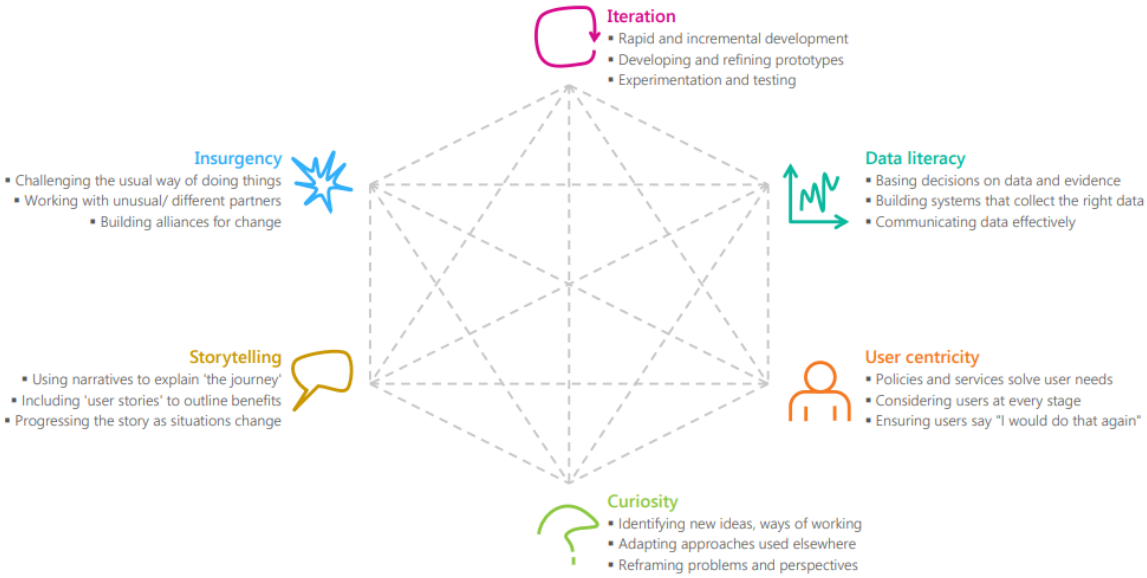
Besides technical skills (e.g. data science, coding, etc.) and the ability to interpret the outputs of algorithms, the emergence of AI increases the value of complementary skills.[109] These are the skills needed to deliver tasks that AI cannot do, such as the ability to make judgements that balance many objectives based on scarce data, creativity and emotional intelligence. OPSI has developed the *Core Skills for Public Sector Innovation* to help guide governments in building 21st century skills. Not all public servants will need to make use of or apply these skills in their day-to-day job. However, for a modern public service, all officials should have at least some level of awareness of these six areas in order to support increased levels of innovation. These skills are:

- **Iteration** – incrementally and experimentally developing policies, products and services
- **Data literacy** – ensuring decisions are data-driven and that data are not an afterthought
- **User centricity** – public services should be focused on solving and servicing user needs
- **Curiosity** – seeking out and trying new ideas or ways of working
- **Storytelling** – explaining change in a way that builds support
- **Insurgency** – challenging the status quo and working with unusual partners.

[108] https://usds.gov.

[109] www.pwc.com/us/en/services/consulting/library/artificial-intelligence-predictions/employer-impact.html.

Figure 4.4: Six core skills for public sector innovation



Source: https://oe.cd/innovationskills.

As AI becomes cheaper and more prevalent, these skills may become more important in government. For instance, staff in operational roles may see their jobs evolve as routine tasks are handled by AI, freeing up time to enable them to focus on the most complex cases and relationships with citizens. Depending on the decisions made by senior leaders, this driver could lead to more personalised, high-quality services or a reduction in the numbers of junior administrative staff whose tasks are more substitutable for AI. In addition, AI could be used to monitor staff productivity, and inform recruitment and performance management decisions (see Box 4.17).

Box 4.17: AI and human resource management

AI has the potential to significantly change how organisations manage their human resources by making them more data-led and responsive. For example, AI could influence recruitment decisions by making predictions about the best candidates based on the characteristics of successful previous candidates. However, the experience of Amazon, where a recruitment algorithm was biased against female applicants, emphasises the importance of effective scrutiny, testing and governance to identify unintended outcomes.

It could also enable faster feedback based on a broader range of data than conventional performance management systems, and counteract managers' conscious and unconscious biases. Again, though, there are risks to be mitigated. The example of an AI system used in Houston, Texas to make recommendations on which teachers to promote or fire based on student test results highlights the importance of systems being well-understood by organisations and decisions being explainable. Unless these issues are addressed, organisations may see a decline in employee morale and may be vulnerable to legal challenge.

Workplace anxiety about the impact of AI may be addressed by developing a narrative about the implications of AI for staff that clarifies how organisations will move people around in response to AI-enable service transformation.

Source: www.forbes.com/sites/insights-intelai/2018/11/29/how-ai-can-help-redesign-the-employee-experience/#19f64c044b34; www.forbes.com/sites/bernardmarr/2017/01/17/the-future-of-performance-management-how-ai-and-big-data-combat-workplace-bias/#1517089a4a0d; https://ainowinstitute.org/AI_Now_2018_Report.pdf, www.bbc.co.uk/news/technology-45809919.

Harness external expertise through partnerships and collaboration

In addition to building internal capacity, governments may draw upon a network of private sector, academic and civil society actors in order to leverage their expertise and resources, and promote knowledge sharing, to improve decision making. Indeed, there are numerous examples of existing cross-sectoral initiatives working to combine capabilities to deliver AI solutions, such as Offices of Data Analytics. These often provide an institutional focal point for collaboration between local and national government, universities, tech firms and non-profits to combine data and address social problems. For example, the New York Mayor's Office for Data Analytics (MODA)[110] actively draws on expertise from Columbia University and NYU to develop data standards and protocols (Mikhaylov, Campion and Esteve, 2018).

In the United Kingdom, the Alan Turing Institute was set up in 2015 by a research council and a group of leading universities as the national institute for data science and artificial intelligence. Its Public Policy Programme enables government agencies to draw on a wealth of external expertise to inform public services and administration (see case study in Annex A). The Institute leverages its reputation to attract academic fellows, appeals to the desire to contribute to the public good and also offers flexible ways to contribute that can be made around other professional commitments. The Public Policy Programme, in particular, has been referenced as a highly successful model in OPSI's interviews with AI stakeholders in a number of countries.[111]

[110] www.nyc.gov/analytics.

[111] www.turing.ac.uk/research/research-programmes/public-policy.

At a strategic level, cross-sectoral collaboration can help government understand existing capabilities and industry priorities, and design better policy. Establishing institutions that facilitate dialogue may develop mutual trust. For example, Canada and the United Kingdom have developed AI advisory committees to enable close working between government, the private sector and academia (Government of Canada, 2019b; Gov.UK, 2019a; see Box 4.18).

Box 4.18: The UK Government's AI Council and Centre for Data Ethics and Innovation

The UK Government has created a senior AI Council as an independent expert committee to advise on how to stimulate the adoption of AI, promote its ethical use and maximise its contribution to economic growth.

The Council consists of leaders from business, academia and civil society. It is envisaged that it will provide a focal point for cross-sectoral collaboration within the AI community to provide solutions to shared priorities, such as data and ethics, adoption, skills and diversity.

The AI Council will sit alongside the Centre for Data Ethics and Innovation, an independent advisory body that will analyse and anticipate the opportunities and risks posed by data-driven technology, and will put forward practical and evidence-based advice to address them. This will include reviews to identify and articulate best practice for the responsible use of data-driven technology within specific sectors or for specific applications of technology.

Source: www.gov.uk/government/news/leading-experts-appointed-to-ai-council-to-supercharge-the-uks-artificial-intelligence-sector, www.gov.uk/government/groups/centre-for-data-ethics-and-innovation-cdei.

At an operational level, delivering AI-based services through a network of academic, private sector, civil society and public sector organisations can help government to leverage external expertise to improve the effectiveness and efficiency of public services. For example, in the United Kingdom, Essex County Council and the University of Essex have partnered to improve public services (see Box 4.19).

Box 4.19: Essex County Council and the University of Essex's Institute for Analytics and Data Science (IADS)

IADS offers an example of the creation of an institutional vehicle for cross-sectoral collaboration at the local government level. This partnership is facilitated by a joint appointment between the two organisations of a Chief Scientific Adviser for the Council who is also Professor of Public Policy and Data Science at the University.

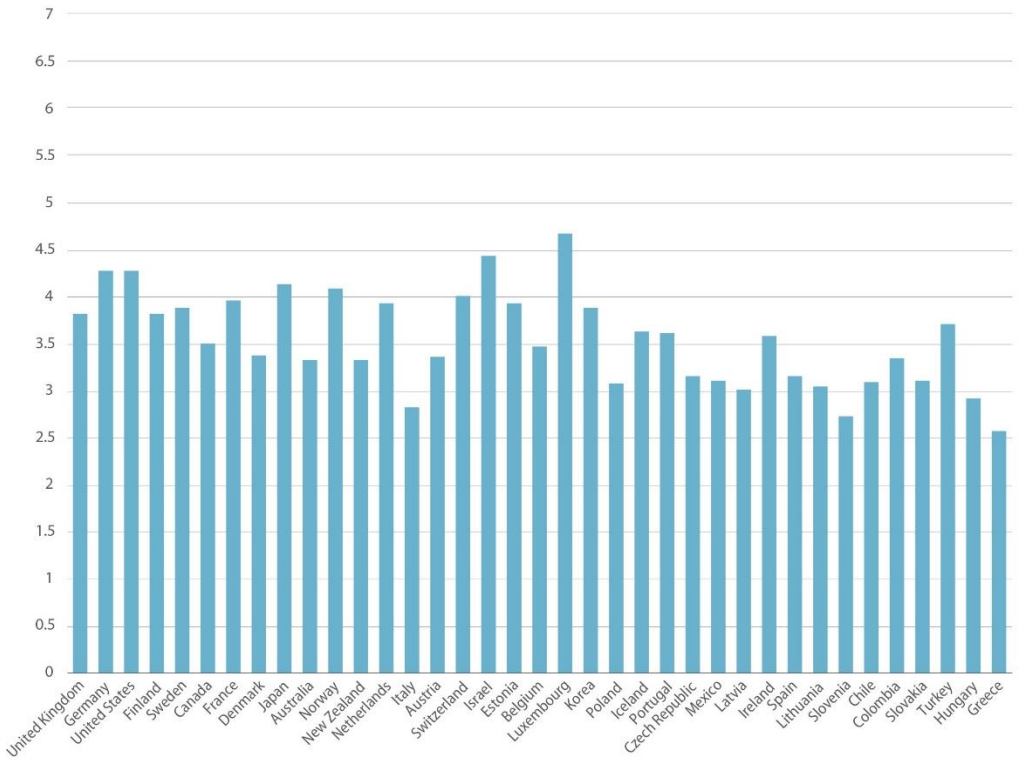
The collaboration enables the combination of public sector data and university and business AI expertise to the benefit of the community in Essex. For instance, operational service improvements include a tool to predict the risk of 14 year olds becoming young people not in employment, education or training by the age of 18. The tool enables targeted early intervention in schools to reduce the risk of this outcome.

Source: Mikhaylov SJ, Esteve M, Campion A. 2018 Artificial intelligence for the public sector: opportunities and challenges of cross-sector collaboration. Phil. Trans. R. Soc. A376: 20170357. http://dx.doi.org/10.1098/rsta.2017.0357.

