EXAMPLES OF AI NATIONAL POLICIES

Report for the G20 Digital Economy Task Force

SAUDI ARABIA, 2020

OECD

BETTER POLICIES FOR BETTER LIVES

This document was prepared by the Organisation for Economic Co-operation and Development (OECD) Directorate for Science, Technology and Innovation, as an input for the discussions in the G20 Digital Economy Task Force in 2020, under the auspices of the G20 Saudi Arabia Presidency in 2020. The opinions expressed and arguments employed herein do not necessarily represent the official views of the member countries of the OECD or the G20.

Table of contents

Executive summary

The international policy debate on artificial intelligence (AI) has gained significant momentum in recent years and in 2019 the G20's global leadership and advocacy brought the opportunities and challenges of AI to the centre of international political discussions. Following stewardship from the Digital Economy Task Force (DETF) and with the aim of fostering public trust and confidence in AI technologies and realising their potential, G20 Digital Ministers in Tsukuba committed to a human-centred approach to AI guided by non-binding G20 AI Principles drawn from the OECD Recommendation on AI. The G20 AI Principles were welcomed by G20 Leaders in Osaka, who noted that the responsible development and use of AI can be a driving force to help advance the Sustainable Development Goals (SDGs) and to realise a sustainable and inclusive society.

There is currently a critical window for G20 members to continue their leadership on AI policy issues and to promote implementation of the G20 AI Principles. Development, diffusion and use of AI technologies are still at a relatively early level of maturity across many countries and firms, and policy-making on AI is in an active experimental phase. By working to operationalise the G20 AI Principles, the G20 can seize this chance to steer AI towards human-centred outcomes that maximise and widely share the benefits from this promising technology.

Under the 2020 Saudi Presidency, the DETF has taken the lead in advancing the G20 AI Principles. This background report, prepared by the OECD, underpinned the development of the *Examples of National Policies to Advance the G20 AI Principles*. It sets out rationales for action on each of the G20 AI Principles and details relevant examples of national strategies and innovative policy practices for AI governance. The compilation drew on country survey responses or information for almost all G20 and guest countries, and on DETF discussions that took place in 2020 under the thematic dialogue on AI. This work by the DETF marks another important milestone for the G20's leadership on AI policy issues.

Key observations from this report include:

- There is considerable activity and experimentation taking place in G20 countries to build and support trustworthy AI ecosystems, and most strategies and policies are very recent.
  - There is significant scope for sharing experiences amongst peers to facilitate learning, and for ex-ante planning for evaluation and review of policies to maximise their relevance and impact in implementing the G20 AI Principles.
- Many strategies and policies address, either explicitly or implicitly, multiple G20 AI Principles at once, reinforcing the strong complementarity of the Principles.
  - Building trustworthy and human-centric AI calls for a mix of policies addressing the full suite of the G20 AI Principles, and including issues of infrastructure, data access, the AI ecosystem and human capacity.
- Currently few policies seem to place a primary focus on the Principles of robustness, security and safety, and accountability, compared to those of inclusive growth or human-centred values.
  - There may be an opportunity to bring more emphasis to these issues as policy approaches mature and experiences grow.

- Many policies and strategic approaches to achieving trustworthy AI leverage policy tools around R&D, fostering a digital ecosystem, shaping an enabling environment, building human capacity and supporting international cooperation for trustworthy AI.
  - The policy recommendations noted by the G20 at the time of welcoming the G20 AI Principles are highly relevant to the achievement of trustworthy AI.
- A significant number of policies are oriented around R&D for AI, highlighting that countries consider that much more progress can be achieved with the technology itself, and its application to various economic and social questions.
  - There is potential for steering public research towards socially oriented applications and issues, and for leveraging R&D activities to make progress on issues such as accountability, explainability, fairness and transparency.

1 Advancing the G20 AI Principles – rationales and illustrative actions

G20 Leaders welcomed the G20 AI Principles in 2019 against a backdrop of burgeoning new and emerging applications of AI systems. In transport, for instance, autonomous vehicles are an active area of research and experimentation, with potential cost, safety and environmental benefits. AI in science has the proven potential to accelerate discovery, facilitate reproducibility, and lower experimentation costs. A recent example of this was the discovery by researchers at MIT of a new antibiotic that kills drug-resistant bacteria. The researchers used AI to screen (within the space of mere hours) a large digital library of existing pharmaceutical compounds.[1] AI in the financial services sector can detect fraud, assess creditworthiness and automate trading. Criminal justice, security, marketing and advertising, and many other activities now use AI systems.

At the same time, AI policy-making in G20 countries is in an experimental and innovative phase. Countries are actively seeking to formulate national strategies and policies that harness the promise of AI while mitigating its challenges and supporting public trust and confidence in the technology. In the first few months of 2020 alone, several major policy initiatives were put to public consultation by G20 economies, including the United States' draft memorandum on guidance for regulation of AI applications, and the European Commission's White Paper on AI (further detail on both policies is provided below).

Both these factors argue for strong and coherent efforts to advance implementation of the G20 AI Principles. Given the breadth of AI applications – and the associated potential impacts on G20 economies and societies if these are not trustworthy – it is critical to progress efforts towards human-centred, transparent, robust and accountable AI systems, and to build a policy environment that facilitates that progress. The time is propitious, as the policy experimentation currently underway offers great scope for innovative approaches and peer learning from national approaches across the G20, and the potential to shape global directions for AI development. Given the centrality of data to AI systems, this effort also dovetails well with the G20's interest in issues of data access and sharing and data flows.