Design effective public sector AI procurement processes

In many instances, internal talent and cross-sector collaboration will not be enough. Governments will need to purchase skills and capabilities from the private sector through public procurement processes. Given the uncertainty of the technology and the lack of existing mature markets and standards, it can be difficult to draft detailed contracts that balance obtaining services and mitigating risk. "Arms-length" procurement from the market, where firms deliver services for government in accordance with detailed legal contracts and technical requirements, is unlikely to work and governments may need to develop longer-term, collaborative relationships with delivery partners. They may wish to adopt innovative procurement approaches to foster innovation and the creation of deep and competitive markets for AI goods and services. Figure 4.5 shows the extent to which government procurement of advanced technology products in OECD countries takes into account innovation as well as price.

Figure 4.5: Government procurement of advanced technology products in OECD countries



Note: The figure is based responses to the question "In your country, to what extent do government purchasing decisions foster innovation? [1 = not at all; 7 = to a great extent].
Source: WEF Executive Opinion Survey 2015 (survey of over 14 000 business executives in more than 140 countries) http://reports.weforum.org/global-information-technology-report-2016/networked-readiness-index; Oxford Insights Government AI Readiness Index, 2019, www.oxfordinsights.com/ai-readiness2019.

Failure to promote diversity, openness and ethically and technically robust standards in AI procurement may lead to sub-optimal technological trajectories for AI that entrench the market power of large firms, limit accountability and undermine social values (Mateos-Garcia, 2018). Intellectual property laws and other rules that protect proprietary software can render "systems opaque and unaccountable, making it hard to assess bias, contest decisions, or remedy errors". In order to maintain public trust and address information asymmetries, firms delivering AI services and goods should be subject to high standards of accountability, transparency, fairness and privacy. The Canadian Government has developed a "source list" (Box 4.20) to help government offices streamline procurement and select vendors with expertise in AI ethics[112] (see case study in Annex A). While external suppliers may not always be required to make their proprietary software public, governments should build in requirements enabling them to access the source code for audit purposes to understand why important decisions were made.

[112] www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

Box 4.20: The Government of Canada's AI Source list for the promotion of innovate procurement

The Government of Canada has created an AI Source List with 73 pre-approved suppliers "to provide Canada with responsible and effective AI services, solutions and products". The framework allows government agencies to expedite procurement from firms that have demonstrated that they are capable of providing quality AI goods and services.

The framework requires suppliers to demonstrate competence in AI ethics, as well as implementation and access to talent. Firms that responded to the "Invitation to Qualify" had to prove to an inter-disciplinary panel that they satisfied these requirements. The framework has three bands with escalating requirements. The lowest band has less stringent requirements, making it easier for small start-ups to qualify, thereby driving innovation and creating a deeper market.

The framework supports mission-driven and iterative innovation by allowing multiple firms to be commissioned to develop early-stage services to address a problem. This enables effective information sharing and an agile approach to mitigate the uncertainty of a disruptive technology.

The process of establishing and maintaining this list of AI service providers is also an important way for the Government of Canada to engage with private companies in longer-term relationships. This dialogue facilitates the development of shared expectations and mutual understanding of the challenges they may be facing that are relevant to public sector organisations.

Source: https://buyandsell.gc.ca/procurement-data/tender-notice/PW-EE-017-34526, https://buyandsell.gc.ca/cds/public/2018/09/21/5e886991ecc74498b76e3c59a6777cb6/ABES.PROD.PW__EE.B017.E33817.EBSU001.PDF.

Bringing it all together: A framework for governments to develop their AI strategy

This section brings together the implications and considerations set out above and places them in a high-level framework to help governments to think about their AI strategy to transform public services and administration.[113]

Governments will want to adopt different strategies based on their strategic context, priorities and baseline capabilities. An AI strategy should include the following:

- **Baselines**: an assessment of the organisation's current strategic situation and challenges that AI might help address.
- **Objectives**: what the organisation wants to achieve using AI and the principles that will underpin the actions it takes to achieve them.
- **Approaches**: the concrete actions that will be undertaken to achieve these objectives.

[113] This chapter draws on a range of sources, notably: www.nesta.org.uk/blog/10-questions-ai-public-sector-algorithmic-decision-making/ https://hbr.org/2018/04/a-simple-tool-to-start-making-decisions-with-the-help-of-ai https://docs.google.com/presentation/d/1TAJ2A4NvMLFi7b0mTvNyL1pMVRy84UhzhgcsXknhR2g/edit#slide=id.p1) Agarwal, A., Gans, J. and Goldfarb, A. (2018). Prediction machines: the simple economics of artificial intelligence. Harvard Business Press, https://faculty.ai/products-services/ai-strategy.

The framework below sets out the elements that governments should consider including in their AI strategy. Effective strategy will require monitoring to provide a clear overview of the current situation. The elements of AI strategy should not be regarded as sequential and linear. However, they will need to be developed concurrently and the strategy will need to be a live document that can be iterated to ensure consistency and adapt as the context evolves.

A framework for an AI strategy

Baseline	**Objectives**	**Approaches**
Determine current strengths and weaknesses by mapping: Internal AI capabilities Government data assets Existing government AI and data science projects. *Assess the strategic context*: Public and workforce attitudes to government and AI, including trust Current legislative framework Existing government and international commitments and institutions Academic and private sector expertise that might be drawn upon. *Identify specific operational problems that AI has the potential to help solve*: Adopt a multi-disciplinary approach to decide whether AI is the best solution to a policy problem. Create mechanisms to match resources to priority problems. Define the specific decision AI will make or support. Consider who will be impacted by this decision and associated risks if it fails. Explore how the service will need to be redesigned to leverage the impact of AI.	*Decide what goals the AI should help government achieve*: Articulate how AI will generate public value and specify missions to which AI can be part of the solution. Engage stakeholders in goal definition. Leave space for experimentation and learning. *Define and communicate to stakeholders the principles that will shape how AI is used in government*: Fairness and unbiasedness Transparency and accountability Privacy and individual autonomy.	*Ensure government access to AI capability and capacity*: Construct talent pipelines, and develop recruitment and retention plans for internal technical expertise. Harness external expertise through partnerships and collaboration. Design effective public sector AI procurement processes. Build a cadre of service managers and senior leaders who understand the legal, ethical, technical and managerial issues around AI. *Secure ethical access to, and use of, quality data*: Determine what data are needed to address the problems. Decide how to obtain input data of sufficient quality and that are sufficiently representative of the target population to make accurate predictions with minimal bias. Develop a data strategy that complies with data protection law and best practice and is consistent with principles to which there is a commitment. *Put in place legal, ethical and technical frameworks to operationalise the principles*: Monitor compliance with principles during implementation to track progress, and identify and respond to emerging issues. Put in place safeguards against bias and unfairness. Clarify the appropriate role for humans in the decision-making process. Develop open and transparent accountability structures.

Annex A: Case Studies

As can be seen throughout this guide, governments are taking an increasingly active role when it comes to designing and implementing AI projects, as well as putting in place the enabling conditions and guidance needed to ensure the projects are executed in an efficient, effective and ethical way. This section presents a number of case studies that illustrate the approaches governments are using to achieve this, in order to bring about innovative new policies and services. They includes cases on specific AI projects, as well as broader methods and frameworks for considering the application of AI. This collection of cases is not exhaustive but helps to create a body of knowledge about different practices around the world, the context in which they emerged, the technologies used and, when possible, the lessons learned from these experiences.

Using AI to crowdsource public decision-making in Belgium

Issue

Governments are increasingly working to develop citizen-driven policies and services. By definition, this requires extensive engagement with citizens and residents in order to understand their perspectives, opinions and needs. Digital participation platforms are important tools for achieving this and improving government responsiveness. However, analysing the high volumes of citizen input collected on these platforms is extremely time-consuming and daunting for government officials, and hinders them from uncovering valuable inputs. Setting up a digital participation platform, therefore, is not enough: the process of data analysis has to be more accessible to enable civil servants to tap into collective intelligence and make better-informed decisions.

Response

Belgium's CitizenLab[114] is a civil society organisation that aims to empower civil servants and provide them with machine-learning augmented processes that will help them analyse citizen input, make better decisions and collaborate more efficiently internally.[115]

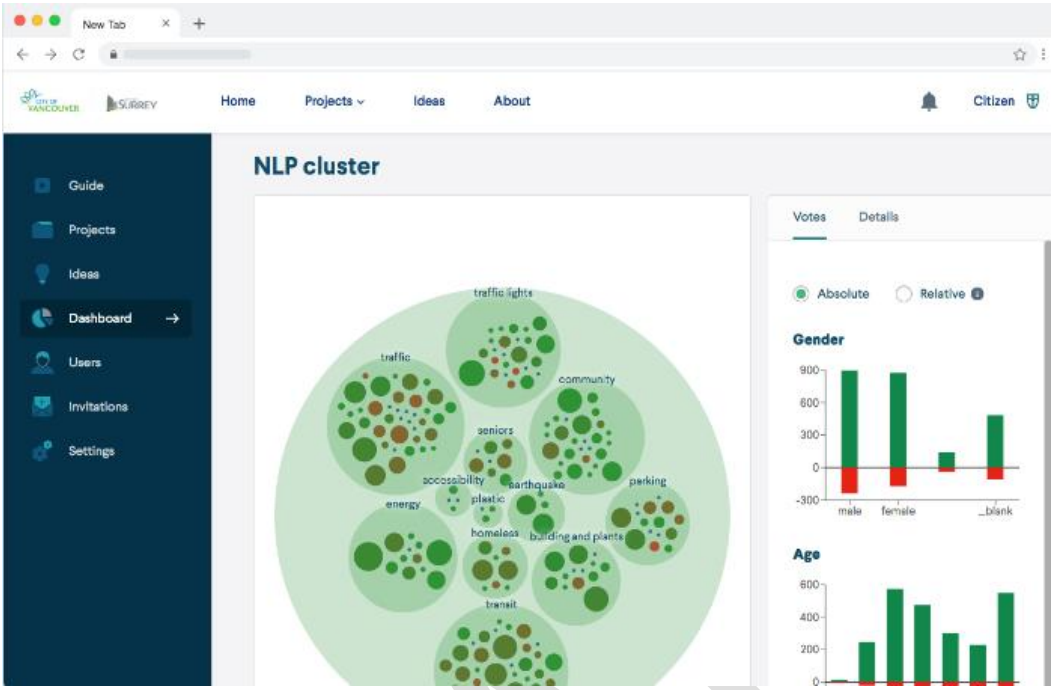
Pursuant to its mission, CitizenLab has developed a public participation platform that uses machine-learning algorithms to help civil servants easily process thousands of citizen contributions and use these insights efficiently in decision-making. The dashboards on the platform can classify ideas, highlight emerging topics, summarise trends, and cluster similar contributions by theme, demographic trait or location.

CitizenLab's platform uses Natural Language Processing (NLP) and Machine Learning techniques to automatically classify and analyse thousands of contributions collected on citizen participation platforms. The algorithms identify the main topics and group similar ideas together into clusters, which can then be broken down by demographic trait or geographic location.

[114] www.citizenlab.co.

[115] See https://oecd-opsi.org/innovations/unlocking-the-potential-of-crowdsourcing-for-public-decision-making-with-artificial-intelligence for additional details on this case.

Figure A.1: Cluster of community interest



Source: https://oecd-opsi.org/innovations/unlocking-the-potential-of-crowdsourcing-for-public-decision-making-with-artificial-intelligence.

Civil servants who manage these citizen participation platforms are able to access this information at a glance through intelligent, real-time dashboards. The "Topic modelling" feature allows them to easily identify citizen's priorities and to make decisions accordingly.