To capitalise on this moment and support the advancement of the G20 AI Principles, the DETF under the 2020 Saudi Presidency compiled *Examples of National Policies to Advance the G20 AI Principles* that provides countries with examples of national strategies and policies as they implement the G20 AI Principles in their particular country contexts.

To support this, this chapter provides examples of national strategies and innovative policy practices that intend to steer towards responsible stewardship of trustworthy AI. For each G20 AI Principle (the five values-based principles and the five recommendations for policy), it also provides additional explanation of the Principle and gives deeper elaboration of the rationale for implementing the Principle.

[1] https://www.theguardian.com/society/2020/feb/20/antibiotic-that-kills-drug-resistant-bacteria-discovered-through-ai?utm_term=RWRpdG9yaWFsX0d1YXJkaWFuVG9kYXlVS19XZWVrZGF5cy0yMDAyMjE%3D&utm_source=esp&utm_medium=Email&CMP=GTUK_email&utm_campaign=GuardianTodayUK

Observations from these existing policy approaches and a stylised (non-exhaustive) mapping of policies against the G20 AI Principles is contained in chapter 2. The examples of national strategies and policies are based on input provided by G20 DETF participants following a request from the Saudi Presidency, regarding their national AI strategies as well as policy settings and experimentation. Countries were requested to highlight selected examples of policies that aim at, or have the effect of, implementing the G20 AI Principles (Box 1 details the core survey questions). Some examples from the health sector, drawn from discussions under the 2020 Dialogue on AI, are highlighted in Box 2 later in the chapter.

Box 1. G20 DETF input request on trustworthy AI

Question 1

In 2019, the G20 supported the G20 AI Principles for responsible stewardship of trustworthy AI, encompassing inclusive growth, sustainable development and well-being; human-centred values and fairness; transparency and explainability; robustness, security and safety; and accountability.

Please provide up to two examples of national policy actions (e.g. strategies, policies, guidance, regulations, legislation) that aim at, or have the effect of, supporting the implementation of one or more of the G20 Principles in your country.

Question 2

In 2019, the G20 took note of recommendations for national policies and international cooperation for trustworthy AI. These addressed:

- Investing in AI research and development
- Fostering a digital ecosystem for AI
- Shaping an enabling policy environment for AI
- Building human capacity and preparing for labour market transformation
- International cooperation for trustworthy AI

Choosing one or more of these categories, please provide up to two examples of national policy actions/orientations that may support the development of trustworthy AI at the national or international level.

Argentina, Australia, Brazil, Canada, China, France, Germany, Indonesia, Italy, Japan, Korea, Mexico, the Russian Federation, Saudi Arabia, Turkey, the United Kingdom, the United States, and the European Commission submitted survey responses. G20 guest countries and regional representatives also provided responses, including Singapore, Spain, Switzerland, and the United Arab Emirates. Policy examples have been grouped under the G20 AI Principles according to the indications of the responding countries or, where not specified, according to key features of the examples. Actions shown from the European Union and India draw on the OECD.AI Policy Observatory; no information was available for South Africa and Jordan.

1. Inclusive growth, sustainable development and well-being

Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.

Explanation and rationale:

This principle recognises that guiding the development and use of AI toward prosperity and beneficial outcomes for people and planet is a priority. Trustworthy AI can play an important role in advancing inclusive growth, sustainable development and well-being and global development objectives. It can be leveraged for social good and can substantially contribute to achieving the Sustainable Development Goals (SDGs) in areas such as education, health, transport, agriculture, environment, and sustainable cities, among others.

This stewardship role should strive to address concerns about inequality and the risk that disparities in technology access increase existing divides within and between developed and developing countries. This principle also recognises that AI systems could perpetuate existing biases and have a disparate impact on vulnerable and underrepresented populations, such as ethnic minorities, women, children, the elderly and the less educated or low skilled. Disparate impact is a particular risk in low- and middle-income countries. This principle emphasises that AI can also, and should, be used to empower all members of society and to help reduce biases.

Responsible stewardship is furthermore a recognition that throughout the AI system lifecycle, AI actors and stakeholders can, and should, encourage the development and deployment of AI for beneficial outcomes with appropriate safeguards. Defining these beneficial outcomes, and how best to achieve them will benefit from multidisciplinary and multi-stakeholder collaboration and social dialogue. A meaningful, well-informed, and iterative public dialogue that is inclusive of all stakeholders can enhance public trust in and understanding of AI.

Along these lines, the DETF engaged in a Dialogue on AI in 2020, to champion a discussion of AI applications that promote the use of trustworthy AI. This focuses on the way in which AI applications at the sector level (including education and health) uphold responsible stewardship of trustworthy AI and the values-based G20 AI Principles, the challenges that arise as these sectors make increasing use of AI, and the role of governments consistent with the Principles to address these challenges.

Considerations for governments:

Failing to work towards advancing this G20 AI Principle on inclusive growth, sustainable development and well-being may not only represent a missed opportunity to harness this technology for positive economic and societal outcomes, but may also lead to widening divides between countries and groups, stoking tensions and deepening inequalities that serve to hold back global progress as a whole. The current global health crisis around COVID-19 is one illustration of the need to harness technologies such as AI towards globally beneficial outcomes.