The platform allows civil servants to break down results by demographic groups and location, which gives them a better overview of variation in priorities. For instance, a certain neighbourhood may prioritise better roads, while its neighbour is petitioning for additional traffic stops.

In one relevant example from early 2019, growing numbers of Belgian youth were protesting inaction against climate change, a movement that evolved into Youth for Climate Belgium. In response, CitizenLab set up a participation platform on the topic entitled Youth4Climate, and invited users to submit ideas on tackling climate change.[116] Over three months, users submitted 1 700 ideas, 2 600 comments and 32 000 votes on initiatives they wanted to support. The AI system analysed these items and surfaced and clustered the most important and supported priorities. The CitizenLab is using the AI-driven findings to develop a report for elected officials with 16 policy recommendations.

Through continual iteration of the platform, CitizenLab is working to ensure that governments are making optimal use of their automated dashboards. In addition, the organisation is exploring ways that this technology could be applied to larger-scale conversations on social media, public forums or other places for online debate.

[116] Details on the process are available at www.citizenlab.co/blog/civic-engagement/youth-for-climate-case-study.

Results and impact

Governments using this platform have experienced positive results. The city of Kortrijk, for instance, uses the intelligent dashboards to easily process contributions by the 1 300 users of their platform. They have clustered ideas from conversations into main topics and shared the results of the analysis with citizens. The outcome is a real dialogue rather than a top-down initiative. In another instance, the city of Temse consulted its citizens on the topic of mobility and located the crowdsourced ideas on a map of the city. This helped the administration identify the areas affected by key issues and make decisions about where to allocate funds.

By automating the time-consuming task of data analysis, the platform frees up time for administrations to meaningfully engage with citizens. It also provides governments with a better understanding of citizens' needs and priorities, which in turn leads to better-informed decisions. Governments using the platform have reported such results.

From the perspective of citizens, this open and transparent process helps to foster trust and increase support for policy decisions. It has also had a positive impact on the willingness of citizens to participate.

Challenges and lessons learned

CitizenLab has faced two main challenges: classification algorithms and human adoption.

The platform uses a classification algorithm that clusters, categorises and summarises input from citizens. It needs to be easily scalable, but must also adapt to different administrations' workflows since the taxonomies used might vary by country or even by region. The classification algorithms also need to support multiple languages on the same platform and make semantic links between languages, which adds an extra layer of technical complexity. When working on the Youth4Climate platform in Brussels, CitizenLab had to analyse thousands of contributions in French, Dutch and English. They found that the best result was obtained by automatically translating comments into a single language, and then working from there.

On the human side, CitizenLab needs to ensure that the technology responds to real user needs in order to maximise adoption by governments. The team has learned that the product should not be promoted without first guiding the users through its benefits. They have also learned that human-machine interaction is crucial. The user needs to learn to interpret and "trust" the output generated by the machine and understand the role this output play can in daily workflow. Governments need to consider these things before deploying such a solution.

The CitizenLab team also noted several conditions for success. The first is promoting adoption of the platform. This involves ensuring that civil servants understand its benefits and feel that they can rely on and trust the results. Explaining the methodology and integrating the public engagement process with existing workflows helps in this regard. It is also important to specify an identified need, as time and resources are scarce in administrations, and civil servants will only invest in a tool if it has proven value.

Secondly, the team found that the quality of inputs (i.e. citizen feedback) is critical to successfully understanding citizens' perspectives and needs. To this end, civil servants need to provide guidance to citizens to ensure they submit useful contributions. The team also conducted user testing regularly and refined the approach based on feedback. At a more strategic level, the CitizenLab team found that highlighting the importance of citizen participation at the most senior levels of government encourages government offices and civil servants to seek out public feedback. This, in turn, promoted continuous improvement of the platform and its results.

Finland's National AI Strategy

Issue

Despite being a small country with a population of 5.5 million, Finland has declared its intention of becoming a world leader in the application of AI. The country is well positioned to achieve this goal due to a number of factors. Its citizens are highly educated and tech savvy, the economy is already technology intensive, the government has amassed high-quality data, and after years of reform its public sector is highly digitised and embraces experimentation and innovation. In addition, research from consulting firm McKinsey indicates that if Finland accelerates development in AI and automation, it can expect a GDP increase of 3% per year and net employment gains of 5% (McKinsey & Company, 2017). The right mix of enablers and incentives is in place. The key question is: What exactly does Finland need to do to meet its potential?

Response

Finland's Age of AI

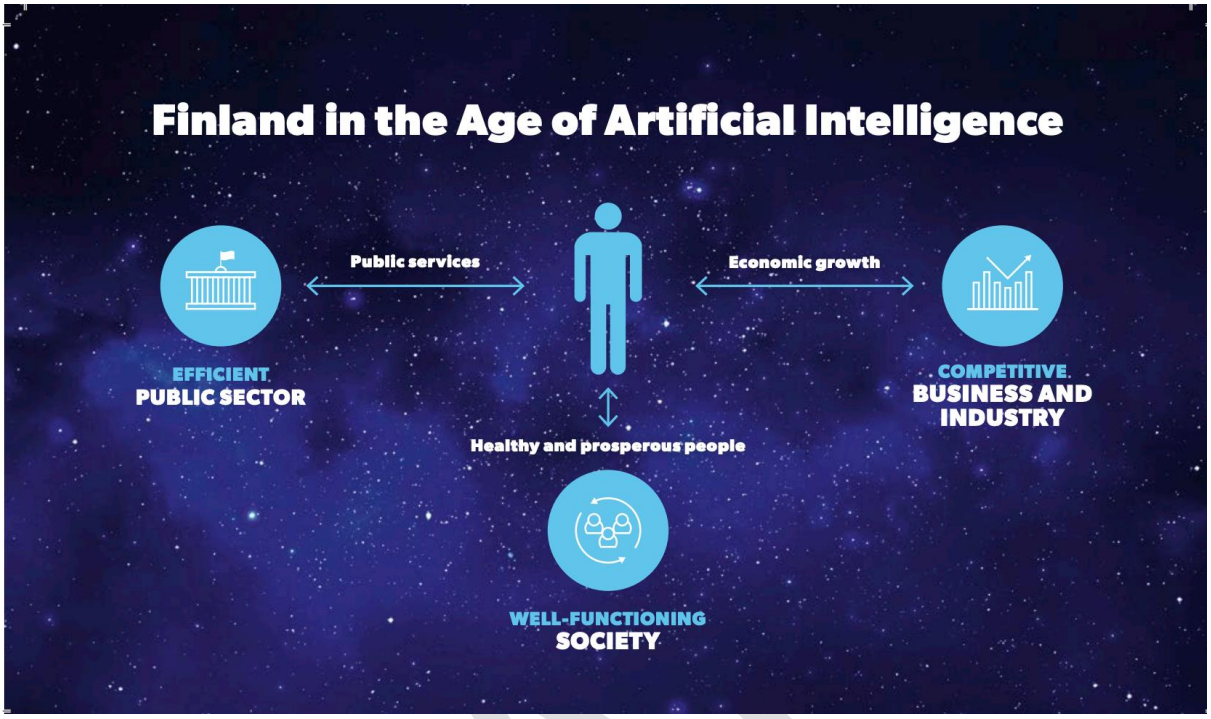
In May 2017, Finland's Ministry of Economic Affairs and Employment created an Artificial Intelligence Programme and a Steering Group to ensure its guidance. The group leveraged a broad network of experts to explore key questions about how best to support the public and private sectors in producing AI-based innovation, how to position government data as resources for economic development, how AI will affect society and what the public sector should do to move Finland towards an AI-driven future. As a consequence of this work, the Steering Group issued two key reports that set forth Finland's approach to AI. *Finland's Age of Artificial Intelligence*[117] (December 2017) and *Leading the Way into the Age of Artificial Intelligence* (June 2019) collectively lay out 11 key actions covering all sectors to help Finland achieve its ambitious goal:

4. Enhance business competitiveness through the use of AI.
1. Effectively utilise data in all sectors
2. Ensure AI can be adopted more quickly and easily.
3. Ensure top-level expertise and attract top experts.
4. Make bold decisions and investments.
5. Build the world's best public services.
6. Establish new models for collaboration.
7. Make Finland a front runner in the age of AI.
8. Prepare for Artificial Intelligence to change the nature of work.
9. Steer AI development in a trust-based, human-centred direction.
10. Prepare for security challenges.

While number six is the action with the clearest implications for the public sector, there is a strong sub-focus on the public sector throughout the document, which envisions a government that provides anticipatory and personalised services to all citizens at all stages of their life in order to support a well-functioning society. Uniquely, when compared to other national strategies, Finland's approach places efficiency of the public sector and the effectiveness of its services on a par with economic growth (Figure A.2).

[117] http://julkaisut.valtioneuvosto.fi/handle/10024/160391.

Figure A.2: AI for achieving a well-functioning society



Source: www.tekoalyaika.fi/en/reports/finland-leading-the-way-into-the-age-of-artificial-intelligence.

Spread across the key action areas, the objectives directly relevant to innovation and transformation of the public sector include the following:

- Develop new operating models to shift from organisation-based activities to systems-wide approaches.
- Adapt the role of government to ensure that citizens have the right to independently determine how their data are used, while protecting the privacy of the citizens.
- Improve the interoperability of government data, and open up this data to fuel innovation in all sectors; encourage companies to share data as well.
- Create a Centre of Excellence for AI, a virtual AI university and a Masters programme in AI to strengthen the talent pool for both the private and public sectors.
- Pursue and build a network for public-private partnerships to allow for collaborative initiatives, knowledge exchange and better adoption of multidimensional thinking.
- Hold a public discussion on AI ethics at in-person events and online.
- Break down silos within and between businesses and public services.
- Revise procurement law to enable effective public-private co-development.

In addition, while some national strategies focus generally on strategies and goals, the Finland approach also identifies specific projects to be adopted by the government to

facilitate the AI transformation, as well as the government components responsible for their implementation. Critically for the public sector, the *Age of AI* report calls for the government to establish Aurora, a network of different smart services and applications to "allow [the] public administration to better anticipate and provide resources for future service needs" and to allow citizens to access high-quality 24/7 digital services.

Expansion into the AuroraAI National AI Programme

Since the initial concept for Aurora was released, it has been expanded significantly into the AuroraAI National AI Programme. AuroraAI seeks to provide a holistic set of personalised AI-driven services for citizens and businesses in a way that is human-centric and works towards their wellbeing as its ultimate goal. AuroraAI, as a broader concept, is intended to enable citizens to access the wide range of services available from various government and cross-sector service providers in a seamless way. The AuroraAI programme is guided by nine principles of digitisation (Figure A.3).

Figure A.3: Nine principles of digitalisation

Source:
https://vm.fi/documents/10623/1464506/AuroraAI+development+and+implementation+plan+2019%E2%80%932023.pdf.

The way that governments tend to operate – and how Finland operated in the past – is by separating functions and services into distinct domains, or ministries, which results in siloed approaches. The AuroraAI programme sees this as antithetical to a human-centric approach and efforts to improve the holistic wellbeing of its citizens, as wellbeing is multi-dimensional and, thus, dependent upon multiple domains. The AuroraAI programme seeks to re-orient the provision of services around citizens and businesses by combining data from multiple domains and building a network of AI citizen-focused applications that provide services when they are needed – around various business activities or life stages and events such as childbirth, buying a home or

retirement.[118] By bringing data together to build human-centred services, "data-based situational awareness facilitates the targeting of effective services based on individuals' real needs and enables people to manage their lives more efficiently in various life circumstances" (AuroraAI, 2019). This is facilitated by the use of reinforcement learning (see discussion in Chapter 2), through which network applications refine themselves based on feedback from users. The AuroraAI network is designed to include not only public services but also private and civil society services.

At present, AuroraAI focuses on three life events for pilots:

- moving away to study
- remaining in the labour market through lifelong learning
- ensuring family wellbeing after a divorce.

Each of these pilots and their associated target problems, pilot tests, and resulting opportunities, findings and outputs are detailed in the AuroraAI implementation plan.[119]

All of this is facilitated by a digital persona of AuroraAI users called DigiMe (Box A.1).

Box A.1: DigiMe

DigiMe refers to the ability of citizens to create a digital twin (or twins) of themselves. These digital personas allow users to manage their own data and use them to create situational profiles in order to access personalised services.

The AuroraAI network uses the collective of these personas in an anonymised way to identify similarities, differences and patterns. These findings are then used to better predict and tailor the resources needed to provide anticipatory and personalised services to citizens.

This is done through the use of reinforcement learning, whereby the system identifies which services are needed for which individuals and which times. Over time, the system collects feedback about what is helpful and what is not and automatically adjusts the services offered to be more precise.

Source:
https://vm.fi/documents/10623/1464506/AuroraAI+development+and+implementation+plan+2019%E2%80%932023.pdf.

Importantly, Finland's general AI strategy when it comes to citizens follows the principles of "MyData", under which the citizen and no one else is the owner of their own data. As the owner, a citizen has full control of their own data. They are empowered to opt in and out of services and to make decisions about with whom they share their data (AuroraAI, 2019).

In April 2019, the government published *AuroraAI – Towards a Human-Centric Society*,[120] which provides a five-year (2019-23) implementation plan for AuroraAI. The plan was developed in partnership with an open network of more than 330 members

[118] See https://youtu.be/IZU_ptEr4eE for a presentation on AuroraAI by programme lead Aleksi Kopponen.

[119] https://vm.fi/documents/10623/1464506/AuroraAI+development+and+implementation+plan+2019%E2%80%932023.pdf.

[120] https://vm.fi/documents/10623/1464506/AuroraAI+development+and+implementation+plan+2019%E2%80%932023.pdf.

from municipalities, provinces, civil society organisations and businesses. Through the plan, the authors propose to the next government a number of actions for more fully implementing the AuroraAI programme. These include the following:

- Allocate funding of EUR 100 million spread across 2020-23 to launch 10-20 services around life events and business practices.
- Launch a consulting process with citizens and businesses to identify the highest priority life events and business practices, which would inform funding and service selection activities.
- Establish a change support team and a central AuroraAI response centre to support organisations in implementing changes that bring about the AuroraAI service model.

The plan also calls for a regulatory sandbox to experiment with citizen-authorised MyData in a controlled way, as well as to explore whether any legislative changes are needed to reach the full potential of AuroraAI.

The short-term adoption and long-term path for AuroraAI have been affirmed in several recent strategic government documents. The June 2019 report *Leading the Way into the Age of Artificial Intelligence* commits the government to work to "ensure human-centric introduction of artificial intelligence and the implementation of ethical principles in the public sector through the AuroraAI project" over the next year. In addition the Prime Minister launched a new government programme entitled *Inclusive and Competent Finland*[121] that, among many things, affirms that "secure and ethically sustainable development of the AuroraAI network will be continued, as permitted by the overall spending limits, in order to make everyday life and business easier".

121 http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/161664/Inclusive%20and%20competent%20Finland_2019.pdf.

Canada's "bomb-in-a-box" scenario: Risk-based oversight by AI

Issue

Transport Canada is the department responsible for the Government of Canada's transportation policies and programmes.[122] It works to promote safe, secure, efficient and environmentally responsible transportation.

Each year, Transport Canada's Pre-load Air Cargo Targeting (PACT) team receives nearly one million pre-load air cargo records per year, containing information such as shipper name and address, consignee name and address, weight and piece count. Each record may include anywhere from 10-100 fields, depending on the air carrier and business model of the shipper. One employee, working at an unrealistic rate of one record per minute, would not even have enough time to review 10 percent. To date, very few governments have the dedicated resources to scan air cargo records for risk before loading, and of those that do, none use AI. Transport Canada decided to improve on this situation and, thereby, enhance the safety and security of air cargo transportation.[123]

Response

Transport Canada is adopting AI to enhance processes and procedures, thereby freeing up employees to work on more highly valued tasks. The department started by exploring the use of AI for risk-based reviews of air cargo records, which could be scaled to other areas if successful.

To achieve this, the department assembled a multi-disciplinary team consisting of members of PACT and the Digital Services and Transformation divisions of the department, one of Canada's Free Agents[124] and partners from an external IT firm with expertise in AI. For the pilot, Transport Canada attempted to answer two questions related to its performance:

- Can AI improve our ability to conduct risk-based oversight?
- How can we improve effectiveness and efficiency when assessing risk in air cargo shipments?

To answer these questions, the innovation team developed and implemented a two-step approach in 2018. As a first step, they used data from previous air cargo records and manual risk assessments to explore unsupervised and supervised approaches (see Chapter 2). Using the supervised approach, the team tried to understand the relationship between the inputs (cargo records) and the outcome (i.e. did this cargo record indicate a greater level of risk, as based on previous manual risk assessments?). Using unsupervised learning, the team sought to understand the relationships between all cargo inputs in order to identity rare or unusual shipments, which could be indicative of risk.

Second, the team developed a proof of concept to test natural language processing (NLP) on a different subset of data. The goal was being able to process air cargo records, and automatically tag a cargo record with a risk indicator based on the contents of the "free text" fields in the air cargo records and other structured fields. This was completed in

---

[122] www.tc.gc.ca/en/transport-canada.html.

[123] Transport Canada submitted details on this innovative project to OPSI's Case Study Platform (https://oecd-opsi.org/innovations). The content for this case is derived from their submission, which can be found at https://oecd-opsi.org/innovations/artificial-intelligence-and-the-bomb-in-a-box-scenario-risk-based-oversight-by-disruptive-technology.

[124] The Government of Canada's Free Agents programme represents an innovative departure from the permanent hiring model of the Public Service, which organises talent and skills for project-based work. See the OPSI report at https://oe.cd/innovation2018 for a full case study.

the first quarter of 2018 and showed that NLP could successfully sort cargo data into meaningful categories in real time.

Both steps led to new insights about hidden patterns that can indicate risk. As a result, the team was able to use AI to automatically generate accurate risk indicators. Through this pilot, Transport Canada learned that AI was indeed a viable solution to address its key questions. The department is now working to implement the approach throughout their risk assessment process. Since the testing phase the team has produced a dashboard and a first version of a targeting interface for identifying potentially high-risk cargo.

The team is careful to point out that AI is not going to replace human activity. The AI will handle triage, filtering and prioritisation, which is currently done using simple Excel filters. The AI is better and more efficiently able to detect anomalies, shifting trade patterns and nuances in a way that a simple Excel sheet could not.

The next step for the team is to conduct A/B testing, which will compare the current methodology with the AI-enhanced methodology. If successful, a production-ready, AI-enhanced targeting system could be ready as early as March 2021.

Results and impact

The initial results were very promising. Because every single cargo record can be addressed, instead of the small subset possible with manual assessments, AI has the potential to increase safety and security 15-fold. In addition, PACT can use AI to increase capacity, while minimising the number of people required to do the work, thus making better use of resources.

Before the introduction of AI, conducting risk assessments was burdensome and very time consuming. It took thousands of hours per year to import, clean and archive data. Dedicated resources had to be in place to analyse cargo records. With the introduction of AI, much of this process has been automated and risk assessments are conducted in real time. Artificial Intelligence helps PACT to meet its security outcomes and makes it possible for them to scan more cargo message from more carriers than ever before.

The innovation team sees this model as highly replicable. There are preliminary discussions underway within the government about adapting the approach to other modes of transportation (e.g. marine, rail, road, etc.) or even expanding it to support the mandate for Canada's agency responsible for customs and the border. The team says that, ideally, all government departments with an interest in the safety and security of Canada – including intelligence, border and law enforcement agencies – would have access to a single database with information that could be used to optimise the process for providing risk-based oversight to cargo shipments. Thinking bigger, it could be something to be leveraged internationally.

Challenges and lessons learned

As with all AI projects, this pilot was fuelled by data. PACT already had access to significant amounts of data; however, these were not in a format that facilitated the use of AI. Before the AI portion of the project could begin, the team had to clean and wrangle the data into a format that could be consumed by the AI algorithm. To support scaling of the project, the team is working to address this challenge by creating a data pipeline which will feed all cargo records received by Transport Canada into a single database in a format that is immediately machine consumable.

Given the risk aversion around disruptive technologies in general, it was also essential for the project to have support from Transportation Canada's senior management. The innovation team found that support from the Deputy Minister

The European Commission's Ethical Guidelines for Trustworthy AI

> *It is essential that trust remains the bedrock of societies, communities, economies and sustainable development. We therefore identify Trustworthy AI as our foundational ambition, since human beings and communities will only be able to have confidence in the technology's development and its applications when a clear and comprehensive framework for achieving its trustworthiness is in place.*
>
> *Ethical Guidelines for Trustworthy AI*

Issue

The European Commission (EC) has set forth a European vision for AI. One of its key goals is to increase public and private investment and boost its uptake across Europe (European Commission, 2018b). Artificial Intelligence, especially some types of Machine Learning, raises new types of ethical and fairness concerns compared to previous technologies. These concerns are likely to grow as Machine Learning becomes more ubiquitous as a result of ever-growing amounts of data and processing power. The OECD states that one of the key challenges of AI is to ensure that systems are trustworthy and human-centric, and has found that national policies are needed to promote trustworthy AI systems (OECD, 2019).

Response

In April 2019, the EC published *Ethics Guidelines for Trustworthy AI*[125] to provide guidance on how to design and implement AI systems in an ethical and trustworthy way.

The *Guidelines* were created by the EC's High-Level Expert Group on Artificial Intelligence (AI HLEG), which consists of 52 AI experts from academia, civil society and industry. One of the core tasks of the AI HLEG has been to propose AI ethics guidelines that consider issues such as fairness, safety, transparency, the future of work, democracy, privacy and personal data protection, dignity and non-discrimination, among others.

The *Guidelines* maintain that trustworthy AI has three components that work in harmony:

- **Lawful**. The AI should comply with all applicable laws and regulations.
- **Ethical**. The AI should adhere to ethical principles and values.
- **Robust**. The AI should avoid unintentional harm from both a technical and social perspective .

In addition, the *Guidelines* were developed under the premise that AI ethics are based on fundamental human rights, as set forth in EU rules and international human rights law. Accordingly, the process surfaced four ethical principles that should be considered when designing and deploying AI (see Box A.2).

[125] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Box A.2: Ethical principles and imperatives identified by the *Guidelines*

1. Respect for human autonomy

Humans interacting with AI systems must be able to retain self-determination. AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement and empower human skills. The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice. This means securing human oversight over AI systems. It should support humans in the working environment and aim for the creation of meaningful work.

2. Prevention of harm

AI systems should not cause or exacerbate harm or otherwise adversely affect human beings. This entails the protection of human dignity and mental and physical integrity. AI systems and their environments must be safe and secure. They must be technically robust and not open to malicious use. Vulnerable persons should receive greater attention and be involved in AI development, deployment and use. Particular attention must be paid to adverse impacts due to asymmetries of power or information, such as between governments and citizens.