Examples of initiatives:

Argentina – National Plan for AI

Argentina's National Plan for AI ("Plan Nacional de Inteligencia Artificial"), created in 2019 and pending implementation, tries to tackle the challenge of designing and promoting AI in a way that benefits society, and supports inclusive growth and sustainable development and well-being, while still guaranteeing fair and inclusive societies and mitigating ethical risks. Its proposed scope and actions demonstrate clear synergies between the values-based G20 AI Principles and the recommendations for national policies also encompassed in the Principles.

The aim of Argentina's National Plan on AI is to guide public policies, initiatives and practices related to AI through to 2030, so as to help reach national development goals linked to the SDGs and to position Argentina as a regional leader in this technology. The National Plan complements two other priority initiatives that also envisage a national strategy for development and adoption of AI in Argentina: the Digital Agenda of Argentina 2030, and the National Strategy of Science, Technology and Innovation – Argentina Innovates 2030. Taking a people-centred approach, the AI plan sets out specific objectives on talent, data and interoperability, infrastructure (notably "hypercompute" capacities [super or quantum computing]), R&D for AI and adoption of AI as a transversal tool, use of AI in the public sector, jobs and skills, ethics and regulation, international cooperation, and an Innovation Lab. It encompasses multiple policy instruments, including creation of bodies (such as an Ethics AI Committee), creation of standards (e.g. for design and use of databases, and ethical use of AI), and establishment of agreements with providers of cloud infrastructure and "hypercompute" to support development and implementation of AI in 10 institutions across Argentina.

Argentina's National Plan on AI aims to benefit all stakeholders, including a special emphasis on under-represented populations. Similarly, the plan benefits from multistakeholder governance, with the current architecture drawing in the Ministry of Education, Ministry of Science, Technology and Innovation, Ministry of Production, Ministry of Employment, Ministry of Foreign Affairs and the Chief of Cabinet's Office, as well as 20 technical teams, a Multi-sectorial Committee of Artificial Intelligence and a Scientific Committee of experts. The Innovation Lab is also involved in the design and development of the National Plan. The Ministry of Science, Technology and Innovation has requested the inclusion of a specific allotted budget in the 2020 National Budget Law to operationalise the National Plan.

Brazil – National AI Strategy

Brazil's National AI Strategy ("Estratégia Brasileira de Inteligência Artificial"), currently out for public consultation, aims to develop a comprehensive whole-of-society approach to AI. Its objective is to extract maximum benefit from the use of AI for scientific development, competitiveness and productivity (including in public services) and well-being. Its breadth of policy interests and potential actions reinforce the relevance of the full suíte of G20 AI Principles.

Brazil's National AI Strategy[2], under the responsibility of the Ministry of Science, Technology, Innovation and Communications, aims to stimulate the development and utilisation of AI technology to promote scientific development and to solve real problems in the country. With regard to the adoption of AI

[2] http://participa.br/estrategia-brasileira-de-inteligencia-artificial/blog/apresentacao-e-instrucoes

technologies, potential areas include health, urban mobility, public safety, and government services, given the need to improve efficiency and reduce costs. Similar to other national AI strategies, issues include the potential impact on jobs, the need to develop talent and skills in the workforce, and the importance of promoting research, development and innovation.

The Strategy proposes developing guidelines and actions along six pillars: education and capacity building in AI; AI research, development, innovation and entrepreneurship; AI applications in the private sector, government and public safety; legislation, regulation and ethical use of AI; governance of AI; and international aspects of AI. Brazil suggests concrete policies can enable the development of the AI ecosystem, including opening government data, establishing regulatory sandboxes, fostering startups in this field, as well as directing R&D investment funds to this area. Additionally, it is essential that nations cooperate in relevant international organisations, in order to achieve a common understanding and develop principles of ethics and responsibility in the use of AI.

China – Governance Principles for the New Generation AI

China's Governance Principles for the New Generation AI – Developing Responsible AI, have the objectives of promoting healthy development of AI, strengthening R&D in legal, ethics and social aspects, and actively participate in global governance of AI. The principles span the five values-based G20 AI Principles.

Released in June 2019 and falling under the responsibility of the Ministry of Science and Technology, China's Governance Principles for the New Generation AI aim to benefit all stakeholders in the AI industry. They are designed to better balance between development and governance of AI, ensure its safety, reliability and controllability, and support sustainable development economically, socially, and environmentally for a community with a shared future. The initiative highlights the theme of developing responsible artificial intelligence with 8 principles – harmony and human-friendly (noting the goal of AI development should be to promote the well-being of humankind), fairness and justice, inclusion and sharing, respect for privacy, safety and controllability, shared responsibility, open and collaboration, and agile governance.

European Union – Ethics Guidelines on AI

The European Union's Ethics Guidelines on AI, developed by its High-Level Expert Group on AI, set out seven key requirements that AI systems should meet in order to be deemed trustworthy. Including concepts such as human agency and oversight, and transparency, the Guidelines not only advance the G20 AI Principle on inclusive growth but also support the other core values-based principles.