3. Fairness

The development, deployment and use of AI systems must be fair. Substantively, this implies a commitment to 1) ensuring equal and just distribution of both benefits and costs, and 2) ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation. If unfair biases can be avoided, AI systems could even increase societal fairness. Equal opportunity to access education, goods, services and technology should also be fostered. The use of AI should never lead to people being deceived or unjustifiably impaired in their freedom of choice. Fairness implies that AI practitioners should respect the principle of proportionality between means and ends, and consider carefully how to balance competing interests and objectives. Procedurally, fairness entails the ability to appeal decisions made by AI systems and the humans operating them.

4. Explicability

Processes need to be transparent, the capabilities and purpose of AI systems must be openly communicated, and the resulting decisions must be explainable to those affected, to the extent possible. Otherwise, a decision cannot be contested. However, an explanation as to why and how a model has generated a particular decision is not always possible. These cases are referred to as "black box" algorithms and require special attention. In those circumstances, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights. The degree to which explicability is needed is dependent on the context and the consequences if that output is inaccurate.

Source: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=58477 (as excerpted by the OECD).

To guide organisations interested in using AI, the *Guidelines* provide seven requirements that AI systems should meet in order to be considered trustworthy (see Box A.3). These requirements are designed to help organisations actualise the four key principles.

Box A.3: Seven requirements for trustworthy AI

1. **Human agency and oversight**
    a) If an AI system has the potential to negatively affect human rights, a fundamental rights impact assessment should be conducted before development.
    b) Humans must be able to make informed autonomous decisions regarding AI systems and, when needed, challenge the systems. Individuals must also have the right to not be subject to a solely automated decision if it significantly affects them.
    c) Humans must have the ability to oversee (to varying degrees based on application area and risk) AI systems.
2. **Technical robustness and safety**
    a) AI systems must be developed in a manner that seeks to prevent risks, promote reliable behaviour and minimise or prevent harm.
    b) Security processes should be in place to protect AI systems against vulnerabilities that can be abused (e.g. hacking) and prevent unintended applications of the system.
    c) Processes should be in place to assess potential risk, and AI systems should have safeguards in case of problems (e.g. requirements for human intervention in some circumstances).
    d) An explicit process should be in place to address unintended risks from inaccurate results and predictions.
    e) AI system results must be reproducible and reliable.
3. **Privacy and data governance**
    a) AI systems must guarantee privacy and data protection through their life cycle to prevent unlawful or unfair discrimination.
    b) Efforts must be made to ensure the quality of data and address any biases, inaccuracies and errors. Data integrity must also be ensured to protect against, for instance, malicious data being fed into a system.
    c) Organisations should have data governance protocols in place that govern data access.
4. **Transparency**
    a) Datasets and decision-making processes should be documented to the extent possible to allow for traceability and transparency.
    b) Where AI systems impact lives, those affected should be able to demand an explanation of the decision-making process.
    c) Humans have the right to be informed that they are interacting with an AI system, and about the system's capabilities and limitations.
5. **Diversity, non-discrimination and fairness**
    a) Processes and procedures should be in place to address and remove biases at the data collection phase when possible, as well as oversight mechanisms to monitor the systems.

b) Depending on the purpose of the AI system, all users should have equitable access to AI products, regardless of their demographics or characteristics.

c) Organisations should consult stakeholders who may be affected by an AI system throughout its life cycle in order to obtain regular feedback.

6. Societal and environmental wellbeing

a) Organisations should make sustainable decisions for AI systems (e.g. energy consumption), taking into consideration their full life cycle and supply chain.

b) The social effects of AI systems (e.g. social agency, social relationships) should be monitored and considered.

c) The societal and democratic effects of AI systems should be considered (e.g. effect on institutions, democracy, and society).

7. Accountability

a) It should be possible to assess algorithms, data and design processes.

b) Organisations should seek to identify, assess, document and minimise the potential negative impacts of AI systems.

c) Methods should be put in place to negotiate, evaluate and document instances where tensions arise among these requirements, and where trade-offs may need to be made. If an ethical trade-off is not possible, the AI system should not proceed in that form.

d) Mechanisms should be in place that ensure individuals have the right to redress when an unjust adverse impact occurs.

The *Guidelines* recommend that these requirements be continuously evaluated and addressed throughout an AI system's life cycle. To help organisations meet these requirements, the *Guidelines* walk describe the technical methods (e.g. systems architecture, explanation methods), and non-technical methods (e.g. stakeholder participation, codes of conduct) necessary to achieve them.

Source: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=58477 (as adapted by the OECD).

Finally, the *Guidelines* provide a concrete "assessment list" designed to help developers of user-facing AI systems operationalise the key requirements laid out in Box A.4. This list is currently undergoing a digital pilot process[126] for testing and validation. Anyone interested in the list is invited to provide feedback on ways in which it can be strengthened. The EC plans to evaluate all feedback received by the end of 2019 for incorporation into a new version in 2020. Box A.4 provides select excerpts from the assessment list. Readers of this guide are encouraged to view and consider the full list in the *Guidelines* document.[127]

[126] https://ec.europa.eu/futurium/en/ethics-guidelines-trustworthy-ai/register-piloting-process-0.
[127] https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=58477.

Box A.4: Selected examples from the Trustworthy AI Assessment List (pilot version)

1. **Human agency and oversight**
   a) Did you consider whether the AI system should communicate to users that a decision or outcome is the result of an algorithmic decision?
   b) Did you consider the task allocation between the AI system and humans for meaningful interactions?
   c) Did you ensure a stop button or procedure to abort an operation if needed? Does this procedure abort the process or delegate control to a human?
2. **Technical robustness and safety**
   a) Did you put measures or systems in place to ensure the integrity and resilience of the AI system against potential attacks?
   b) Did you ensure that your system has a sufficient fallback plan if it encounters attacks or other situations (e.g. technical switching procedures or asking for a human operator before proceeding)?
   c) Did you verify what harm would be caused if the AI system makes inaccurate predictions?
   d) Did you test whether specific contexts or particular conditions need to be taken into account to ensure reproducibility?
3. **Privacy and data governance**
   a) Did you consider ways to develop the AI system or train the model with minimal use of potentially sensitive data?
   b) Did you align your system with relevant standards (e.g. ISO, IEEE) or widely adopted protocols for daily data management and governance?
   c) Does the system log when, where, how, by whom and for what purpose data were accessed?
4. **Transparency**
   a) Did you establish measures that can ensure traceability, such as documenting methods for training the algorithm?
   b) Did you assess to what extent the decisions and hence the outcome made by the AI system can be understood?
   c) Did you clearly communicate characteristics, limitations and potential shortcomings of the AI system?
5. **Diversity, non-discrimination and fairness**
   a) Did you consider diversity and representativeness of users in the data? Did you test for specific populations or problematic use cases?
   b) Did you assess whether the team involved in building the AI system is representative of your target user audience?
   c) Did you consider a mechanism to include the participation of different stakeholders in the AI system's development and use?
6. **Societal and environmental wellbeing**

a) Did you ensure measures to reduce the environmental impact of your AI system's life cycle?

b) Did you ensure that the AI system signals that its social interaction is simulated and that it has no capacities of "understanding" and "feeling"?

c) Did you assess the broader societal impact of the AI system's use beyond the individual user (e.g. indirectly affected stakeholders)?

7. Accountability

a) Where application affects fundamental rights did you ensure that the AI system can be audited independently?

b) Did you establish processes for third parties or workers to report potential vulnerabilities, risks or biases?

c) How do you decide on trade-offs? Did you ensure that the trade-off decision was documented?

d) Did you establish an adequate set of mechanisms that allows for redress in case of the occurrence of any harm or adverse impact?

Source: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=58477 (as adapted by the OECD).

While quite comprehensive, AI HLEG authors are careful to note that the *Guidelines* will need to be adapted for each specific situation. This is important, for instance, because some AI applications are more sensitive than others.

The *Guidelines* were developed through an open and participatory process. An initial draft was published for public consultation in December 2018, and received over 500 comments from a diverse set of respondents, ranging from businesses and civil society organisations to members of the general public.[128] This comments were taken into consideration in the creation of the final *Guidelines* document.

[128] See https://ec.europa.eu/digital-single-market/en/news/over-500-comments-received-draft-ethical-guidelines-trustworthy-artificial-intelligence for details on the public consultation, including a summary of the feedback received.

Canada's Directive on Automated Decision-Making

These are your guard rails for responsible automation.[129]

Alex Benay, Chief Information Officer, Government of Canada

Issue

The Government of Canada (GC) is increasingly looking to utilise Artificial Intelligence to make or help make administrative decisions to improve service delivery.[130] A key issue in this regard, however, is the potential for biases, ethical issues and other considerations as government organisations advance in their adoption of AI, as well as questions about the extent to which humans should be involved in AI-based decision making. Existing laws and policies were unclear on how to handle these scenarios.

These problems first manifested with the launch of government pilot projects to develop advanced algorithms to help triage temporary resident visa applications from China and India. The number of applications had increased significantly, putting a strain on processing. The projects drew the attention of the media leading to a public discussion about the appropriateness of using automation to make decisions that can affect people.[131] In the absence of guidance to determine acceptable practices, the pilot projects ground to a halt and were unable to determine a way forward (Wright, 2018).[132]

With the growth in AI-enabled services, including those making automatic decisions that affect the lives of humans, the government is seeking to ensure that such decisions minimise the risks to citizens as well as to public sector organisations at the Federal level. The aim is to provide a clear ethical baseline for government organisations in cases related to automated decision making, in order to prevent situations such as the one above.

Response

The Government of Canada crowdsourced research from hundreds of computer science experts and government officials to develop the white paper *Responsible Artificial Intelligence in the Government of Canada*,[133] and on 1 April 2019 published its Directive on Automated Decision-Making.[134] The Directive provides a risk-based approach to ensuring the transparency, accountability, legality and fairness of automated decisions that affect Canadians, and imposes certain requirements for the government's use of decision-making algorithms and systems. The Directive is the first of its kind in the world, and will take effect across the Federal Government (with the exception of a few exempted agencies) from April 2020.

The Directive only applies to automated decision-making systems that are public-facing, such as benefits programmes that decide whether applicants meet qualification requirements. It does not yet cover internal government services or national security issues.

The backbone of the Directive is the Algorithmic Impact Assessment (see Box A.5), which agency leaders will be responsible for completing before producing or significantly changing an automated decision system. This helps them to pre-empt

[129] https://askai.org/blog-podcast-canada-cio-alex-benay-is-on-a-mission-to-modernize.
[130] www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.
[131] www.cbc.ca/news/politics/human-rights-ai-visa-1.4838778.
[132] Interview with Michael Karlin, Team Lead, Data Policy at Department of National Defence, Government of Canada, 5 June 2019.
[133] https://docs.google.com/document/d/1Sn-qBZUXEUG4dVk909eSg5qvfbpNlRhzIefWPtBwbxY.
[134] www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

issues and put in place systems and processes to monitor implementation. Leaders are required to publicly publish the results of their assessments online as open government data. GC officials emphasise the importance of this component, which provides a comprehensive, publicly available overview of existing automated decision programmes.

Box A.5: Algorithmic Impact Assessment

The Algorithmic Assessment is a digital questionnaire that evaluates the potential risk of a public-facing automated decision system. It assesses the decisions the system has the capacity to make or inform and the potential harm to citizens. The results of the questionnaire generate a risk rating on a scale of 1-4 for the decision-making system: 1 indicates decisions leading to impacts that are brief and reversible, and 4 indicates decisions leading to potential impacts that are irreversible and significant. This rating establishes the minimum level of responsibility for the organisation, and assigns mandatory governance, oversight and reporting requirements.