In April 2019 the EU published its Ethics Guidelines on AI[3], following a process of development by the 52-member High-Level Expert Group on AI (HLEG) and an extensive public consultation. The Guidelines state that trustworthy AI should be lawful (respecting all applicable laws and regulations), ethical (respecting ethical principles and values), and robust (both from a technical perspective while taking into account the social environment). They set out seven key requirements that AI systems should meet:

- Human agency and oversight
- Technical robustness and safety

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

- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental well-being
- Accountability

While these requirements are intended to apply to all AI systems in different settings and industries, the specific context in which they are applied should be taken into account for their concrete and proportionate implementation, taking an impact-based approach. In this light, the Guidelines provide an assessment list that operationalises the key requirements and offers guidance to implement them in practice. This assessment list was piloted from June to December 2019, with stakeholders invited to test and provide feedback, and a revised document will be issued in 2020. The Guidance will serve as input for a possible legal framework on AI.

European Union – White Paper on AI: A European approach to excellence and trust

The European Commission is consulting on policy options to achieve the twin objectives of promoting the uptake of AI and of addressing the risks associated with certain uses of this technology. Focused on enabling a trustworthy and secure development of AI in Europe, the White Paper has a strong orientation to inclusive growth, sustainable development and well-being, as well as contributing to a strong AI eco-system.

On 19 February 2020, the European Commission published a White Paper [4]aiming to foster a European ecosystem of excellence and trust in AI. The consultation period runs to 14 June 2020. The White Paper posits that a common European approach to AI is necessary to reach sufficient scale and avoid the fragmentation of the single market. It also proposes that the introduction of national initiatives could endanger legal certainty, weaken citizens' trust and prevent the emergence of a dynamic European industry. The White Paper presents policy options to enable a trustworthy and secure development of AI in Europe, in full respect of the values and rights of EU citizens, and builds further on Europe's broad policy mix[5]. The main building blocks are:

- A policy framework setting out measures to align efforts at European, national and regional level. In partnership between the private and the public sector, the aim of the framework is to mobilise resources to achieve an 'ecosystem of excellence' along the entire value chain, starting in research and innovation, and to create the right incentives to accelerate the adoption of solutions based on AI, including by small and medium-sized enterprises (SMEs).
- The key elements of a future regulatory framework for AI in Europe that will create a unique 'ecosystem of trust'. This would ensure compliance with EU rules, including the rules protecting fundamental rights and consumers' rights, in particular for AI systems operated in the EU that pose a high risk. Building an ecosystem of trust is a policy objective in itself, giving citizens the confidence to take up AI applications and give companies and public organisations the legal certainty to innovate using AI. The effort is strongly supportive of a human-centric approach based on the Communication on Building Trust in Human-Centric AI and will also take into account the

[4] https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

[5] https://ec.europa.eu/digital-single-market/en/artificial-intelligence

input obtained during the piloting phase of the Ethics Guidelines prepared by the High-Level Expert Group on AI.

The White Paper is accompanied by a European Strategy for Data[6], which aims to enable Europe to become the most attractive, secure and dynamic data-agile economy in the world – empowering Europe with data to improve decisions and better the lives of all its citizens. The strategy sets out a number of policy measures, including mobilising private and public investments, needed to achieve this goal.

France – French AI Strategy: Economic Action Plan

Building from the French AI Strategy announced a year earlier, the Economic Action Plan brings a focus to funding, talent and access to data to help develop a data-driven French and European economy. As well as contributing to implement the AI Principle of inclusive growth, sustainable development and well-being, this Plan also addresses transparency and explainability.

Presented by the Minister of Economy and Finance Bruno Le Maire in July 2019, the Economic Action Plan of the French AI Strategy[7] identifies three top priorities: building and structuring a strong French AI system; making AI accessible to all companies; and developing a data-driven French and European economy. To achieve this, the Plan uses the concept of "AI challenges" to give access to public funding to AI providers and users who propose solutions to concrete problems in four key sectors (health, transportation/mobility, environment, and defence/security). Already more than 550 French start-ups are active in the field of AI. In addition, high priority will be given to training of data scientists, engineers, developers and AI researchers, as demand for such skills will become increasingly critical. Relevant also to the G20 AI Principle on fostering a digital ecosystem for AI, a call for projects was launched to co-fund data sharing initiatives for the development of new AI solutions. By leveraging elements of project funding and education/training policy, the Plan is aimed at private companies.

One achievement already recorded is the signature by eight top industrial corporations of a joint Manifesto on Artificial Intelligence for Industry (July 2019), in which they put forward a common strategic vision of AI. The signatories have agreed to conduct a joint review, to share findings with policymakers, to establish a coordinated plan of action with the French AI ecosystem, and to encourage the participation of all public and private stakeholders who share this common strategic vision of AI.

[6] https://ec.europa.eu/info/files/communication-european-strategy-data_en

[7] See https://www.aiforhumanity.fr/en/ and https://www.entreprises.gouv.fr/numerique/intelligence-artificielle-au-service-des-entreprises

Germany – AI Strategy

Germany's Federal AI Strategy sets out a framework for a holistic policy on future development and application of AI in Germany. Benefitting from a nationwide public consultation, the Strategy places a focus on the benefits for people and the environment, and the intensive dialogue underway with all sections of society about AI is to be strengthened. The Strategy aims to address all five values-based G20 AI Principles.

Germany's AI Strategy[8] was adopted in November 2018 and was jointly developed by the Federal Ministry for Economic Affairs and Energy, the Federal Ministry of Education and Research and the Federal Ministry of Labour and Social Affairs. It aims to safeguard Germany's outstanding position as a research centre, to build up the competitiveness of German industry, and to promote the many ways to use AI in all parts of society. It seeks to build on existing strengths and transfer them to areas where no or little use has been made of AI's potential. The Strategy has three core objectives:

- Making Germany and Europe a leading centre for AI and thus help safeguard Germany's competitiveness in the future
- Safeguarding the responsible development and use of AI that serves the good of society
- Integrating AI in society in ethical, legal, cultural and institutional terms in the context of a broad societal dialogue and active political measures.

Under each objective is nested a number of specific goals, with the focus on twelve fields of actions starting from fostering research and development to support for SMEs, better conditions for start-ups, managing structural changes in the labour market and deeper international cooperation. In November 2019, the Federal Government published an interim report, describing the implementation of the AI Strategy in the individual fields of action so far. Funding from the federal budget totalled EUR 1 billion for 2019 and 2020, and up to 2025 the Federation intends to provide around EUR 3 billion for the implementation of the Strategy. Leveraging business, science and the Länder will double the overall amount available.

India – National Strategy on AI – "#AIForAll"

India's AI Strategy is focused on leveraging AI for inclusive growth in line with the government policy of "Sabka Saath Sabka Vikas". In achieving this, the role of the government is aimed towards developing the research ecosystem, promoting adoption and addressing skill challenges. The Strategy and its implementation align strongly with many aspects of the G20 AI Principles.

In 2018, India's Finance Minister mandated NITI Aayog to establish a National Programme on AI[9], with a view to guiding R&D in new and emerging technologies. NITI Aayog has advanced on multiple fronts, with one prong being to craft a national strategy for building a vibrant AI ecosystem in India. The Strategy flags important issues including bias, ethics and privacy, and envisions the government undertaking R&D in technology to address these concerns. The focus of the Strategy is particularly on agriculture, health and education, where public investment and leadership is needed. A discussion paper[10] issued in June 2018

[8] https://www.ki-strategie-deutschland.de/ (in German only, but English, French and Japanese PDF version of the Strategy are available under 'downloads').

[9] https://niti.gov.in/national-strategy-artificial-intelligence

[10] https://niti.gov.in/sites/default/files/2019-01/NationalStrategy-for-AI-Discussion-Paper.pdf

recommended multiple actions for government including *inter alia* setting up centres of excellence, establishing an attractive intellectual property regime for AI, introducing AI and machine learning in schools, and collaborating on privacy-preserving technology research in AI.

Indonesia – National Strategy for Artificial Intelligence

Indonesia is currently finalising a National Strategy for AI as a roadmap for AI development. Supporting a focus on AI talent development, ethics in AI, AI infrastructure and data, and AI research and industrial innovation, and seeking to align with Indonesia's national ideology (including principles of social justice and humanity), the Strategy touches on multiple G20 AI Principles.

With the development of its National Strategy on AI, Indonesia aims for the development of AI to be in line with its national ideology (Pancasila) and interests, which include principles of social justice, citizen protection, humanity and the pursuit of democratic approaches. It takes the position that the state must provide a data ecosystem and infrastructure that support AI's contribution to national development, and that the ecosystem should provide a secure, safe and accountable system for data protection.

The development of human resources is an integral factor for implementing a successful AI ecosystem, and through the Strategy Indonesia is setting a mission to build a learning and innovation ecosystem involving a quadruple helix co-operation between government, academics, industry and community. This aims to strengthen education and improve the quality and quantity of talent on AI in Indonesia, producing people that can work as technicians, researchers and business practitioners to counter the current shortage of skilled workers in the AI field. It also aims to create talent with strong soft skills.

Indonesia is also establishing a research ecosystem and accelerating the growth of AI industry innovations that have a wide-ranging impact on society. AI development will focus on realising the national vision in five priority sectors, namely healthcare services, bureaucratic reformation, research and education, food security, and smart cities and mobility. Finally, Indonesia also welcomes international stakeholders to collaborate and exchange ideas on how to collectively develop a trustworthy AI ecosystem.

Italy – development of an AI Strategy

Italy is currently developing a holistic and multi-perspective approach to AI, through the development of an AI Strategy. As well as seeking to evaluate the socio-economic impact of the development and widespread adoption of AI systems, it will also include a focus on safety and responsibility, in alignment with several of the G20 AI Principles.

In September 2018, the Italian Ministry of Economic Development formed a 30-member group of experts to draft a national strategy on AI.[11] The group, chaired by the Minister of Economic Development, comprised ten representatives of enterprises operating in the field of AI, ten representatives of research centres / think tanks or academia, and ten representatives of the labour market, professions, consumers and civil society. It was tasked to provide guidance on issues including:

- improving, coordinating and strengthening the research in the AI field;
- promoting public and private investments in AI, also benefitting from the dedicated EU funds;

[11] https://www.mise.gov.it/index.php/en/news/2038605-artificial-intelligence-ai-call-for-experts

- attracting talent and developing business in the field of AI;
- encouraging the development of the data-economy, paying particular attention to the spreading and valorisation of non-personal data, adopting the better standards of interoperability and cybersecurity;
- the legal framework with specific regard to safety and responsibility related to AI-based products and services;
- the socio-economic impact of development and widespread adoption of AI-based systems, along with proposals for tools to mitigate the encountered issues.