Sample questions include:

- Is the project within an area of intense public scrutiny?
- Are clients in this line of business particularly vulnerable?
- Will the algorithmic process be difficult to interpret or to explain?
- Will the system be replacing a decision that would otherwise be made by a human?
- Are the impacts resulting from the decision reversible?
- Who collected the data used for training the system?
- Will you have documented processes in place to test datasets against biases and other unexpected outcomes?
- Will the system provide an audit trail that records all the recommendations or decisions made by the system?
- Will the system be able to produce reasons for its decisions or recommendations when required?
- Will the system enable human override of system decisions?

Based on the answers to these and other questions, the assessment specifies the required response based on potential risk. For example, it determines the extent to which there is a need for:

- peer review of the system
- public notice about the system
- human involvement during the decision-making process
- explanation of how decisions are made
- testing the system and monitoring outcomes for unexpected outcomes (e.g. bias)
- training staff so they understand and oversee the system
- contingency planning
- mitigation measures.

Use of the assessment tool will be mandatory in Canada as of April 2020.

The tool is available as free and open source software (FOSS) on GitHub. The Government of Canada is encouraging governments of other countries, experts and community groups to participate in the continuous development and evolution of the

tool, and inciting them to adapt the tool to fit their own institutional and cultural contexts.

Source: https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592, https://canada-ca.github.io/aia-eia-js; interviews with GC officials.

In addition, the Directive requires government organisations to release the custom source code of the algorithms to the extent possible, and to provide clients with applicable recourse options to appeal against decisions, among other things.

Importantly, the Directive and the Assessment Tool have been developed in an open and participatory manner. Stakeholders from all sectors, as well as members of the public, were invited to provide comments. This enabled feedback from academia, civil society organisations, private sector businesses and interested individuals to be incorporated during the development process (Government of Canada, 2019a).

Results and impact

Although the Directive will not come into full effect until April 2020, government organisations have already changed their behaviours to ensure compliance. This includes completing the Algorithmic Impact Assessment and following the risk-based requirements set forth by the risk rating determined by the Assessment. GC officials believe that all recently initiated or updated automated decision systems are already in compliance with the Directive and the draft Assessment.

The impact of Canada's approach is already diffusing internationally. Every member of the D9[135] – a network of the world's most advanced digital nations – is considering adopting the Algorithmic Impact Assessment (Greenwood, 2019). The governments of Germany and Mexico have both moved to adopt modified versions of the Algorithmic Impact Assessment tailored to their own contexts.

Challenges and lessons learned

The GC team who designed the Directive and Impact Assessment did not face many significant challenges during the development and approval stages. They attribute this to a few key factors in their approach:[136]

- **Timing**: The team started their work at a moment when AI was receiving less attention. This allowed them to formulate initial concepts and iterations without undue attention from many intensely interested actors.
- **Working in the open**: According to GC officials, openness was critical to securing broad acceptance of the Directive, as well as the risk-based classifications determined by the Algorithmic Impact Assessment. This open process also quickened development of the Directive and the Assessment tool by leveraging the expertise of outside contributors.
- **Doing before showing**: To secure political support, the government team created an early draft of the Directive and a prototype of the Algorithmic Impact Assessment, and presented them to senior leaders. This approach worked better than trying to describe the concept and associated need in a proposal.

[135] See www.digital.govt.nz/digital-government/international-partnerships/the-digital-9.

[136] Interview with Michael Karlin, Team Lead, Data Policy at Department of National Defence, Government of Canada, 5 June 2019.

Further broader lessons can be drawn from the government's experience in developing this Directive:

- **Guide and empower instead of constrain and restrict**: While public organisations are required to comply with the Directive, the Directive itself does not prevent them from undertaking (or restrict) the development of their own automated decision-making system. The Directive does, however, empower authorities and third parties (private sector or civil society) to raise important questions about the proposed system which developers then are required to answer (e.g. what data is used, how it is used and does the system comply with existing regulations).
- **AI is not the end game**: Although the advent of Artificial Intelligence and its challenges provided the impulse for the Directive, the initiative must be seen as part of a broader effort from the government to protect its citizens while also promoting innovation for better public services.

United States Federal Data Strategy and Roadmap

> *The mission of the Federal Data Strategy is to leverage the full value of Federal data for mission, service, and the public good by guiding the Federal Government in practicing ethical, governance, conscious design, and a learning culture.*
>
> *US Federal Data Strategy*

Issue

The United States Government is one of the largest entities in the world, which can make it challenging to manage data as an asset in a consistent manner at an enterprise level. Several new laws have been passed recently that seek to address this situation, which could be complemented by uniform policy guidance in the executive branch to help ensure consistent implementation. In addition to new laws, the US President has identified "Leveraging Data as a Strategic Asset" as a presidential priority area and a Cross-Agency Priority Goal (CAP Goal), which necessitates a systems approach to data in government.[137]

Response

On 4 June 2019, the White House Office of Management and Budget (OMB) launched the Federal Data Strategy (Strategy) as a government-wide framework to help promote consistency and quality in data infrastructure, governance, actions, protection and security. The Strategy was created by a cross-government team[138] and represents a ten-year vision for how the government will "accelerate the use of data to support the foundations of democracy, deliver on mission, serve the public, and steward resources while protecting security, privacy and confidentiality".

The Strategy consists of **10 principles** organised around three categories that serve as motivational guidelines for government agencies (see Box A.6).

[137] www.performance.gov/CAP/leveragingdata. CAP Goals are long-term goals used by leadership to accelerate progress on a limited number of Presidential priority areas where implementation requires active collaboration among multiple agencies. They seek to drive cross-government collaboration to tackle government-wide management challenges.

[138] https://strategy.data.gov/team.

Box A.6: Strategy principles

Ethical governance

1. **Uphold ethics**: Monitor and assess the public implications of leveraging data. Design checks and balances to protect and serve the public good.
2. **Exercise responsibility**: Practise effective stewardship and governance. Ensure that security, privacy, promised confidentiality, and appropriate access and use practices are in place.
3. **Promote transparency**: Articulate the purpose and uses of all data to engender public trust. Document all processes and products to inform data providers and users.

Conscious design

1. **Ensure relevance**: Protect data quality and integrity. Ensure that data are appropriate, accurate, objective, useful, understandable and timely.
2. **Harness existing data**: Identify needs to inform research and policy questions, reusing data if possible.
3. **Anticipate future uses**: Consider and plan for reuse and interoperability from the start.
4. **Demonstrate responsiveness**: Improve data with input from users. Using a cyclical feedback process, establish a baseline, gain support, collaborate and refine continuously.

Learning culture

1. **Invest in learning**: Promote a culture of continuous and collaborative learning through ongoing investment in infrastructure and human resources.
2. **Develop data leaders**: Cultivate leadership at all levels by investing in training on and development of data value for missions, service and the public good.
3. **Practice accountability**: Assign responsibility, audit data practices, document and learn from results, and make needed changes.

Source: www.whitehouse.gov/wp-content/uploads/2019/06/M-19-18.pdf (edited for brevity).

Alongside the principles, the Strategy elaborated 40 practices to guide agencies on how to leverage the value of federal data, as well as data sponsored by the federal government. The practices take into account the different uses of data available to better achieve public value. In so doing, the practices seek to align data management with these uses in order to address the needs of both the government and stakeholders/users. The practices are clustered around three core categories. Box A.7 lists these categories and presents a sample of practices set forth in the Strategy. All 40 practices are listed in the full Strategy.

Box A.7: Selected Strategy Practices

Building a culture that values data and promotes public use: This practice focuses on using data for government decision making and supporting external use.

1. **Identify data needs to answer key agency questions**: Identify and prioritise key questions and the data needed to answer them.
2. **Monitor and address public perception**: Regularly assess public confidence in terms of value, accuracy, objectivity and privacy protection in order to make improvements and advance missions.
3. **Connect data functions across agencies**: Establish communities of practice for common functions (e.g. data management, analytics), and to promote efficiency, collaboration and coordination.

Governing, managing and protecting data: This practice focuses on data governance across agencies.

1. **Prioritise data governance**: Ensure efficient authorities, roles, structures, policies and resources to transparently support the management, maintenance and use of data.
2. **Allow amendment**: Establish clear procedures to allow members of the public to access and amend data about themselves, as appropriate.
3. **Share data between state, local and tribal governments and Federal agencies**: Facilitate data sharing, particularly for programmes that are Federally funded and locally administered.

Promoting efficient and appropriate data use: This practice focuses on providing access to data resources (e.g. sharing data, open data), promoting their appropriate use (documenting and protecting data), and providing guidance on data augmentation (data quality, metadata, secure linkages).

1. **Increase capacity for data management and analysis**: Educate the workforce through training, tools, communities and expanding capacities.
2. **Align quality with intended use**: Ensure that data likely to inform important policy or private sector decisions are of appropriate utility, integrity and objectivity.
3. **Diversify data access methods**: Invest in multiple tiers of access to make data as accessible as possible.

Source: www.whitehouse.gov/wp-content/uploads/2018/10/M-19-01.pdf.

In order to make these practices actionable, the Strategy requires that agencies adhere to the requirements of annual government-wide action plans, which prioritise practices for a given year and provide timelines and designate responsibilities. Alongside publication of the Strategy, the United States released a draft version of the first of these plans, the *2019-2020 Federal Data Strategy Action Plan*.[139] The draft, which involved a three-week public and stakeholder consultation period lasting until 8 July 2019, listed

[139] The draft plan is available at https://strategy.data.gov/assets/docs/draft-2019-2020-federal-data-strategy-action-plan.pdf. It is expected that the final plan will be published at https://strategy.data.gov/action-plan.

a series 16 concrete actions across government, including a focus on Artificial Intelligence. It consists of three types of actions:

5. **Shared** – led by a single agency for the benefit of all agencies
6. **Community** – actions taken by a group of agencies around a common topic
7. **Agency-specific** – actions for a single agency to build capacity in that agency.

Figure A.4: Relationship between the Strategy and the Action Plan

Source: https://strategy.data.gov/assets/docs/draft-2019-2020-federal-data-strategy-action-plan.pdf.

Each action listed in the plan explicitly sets forth:

- The practice with which the action is associated
- The responsible office
- The timeline for completion
- How success will be measured.

Several sample actions are included in Table A.1. The US government anticipates that a final action plan will be released in September 2019.

Table A.1: Sample actions

Action	Description	Deadline
Improve data resources for AI research and development	Improve data and model inventory documentation to enable discovery and usability, and prioritise improvements to access and quality of AI data and models based on user feedback from the AI research community.	February 2020
Develop a data ethics framework	Establish a consistent framework for evaluating ethical repercussions and trade-offs associated with data management and use.	November 2019
Identify opportunities to increase staff data skills	Identify critical data skills for each agency. Assess current staff capacity for critical data skills. Develop an initial plan to address gaps between critical data skill needs and current capacity.	May 2020

Source: https://strategy.data.gov/action-plan.

The Public Policy Programme at The Alan Turing Institute (United Kingdom)

Issue

Governments have access to vast quantities of data. These data may be collected through processes to develop official statistics, such as through surveys. However, a great deal of data is also generated through government's day-to-day interactions and transactions with citizens and businesses. AI has enormous potential for governments to harness these vast quantities of data and provide officials with unprecedented insights, enabling them to improve policy-making processes and to make public services more efficient (Margetts and Dorobantu, 2019). For instance, AI could:

- provide more accurate forecasts and predictions, enabling governments to plan more effectively and to target resources and services where they are most needed
- tailor public services to user need, allowing governments to adapt the services they offer to individual circumstances
- simulate complex systems, from military operations to housing markets, giving governments the opportunity to experiment with policy options and spot unintended consequences before committing to new measures.