The draft Strategy has undergone a period of public consultation and is now being finalised.

Japan – Human-centric Principles and AI Strategy

Japan's overarching "Society 5.0" vision has been bolstered by the introduction of Principles of Human-centric AI Society and an AI Strategy. These policy actions, both introduced in 2019, have a focus on the social side of AI and are an important channel for Japan to implement the G20 AI Principles, especially on inclusive growth, sustainable development and well-being, and human-centred values and fairness.

In March 2019, the Japanese government introduced Principles of Human-centric AI Society, formulated to mitigate people's concerns about AI and promote active social implementation of AI. Shortly after, in June 2019, it introduced Japan's AI Strategy, which focused on measures that the national government should take in the immediate term. Under the Strategy, it established specific goals in the fields of "building the foundation for the future", "building the foundation of industry and society", and "ethics" (including education reform, research and development, social implementation, data, digital government, support for SMEs and start-ups, and social principles). These complementary policies both support Japan's Society 5.0 goal to balance economic development with resolution of social issues through deep integration of emerging technologies. The Principles aim to increase social adoption of AI, while the Strategy aims to steer AI and the structure of society toward digital transformation.

Korea – National Strategy for AI

Korea has just released its National Strategy for AI, laying out its vision for the era of AI and setting measurable goals for its performance to 2030. The product of inter-Ministerial collaboration, the Strategy aims to address all aspects of the G20 AI Principles.

Korea's National Strategy for AI[12] sets out its vision to transition from an IT superpower to an AI superpower. It includes objectives aimed at both economic and social advancement, namely: to be ranked third in the IMD World Digital Competitiveness Ranking by 2030 (currently 10th); additional economic output of KRW 455 trillion delivered by AI; and to be ranked among Top 10 nations in terms of the quality of life (currently among Top 30, OECD). The Ministry of Science and ICT holds responsibility for the

[12] The English version of the National Strategy for Artificial Intelligence can be found on https://www.msit.go.kr/english/msipContents/contents.do?mld=NDU4

strategy, which incorporates direct financial support, collaborative infrastructures, guidance and incentives. The strategy entails annual budget expenditure of KRW 1.4 trillion.

The National Strategy for AI lays out nine strategies and 100 tasks in three pillars: 1) fostering a world-leading AI ecosystem, 2) becoming a nation unrivalled for its use of AI, and 3) realising human-centred AI.

- The first pillar seeks to expand high-performance computing resources and make pre-emptive investments in next-generation AI technology that can surpass the limits of existing AI technology. In particular, a comprehensive plan will be developed for the semiconductor industry to boost AI chip-making by investing in AI processors and new semiconductor devices.
- The second pillar outlines strategies to nurture world-class AI talent and provide AI education for all citizens, apply AI across all industries, and become the world's best e-government. It includes actions to: expand AI graduate programs; implement class-specific, job-specific and age-specific software and AI education for all citizens; apply AI across diverse industries (including manufacturing, healthcare, energy, agriculture and fisheries, arts and culture, national defense, transportation and environment), as well as national infrastructure; and adopt public cloud and use AI to innovate public services, such as welfare, public safety and climate change responses.
- For the third pillar the Korean government will build an inclusive employment safety net so that citizens can enjoy the benefits of AI amid rapid technological and social changes. The Korean government will strengthen the employment safety net, such as social insurance, and aim to ensure lifelong employment for people by providing job training and reskilling in emerging technology sectors. Another priority is innovating cybersecurity in response to potential cyber threats caused by proliferation of AI, by analysing vulnerabilities in devices and networks while verifying password safety. There are also actions via R&D and inter-ministerial cooperation to counter adverse effects of AI, such as deepfakes. The Korean government also aims to set AI ethics guidelines and action plans that are in line with the discussions on AI ethics at international fora such as the G20 and OECD.

In addition to the National Strategy for Artificial Intelligence, the ministries of the Korean government each developed sector-specific action plans and reflected them in the 2020 Action Plan. The Korean government plans to implement the main tasks of the Strategy and keep track of their progress.

Singapore – Model AI Governance Framework

Singapore's Model AI Governance Framework is a living document that translates key ethical AI principles into implementable practices for industry. Accompanied by an Implementation and Self-Assessment Guide for Organisations, the Model Framework intends to guide organisations to deploy AI responsibly and addresses all values-based G20 AI Principles.

Singapore's approach to AI governance is outcome-driven, principle-based and multistakeholder, to ensure the approach is informed and practical for industry to adopt. Its Model AI Governance Framework[13] was first launched in January 2019, under the primary responsibility of the Infocomm Media Development Authority of Singapore (IMDA) and the Personal Data Protection Commission (PDPC). A second edition was launched by the Minister for Communications and Information in January 2020. The ethical principles embodied in the Model Framework are:

- Decisions made by or with the assistance of AI should be explainable, transparent and fair; and

[13] https://go.gov.sg/ai-gov-mf-2

- AI systems and decisions made using AI should be human-centric and safe.

In line with these principles, the Model Framework maps out the key governance considerations and practices that apply to common AI deployment processes in these four areas:

- Internal governance structures and measures;
- Determining the level of human involvement in AI-augmented decision-making;
- Operations management; and
- Stakeholder interaction and communication.

The second edition of the Model Framework includes new considerations for effective AI governance (such as robustness and reproducibility), and refinements to enhance its relevance and usability (such as expanding the section on customer relationship management to include interactions and communications with a broader network of stakeholders). In addition, it contains seven practical examples from selected organisations[14] to illustrate actual implementation of the AI governance practices described in the framework.

The Model Framework is accompanied by an Implementation and Self-Assessment Guide for Organisations (ISAGO)[15], developed in partnership with WEF and intended to guide organisations to assess the alignment of their AI governance practices with the Model Framework. It also provides industry examples and practices to better guide organisations in implementing the Model Framework. Over 60 Singapore and international organisations contributed to the ISAGO, including DataRobot, DBS Bank, Google, KPMG, Mastercard, Microsoft, Salesforce, Suade Labs and Visa. The Model AI Governance Framework, ISAGO, and Compendium of Use Cases, contain use cases and illustrations on how various organisations of different size, sector and geographical reach have either implemented or aligned their AI governance practices with the Model Framework. In particular, the Compendium shares information on how organisations have effectively put in place accountable AI governance practices and have benefited from doing so. These initiatives serve to encourage more organisations to adopt similar practices and expand the ecosystem of organisations embarking on trustworthy AI.

Together, the Model AI Framework and ISAGO form one of three inter-linked initiatives under Singapore's National AI Strategy (NAIS). This Strategy sets out Singapore's plans to deepen the use of AI technologies to transform Singapore's economy, and to position Singapore at the forefront of developing and deploying scalable AI solutions in sectors of high value and relevance to Singapore's citizens and businesses. To achieve this, the NAIS envisions: (a) Singapore as a global hub for developing, test-bedding, deploying and scaling AI solutions; (b) the use of AI by government and businesses to generate economic gains and improve lives; and (c) Singaporeans understanding AI technologies and its benefits, and Singapore's workforce being equipped with the necessary competencies to participate in the AI economy. In this regard, the creation of a progressive and trusted environment for AI – where citizens trust the responsible use of AI – has been identified as a key enabler in the NAIS.

Spain – National Strategy on AI

Spain is in the final stages of developing its National Strategy on AI, with a suite of objectives from promotion of research to the prevention of discrimination and respect for human rights. This follows a recommendation

[14] The organisations are Ayasdi, CUJO AI, Grab, Mastercard, MSD International GmBH (Singapore Branch), pymetrics and Suade Labs.

[15] https://go.gov.sg/ai-gov-mf-2

of the Spanish Strategy on Research, Development and Innovation, which also proposed establishing an AI Observatory. With a call to use AI to achieve the Agenda 2030 and a strong focus on ethics and society, the initiatives address important aspects of the G20 AI Principles.

Spain's National Strategy on AI (Estrategia Nacional de IA) is being developed under the responsibility of the Ministry of Economic Affairs and Digital Transformation. With implementation of the Strategy scheduled for April, it is currently pending approval and budget allocation. Its objectives include promotion of basic and applied research in the field of AI, introduction of AI into production processes and public sector activity, protection of constitutional values and prevention of discrimination, ensuring reliable and quality data sources, full respect for individual rights, and protection of security. The strategy will have actions structured in the following areas: scientific research; education & training; data & digital infrastructure; economic and industrial transformation; AI in government; security; ethical standards, inclusion and social welfare.

The Strategy was one of the recommendations of the Spanish Strategy on Research, Development and Innovation[16] (Estrategia Española de I+D+i en Inteligencia Artificial) under the responsibility of the Ministry of Science, Innovation and Universities. The RDI Strategy describes six priorities and seven recommendations to guide the further development of a national strategy, including establishing strategic areas to focus R&D activities, facilitating knowledge transfer, planning training and professionalisation in the field of AI, developing a data ecosystem, and analysing AI ethics. The recommendations include a proposal to create a National Data Institute to plan and define governance over data from different levels of government.

Turkey – Eleventh Development Plan (2019-2023)

Published by the Presidency of the Republic of Turkey, the Development Plan looks to transform Turkey's economic structure and improve the welfare of its people, including through breakthroughs in technology and innovation. It includes a number of measures specific to AI and as well as its overarching objectives related to inclusive growth, the Plan also has a strong focus on increasing R&D as well as strengthening the innovation ecosystem.

This development plan[17] creates nationwide awareness and coordination on science, technology and innovation and employs a range of policy instruments. The main objective related to "Science, Technology and Innovation" is to improve the capacity to produce and use knowledge by establishing an efficiently functioning research and innovation ecosystem, and to increase R&D and innovation activities that would support high value-added products and services. In addition to policies and measures for all digital technologies, the policies and measures related to specifically AI include:

- Preparing a roadmap for national-scale studies aimed at the development of domestic technology production capabilities and roll out of these technologies across the whole economy, in the field of AI technologies.
- Establishing an educational data bank where databases related to the education system are integrated and the data will be processed by AI technologies.