Despite the promise that AI holds for policy making and service delivery, developing and retaining internal expertise on emerging technologies and their applications within the public sector is a challenge for every government. In the past, governments have struggled to adopt much simpler technologies, such as electronic payroll systems or online appointment booking systems (Margetts and Dorobantu, 2019).

Moreover, drawing on external expertise comes with its own set of difficulties, such as effective contract management, moving data across organisational boundaries, and wider cultural issues around different organisational priorities and ways of working. When buying algorithms and AI technologies from external providers, governments struggle to determine value-for-money and to evaluate how well these complex technologies perform.

Response

The Alan Turing Institute is the United Kingdom's national institute for data science and artificial intelligence. Founded as a charity in 2015, with a focus on data science, the Institute added AI to its remit in 2017.[140] The goals of The Alan Turing Institute are set out in Box A.8.

[140] www.turing.ac.uk/about-us.

Box A.8: The Alan Turing Institute's goals

- **Advance world-class research and apply it to real-world problems**: innovate and develop world-class research in data science and artificial intelligence that supports next-generation theoretical developments and is applied to real-world problems, generating the creation of new businesses, services and jobs.
- **Train the leaders of the future**: provide training for new generations of data science and AI leaders with the necessary breadth and depth of technical and ethical skills to match the United Kingdom's growing industrial and societal needs.
- **Lead the public conversation**: through agenda-setting research, public engagement and expert technical advice, drive new and innovative ideas which have a significant influence on industry, government, regulation or societal views, or which have an impact on how data science and artificial intelligence research are undertaken.

Source: www.turing.ac.uk/about-us.

The Alan Turing Institute partners with 13 leading research universities from across the United Kingdom, as well as the Engineering and Physical Sciences Research Council. Close to 500 academics hold appointments at the Institute, giving it unparalleled access to expertise in a variety of disciplines, from computer science and mathematics to social science and philosophy.[141]

The Alan Turing Institute's mission is "to make great leaps in data science and artificial intelligence research in order to change the world for the better". As an important step towards fulfilling its mission, the Institute launched a Public Policy research programme in May 2018. The programme works alongside policy makers to develop AI research, tools and techniques that have a positive impact on the lives of as many people as possible.[142] The Public Policy Programme's challenges are set out in Box A.9.

[141] www.turing.ac.uk/about-us.

[142] /www.turing.ac.uk/research/research-programmes/public-policy.

Box A.9: The Alan Turing Institute's Public Policy Programme Challenges

- **Use data science and Artificial Intelligence to inform policy making**. In a world of changing and interlinked policy measures, data science and AI can provide policy makers with unprecedented insights ranging from identifying policy priorities by modelling complex systems and scenarios, to evaluating hard-to-measure policy outcomes. The Public Policy Programme's aim is to equip policy makers across all levels of government with the tools they need to not only design effective public policy, but also to track and measure policy impacts.
- **Improve the provision of public services**. Governments today are major holders of data which data science and AI can harness to improve the design and provision of public services. The Public Policy Programme brings researchers and policy makers together in order to develop innovative ways to provide public services. The programme's aim is to change everyday life for the better from allocating resources in the fairest and most transparent way, to designing personalised public services tailored to people's individual needs and situations.
- **Build ethical foundations for the use of data science and AI in policy making**. Understanding the ethical and societal implications of data science and AI is an integral part of the development of these technologies. The Public Policy Programme works with policy makers to develop the ethical foundations for the use of data science and AI in the public sector, with the aim of securing the benefits and addressing the risks these technologies pose.
- **Contribute to policy that governs the use of data science and AI**. The effects of data science and AI on society are already being felt, and their impact will only grow in the years to come. The Public Policy Programme works alongside governments and regulators to develop well-crafted laws and sensible regulation, with the aim of ensuring that the impact of these powerful technologies is as beneficial and equitable as possible.

Source: www.turing.ac.uk/research/research-programmes/public-policy.

A core team of researchers with a social science background oversee the Public Policy Programme's activities. Their academic expertise ranges from philosophy and law to economics and international relations.

The programme is uniquely positioned to work alongside policy makers on the use of data science and AI for the greater good. In the United Kingdom, the programme provides a single point of contact for the government to draw on the country's leading academic experts in data science and AI. By providing impartial advice, independent academic researchers are particularly well placed to help governments maximise the potential of these technologies to solve public policy problems (Margetts and Dorobantu, 2019).

Through the Public Policy Programme, civil servants have access to The Alan Turing Institute's growing network of academics. The programme's core team of researchers link up policy makers with academics to discuss the problems they are facing and to ascertain the best available methodologies and tools to address these issues. Academic researchers can then work at the Institute to develop the relevant tools or can be embedded in public sector teams to deliver a project. At the end of each collaboration, the programme hands over the developed tools and techniques to civil servants, who are trained by the academics to take over the ownership of the projects.

Results and impact

The Alan Turing Institute is an integral part of the UK government's approach to AI, which was set out in the AI Sector Deal as part of the country's Industrial Strategy (Gov.UK, 2019b). Public sector bodies can consult The Turing's Public Policy Programme to obtain to trusted, independent advice on AI and data science, including ethical issues. The European Commission has noted the value of such institutes and programmes in its flagship policy report on AI, entitled *Artificial Intelligence: A European Perspective*.[143]

The Public Policy Programme has been in existence for a little over a year but it has already had a substantial impact. The programme has provided advice and guidance to hundreds of policy makers, representing more than 80 organisations, from local councils and police forces to central government departments, regulators and international organisations. The programme has also contributed to key policy initiatives in the United Kingdom, including the set-up of the UK Government's Centre for Data Ethics and Innovation[144] and the Government's Technology Innovation Strategy.

In addition to its advisory role, the Public Policy Programme is home to over 20 multi-year research projects, involving 60 plus academic researchers from 10 universities. Each project addresses a specific policy challenge and is led by a senior academic. The research projects cover a broad range of topics, from measuring and countering online hate speech, to increasing the number of women in data science and AI, and identifying the policies that developing countries need to prioritise in order to reach the UN Sustainable Development Goals.

Some of the programme's projects are already having a direct impact in the policy world. In 2019, the Institute's Public Policy Programme partnered with the UK Government's Office for Artificial Intelligence and the Government Digital Service to produce guidance on the responsible design and implementation of AI systems in the public sector. The guide, *Understanding Artificial Intelligence Ethics and Safety*,[145] was written by Turing researchers and launched by the UK's Minister for Implementation in June 2019. It represents the world's most comprehensive guidance on AI ethics and safety for the public sector; and identifies the potential harm caused by AI systems and proposes concrete, operationalisable measures to counteract them.

The Turing's Public Policy Programme also collaborates with regulators in the United Kingdom on AI explainability and transparency. The programme is also working with the Information Commissioner's Office to develop guidance to assist organisations in explaining AI decisions to the individuals affected. An interim report was published in June and final guidance will be produced in the third quarter of 2019 (ICO, 2019). The programme also recently announced a new collaboration with the Financial Conduct Authority on a research project that will examine current and future uses of AI across the financial services sector, analyse ethical and regulatory questions that arise in this context, and provide advice on potential strategies to address them.[146]

Challenges and lessons learned

The programme is a trusted advisor and collaborator to government departments and regulators. The ability to maintain its independence in that role is of paramount

[143] https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/artificial-intelligence-european-perspective.

[144] www.turing.ac.uk/research/publications/dcms-consultation-centre-data-ethics-and-innovation.

[145] www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety.

[146] www.turing.ac.uk/news/new-collaboration-fca-ethical-and-regulatory-issues-concerning-use-ai-financial-sector.

importance, which is why the programme relies solely on public funds to support its work. The programme's team of researchers is supported through a combination of core funding from The Alan Turing Institute and grants from UK Research and Innovation, a national funding agency that invests in science and research. While this funding model has proved successful, it can sometimes limit the programme's ability to work on projects that fall outside the scope of its grants.

For research projects that lie outside grant conditions, the programme goes through public procurement processes. However, public sector procurement is not always well adapted to AI research projects. The timelines for the projects are sometimes too short, especially when government departments need to deliver concrete results within tight timeframes. In academia, post-doctoral researchers must be recruited for each new project – a lengthy process in contrast with private consultancies, which have employees that can be assigned immediately to a new project. Government contracts may also come with strict conditions that undermine academic career advancement, which is closely tied to the publication of research findings. For example, some procurement contracts do not allow for any form of publication, which diminishes the attractiveness of the work to academic researchers.

None of these challenges are insurmountable. In recognition of the need for a new framework for effective and responsible design, procurement and deployment of AI by the public sector, the UK government's Office for AI partnered this year with the World Economic Forum to co-design guidelines for AI procurement (Gerdon, 2019).

The Public Policy Programme is exploring additional avenues to fund data science and AI research and to encourage research institutions to play a greater role in public sector innovation. Potential solutions include setting up a government innovation fund or developing new funding models that allow government departments and agencies to collaborate with academic partners.

Annex B. Glossary

This is being developed during the open consultation.

References

Agrawal, A., J. Gans and A. Goldfarb (2018), *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Review Pres, Boston, MA.

AI NOW (2018), "After a Year of Tech Scandals, Our 10 Recommendations for AI", *Medium*, 6 December, https://medium.com/@AINowInstitute/after-a-year-of-tech-scandals-our-10-recommendations-for-ai-95b3b2c5e5.

Alom, Z. Md, T.M. Taha, C Yakopcic, S. Westberg, P. Sidike, M.S. Nasrin, B.C. Van Essen, A.A.S. Awaal and V.K. Asari (2018), "The history began from AlexNet: A comprehensive survey on deep learning approaches", https://arxiv.org/ftp/arxiv/papers/1803/1803.01164.pdf.

Anastasopoulos, L.J. and A.B. Whitford (2019), "Machine learning for public administration research, with application to organizational reputation", *Journal of Public Administration Research and Theory*, Vol. 29/3, pp. 491-510, https://doi.org/10.1093/jopart/muy060.

Andrews, M. (2018), *How Do Governments Build Capabilities to Do Great Things? The Oxford Handbook of the Politics of Development*, Oxford University Press, https://doi.org/10.1093/oxfordhb/9780199845156.013.34.

AuroraAI (2019), *AuroraAI – Towards a Humancentric Society*, https://vm.fi/documents/10623/1464506/AuroraAI+development+and+implementation+plan+2019%E2%80%932023.pdf.

Balaram, B., T. Greenham and J. Leonard (2018), *Artificial Intelligence: Real Public Engagement*, London, RSA, www.thersa.org/globalassets/pdfs/reports/rsa_artificial-intelligence---real-public-engagement.pdf.

Bansak, K., J. Ferwerda, J. Hainmueller, A. Dillon, D. Hangartner, D. Lawrence and J. Weinstein (2018), "Improving refugee integration through data-driven algorithmic assignment", *Science*, Vol. 359/6373, pp. 325-329, http://science.sciencemag.org/content/359/6373/325.

BBC News (2019), "Could an algorithm help prevent murders?", 24 June, www.bbc.com/news/stories-48718948.

Bryman, A. (2016), *Social Research Methods*, Oxford University Press.

Case, N. (2018), "How to become a centaur", https://jods.mitpress.mit.edu/pub/issue3-case.

Dencik, L., A. Hintz, J. Redden and H. Warner (2018), *Data Scores as Governance: Investigating Uses of Citizen Scoring in Public Services. Project Report*, Data Justice Lab/Cardiff University/Open Society Foundations, https://datajustice.files.wordpress.com/2018/12/data-scores-as-governance-project-report2.pdf.

du Preez, D. (2018), "Professor Dame Wendy Hall – 'We need to put diversity at the centre of the AI ethics debate'", *Diginomica*, 3 December, https://diginomica.com/professor-dame-wendy-hall-we-need-to-put-diversity-at-the-centre-of-the-ai-ethics-debate.

European Commission (2018a), *Artificial Intelligence: A European Perspective*, European Commission, Brussels, https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/artificial-intelligence-european-perspective.

European Commission (2018b), *Artificial Intelligence for Europe*, European Commission, Brussels, https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF.

Gardner, J. (1983, 2011), *Frames of Mind: The Theory of Multiple Intelligences*, Basic Books, New York.

Gerdon, S. (2019), "How the public sector can procure AI-powered solutions more effectively and responsibly", *DCMS blog*, 24 July, https://dcmsblog.uk/2019/07/how-the-public-sector-can-procure-ai-powered-solutions-more-effectively-and-responsibly.

Goodfellow, I, Y. Bengio and A. Courville (2016), *Deep Learning*, MIT Press, Cambridge, MA, www.deeplearningbook.org.

Government of Canada (2019a), "Ensuring responsible use of artificial intelligence to improve government services for Canadians", Press release, 4 March, www.canada.ca/en/treasury-board-secretariat/news/2019/03/ensuring-responsible-use-of-artificial-intelligence-to-improve-government-services-for-canadians.html.

Government of Canada (2019b), "Government of Canada creates Advisory Council on Artificial Intelligence", Press release, 14 May, www.canada.ca/en/innovation-science-economic-development/news/2019/05/government-of-canada-creates-advisory-council-on-artificial-intelligence.html.

Gov.UK (2019a), "Leading experts appointed to AI Council to supercharge the UK's artificial intelligence sector", Press release, 16 May, www.gov.uk/government/news/leading-experts-appointed-to-ai-council-to-supercharge-the-uks-artificial-intelligence-sector.

Gov.UK (2019b), *AI Sector Deal*, Policy paper, 21 May, www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal.

Greenwood, M. (2019), "Canada's new Federal Directive makes ethical AI a national issue", *Techvibes*, 8 March, https://techvibes.com/2019/03/08/canadas-new-federal-directive-makes-ethical-ai-a-national-issue.

Haugeland, J. (1985), *Artificial Intelligence: The Very Idea*, https://philpapers.org/rec/HAUAIT.

Herd, P. and D.P. Moynihan (2018), *Administrative Burden: Policymaking by Other Means*. Russell Sage Foundation, New York, www.jstor.org/stable/10.7758/9781610448789.

ICO (Information Commissioner's Office) (2019), *Project ExplAIn: Interim Report*, ICO, Wilmslow, UK, https://ico.org.uk/media/2615039/project-explain-20190603.pdf

IDIA (International Development Innovation Alliance) (2019), *Artificial Development in International Development: A Discussion Paper*, IDIA/AI & Development Working Group, https://static1.squarespace.com/static/5b156e3bf2e6b10bb0788609/t/5d3f283a3ee5d60001fcf184/1564420165214/AI+and+international+Development_FNL.pdf.

Kattel, R. (2019), "Do coders need a code of conduct?", *New Statesmen*, 6 June, www.newstatesman.com/spotlight/emerging-technologies/2019/06/do-coders-need-code-conduct.

Kortz, M. and F. Doshi-Velez (2017), *Accountability of AI Under the Law: The Role of Explanation.* Berkman Klein Center, Cambridge, MA, https://cyber.harvard.edu/publications/2017/11/AIExplanation.

Janssen, M. and J. van den Hoven (2015), "Big and Open Linked Data (BOLD) in government: A challenge to transparency and privacy?", *Government Information Quarterly*, Vol. 32/4, pp. 363-368, https://doi.org/10.1016/j.giq.2015.11.007.

Lewin, A. (2019), "Shiny moonshot technology will not save healthcare — yet". *Sifted*, 10 June, https://sifted.eu/articles/health-tech-startups-europe-doctolib-kry-accurx.

Lighthill, J. (1973), *Artificial Intelligence: A General Survey*, www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm.

Manning, C.D., P. Raghavan and H. Schütze (2008), *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, https://nlp.stanford.edu/IR-book/information-retrieval-book.html.

Marcus, G. (2018), "In defense of skepticism about deep learning". *Medium*, 14 January, https://medium.com/@GaryMarcus/in-defense-of-skepticism-about-deep-learning-6e8bfd5ae0f1.

Margetts, H. and C. Dorobantu (2019), "Rethink government with AI", *Nature*, 9 April, www.nature.com/articles/d41586-019-01099-5.

Marr, B. (2018), "How much data do we create every day? The mind-blowing stats everyone should read", *Forbes*, 21 May, www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#729e7ad460ba.

Mateos-Garcia, J. (2018), "The complex economics of artificial intelligence". *Nesta* (blog), 13 December, www.nesta.org.uk/blog/complex-economics-artificial-intelligence.

Mateos-Garcia, J. (2017), "Algorithmic fallibility and economic organisation". *Nesta* (blog), 10 May, www.nesta.org.uk/blog/to-err-is-algorithm-algorithmic-fallibility-and-economic-organisation.

Mazzucato, M. (2011), *The Entrepreneurial State: Debunking Public Vs. Private Sector Myths*. Anthem Press, London.

McKinsey & Company (2017), *Digitally-enabled Automation and Artificial Intelligence: Shaping the Future of Work in Europe's Digital Front-Runners*, McKinsey & Company, www.mckinsey.com/~/media/mckinsey/featured%20insights/europe/shaping%20the%20future%20of%20work%20in%20europes%20nine%20digital%20front%20runner%20countries/shaping-the-future-of-work-in-europes-digital-front-runners.ashx.

Meyer, C. (2019), "How self-driving cars could make or break a green future of transportation". *Forbes*, 12 June, www.forbes.com/sites/christophmeyereurope/2019/06/12/a-green-future-of-transportation-how-self-driving-cars-will-be-make-or-break/#3c8961522337.

MGI (McKinsey Global Institute) (2018), *Notes from the AI Frontier: Applying AI for Social Good*. McKinsey & Company. www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Applying%20artificial%20intelligence%20for%20social%20good/MGI-Applying-AI-for-social-good-Discussion-paper-Dec-2018.ashx.

Miaihle, N. and C. Hodes. (2017), "Making the AI revolution work for everyone", The Future Society at the Harvard Kennedy School of Government, Cambridge, MA, http://ai-initiative.org/wp-content/uploads/2017/08/Making-the-AI-Revolution-work-for-everyone.-Report-to-OECD.-MARCH-2017.pdf.

Mikhaylov, S., M. Esteve A. Campion (2018), "AI for the public sector: Opportunities and challenges of cross-sector collaboration", *Philosophical Transactions of the Royal Society* A, 376: 20170357.

Moneycontrol News (2019), "Gartner debunks five Artificial Intelligence misconceptions". *Moneycontrol*, 15 February, www.moneycontrol.com/news/business/companies/gartner-debunks-five-artificial-intelligence-misconceptions-3545891.html.

Mulgan, G. (2019), "Intelligence as an outcome not an input: How can pioneers ensure AI leads to more intelligent outcomes?" *Nesta* (blog), 11 June, www.nesta.org.uk/blog/intelligence-outcome-not-input/?utm_source=Nesta+Weekly+Newsletter&utm_campaign=848de748ed-EMAIL_CAMPAIGN_2019_06_07_10_07&utm_medium=email&utm_term=0_d17364114d-848de748ed-182049541.

Nelson, R. (2019), "AI beats radiologists for accuracy in lung cancer screening", *Medscape*, 23 May, www.medscape.com/viewarticle/913428.

OECD (forthcoming), "State of the art in the use of emerging technologies in the public sector", Working Paper.

OECD (2019), *Artificial Intelligence in Society*, OECD Publishing, Paris, https://doi.org/10.1787/eedfee77-en.

OECD (2018a), *Science, Technology and Innovation Outlook 2018*, OECD Publishing, Paris, www.oecd.org/sti/oecd-science-technology-and-innovation-outlook-25186167.htm.

OECD (2018b), *IoT Measurement and Applications*, DSTI/CDEP/CISP/MADE(2017)1/FINAL, www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/CDEP/CISP/MADE(2017)1/FINAL&docLanguage=En.

OECD (2017), *Fostering Innovation in the Public Sector*, OECD Publishing, Paris, www.oecd.org/gov/fostering-innovation-in-the-public-sector-9789264270879-en.htm.

OECD (2015a), *Data-Driven Innovation: Big Data for Growth and Well-Being*, OECD Publishing, Paris.

OECD (2015b), *The Innovation Imperative in the Public Sector: Setting an Agenda for Action*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264236561-en.

Partnership for Public Service/IBM Center for the Business of Government (2019), *More than Meets AI: Assessing the Impact of Artificial Intelligence on the Work of Government*, Washington, DC, www.businessofgovernment.org/sites/default/files/More%20Than%20Meets%20AI.pdf.

Partnership for Public Service/IBM Center for the Business of Government (2018), *The Future Has Begun*, Washington, DC, www.businessofgovernment.org/sites/default/files/Using%20Artificial%20Intelligence%20to%20Transform%20Government.pdf.

Pencheva, I., M. Esteve and S.J. Mikhaylov (2018), "Big Data and AI – A transformational shift for government: So, what next for research?", *Public Policy and Administration*, https://doi.org/10.1177/0952076718780537.

Raja, A. (2018), "How will GDPR affect AI?" *Medium*, 30 October, https://medium.com/datadriveninvestor/how-will-gdpr-affect-ai-3f10ed25e4c4.

Russell, S. and P. Norvig (2016), *Artificial Intelligence: A Modern Approach*, 3rd Edition, Pearson Education, London, http://aima.cs.berkeley.edu.

Shafique, A. (2018), "Forget jobs. Will robots destroy our public services?" *RSA*, 12 September, www.thersa.org/discover/publications-and-articles/rsa-blogs/2018/09/forget-jobs.-will-robots-destroy-our-public-services.

van Ooijen, C., B. Ubaldi and B. Welby (2019), *A Data-Driven Public sector: Enabling the Strategic Use of Data for Productive, Inclusive and Trustworthy Governance*, OECD Working Paper, Paris, OECD Publishing, https://doi.org/10.1787/09ab162c-en.

Viechnicki, P. and W.D. Eggers (2017), *How much time and money can AI save government? Cognitive technologies could free up hundreds of millions of public sector worker hours.* Deloitte University Press, www2.deloitte.com/content/dam/insights/us/articles/3834_How-much-time-and-money-can-AI-save-government/DUP_How-much-time-and-money-can-AI-save-government.pdf.

Vincent, J. (2019), Forty percent of 'AI startups' in Europe don't actually use AI, claims report, *The Verge*, 5 March, www.theverge.com/2019/3/5/18251326/ai-startups-europe-fake-40-percent-mmc-report.

Whittaker, M., K. Crawford, R. Dobbe, G. Fried, E. Kaziunas, V. Mathur, S. Myers West, R. Richardson, J. Schultz and O. Schwartz (2018), *AI Now Report 2018*, AI Now, New York University, New York, https://ainowinstitute.org/AI_Now_2018_Report.pdf.

Wingfield, T., L. Kostopoulos, C. Hodes and N. Miailh (2016), "Artificial Intelligence and the Law of Armed Conflict: Parameters for Discussion", The Future Society at the Harvard Kennedy School of Government, Cambridge, MA, http://ai-initiative.org/wp-content/uploads/2016/08/AI_MSC.-FINAL.pdf.

Wright, T. (2018), "Canada's use of artificial intelligence in immigration could lead to break of human rights: study", *Global News*, 26 September, https://globalnews.ca/news/4487724/canada-artificial-intelligence-human-rights.