[16] http://www.ciencia.gob.es/stfls/MICINN/Ciencia/Ficheros/Estrategia_Inteligencia_Artificial_IDI.pdf

[17] http://www.sbb.gov.tr/wp-content/uploads/2020/03/On_BirinciPLan_ingilizce_SonBaski.pdf

- Improving processes and technological infrastructure in order to utilise new technologies such as big data, cloud computing, mobile platforms, internet of things, AI and blockchain in the development of public services.
- Encouraging technology suppliers to develop applications and services that can be offered on the industrial cloud platform such as AI, advanced data analytics, simulation and optimisation, product lifecycle and production management systems. The use of this platform by companies will be encouraged through support for digital transformation.
- Preparing road maps for AI, internet of things, augmented reality, big data, cybersecurity, energy storage, advanced material, robotics, micro/nano/optoelectronics, biotechnology, quantum, sensor technologies and additive manufacturing technologies. These will address infrastructure, human capital and social considerations.
- Developing agricultural information systems with digitalisation, AI and data-based business models and making them widely available.

Turkey – AI Technology Roadmap

Turkey is currently preparing an AI Technology Roadmap to identify scientific themes, sub-technology areas and sectoral applications on which Turkey will focus its research, technological development and innovation for the next 5-10 years, in order to achieve the most added value from AI technologies. This multistakeholder effort particularly supports implementation of the G20 AI Principles on inclusive growth, robustness and accountability.

In January 2020, Turkey established a stakeholder working group, comprised of all the relevant stakeholders from academia, private sector and major umbrella NGOs, to assist in developing its AI Technology Roadmap. The exercise is being conducted under the Science, Technology and Innovation Policy Council of the Turkish Presidency, via the technical contribution of TUBITAK[18] (The Scientific and Technological Research Council of Turkey). Frontier scientific themes, priority sub-technology areas and priority sectoral applications of AI technologies that have critical added value in terms of technological advancement and/or competitiveness of major sectors in Turkey will be identified. The stakeholder working group will help ensure effective intergovernmental coordination and broad inputs to the process. The exercise aims to benefit the broad AI community in Turkey, and will be implemented through direct financial support, incentives and infrastructure. As well as inclusive growth, sustainable development and well-being, the AI Technology Roadmap will address the Principles of robustness, security and safety and accountability.

[18] www.tubitak.gov.tr

United Arab Emirates – AI Principles and Guidelines for the Emirate of Dubai

Dubai's AI Principles and Guidelines aim to help AI developers, government and society develop AI in a safe, responsible and ethical way. With objectives including pursuit of sustainable goals, the initiative addresses the suite of values-based G20 AI Principles.

Launched in January 2019, and part of Dubai's broader approach to ethical AI, the AI Principles and Guidelines[19] aim to:

- provide commissioners, developers and users of artificial intelligence with an understanding of how AI should be developed and monitored in a way that is fair, transparent, accountable and explainable,
- ensure that the innovation potential is optimised whilst societal risks are minimised,
- capture economic and social value generated through AI and boost its use in pursuit of sustainability goals,
- grow AI as a segment within Dubai's broader economy.

Accompanying the Principles and Guidelines is an Ethical AI Self-Assessment Tool built to enable AI developer or operator organisations to evaluate the ethics level of an AI system. It gives an assessment (from proof of concept to production) of the potential ethical issues that may arise throughout the development process and how specific AI applications could be improved to ensure fairness, transparency, accountability and explainability. The tool also aims to ensure careful adoption of AI in which innovation potential is optimised and where economic and social value is captured.

Dubai established a governing AI ethics board, comprising government policy, academic, legal and industry experts to oversee and guide the strategic development of the tool and guidance. The Executive Council of Dubai has directed government entities to use the principles and guidelines when considering AI development, and entities including the Road and Transport Authority and the Dubai Police have formally acknowledged their adoption of the self-assessment tool when developing AI.

United Kingdom – Response to independent AI review

The UK government is considering its response to an independent AI review, with the aim of developing options to build the UK's strengths in AI and support the safe and innovative adoption of AI technologies for the benefit of the economy.

The UK government is considering its response to the independent AI review developed by Dame Wendy Hall and Jerome Pesenti.[20] The objective is to implement the AI Review recommendations (numbering 18) as part of the Industrial Strategy and Sector Deal. Options include a range of measures for increasing supply of skills and talent at all levels, increasing trusted access to data, and a leadership body to support

[19] https://www.smartdubai.ae/initiatives/ai-principles-ethics

[20] https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk

this developing industry. The government is also taking forward a manifesto commitment to develop a Data and AI Ethics body to work with government and a range of institutions on governance frameworks.

In parallel, the AI UK Sector Deal (2018-2027) provides GBP 1 billion support from government and industry to boost the UK's global position as a leader in developing AI technologies. It will take actions to advance the Industrial Strategy's AI and Data Grand Challenge and ensure the UK is the leading destination for AI innovation and investment. Objectives relate to:

- Ideas - the world's most innovative economy;
- People - good jobs and greater earning power for all;
- Infrastructure - a major upgrade to the UK's infrastructure;
- Business environment - the best place to start and grow a business;
- Places - prosperous communities across the UK.

The Sector Deal builds on the Hall and Pesenti review. Ongoing action areas from previous strategies include finding ways to use data and AI for prevention, early diagnosis and treatment of chronic diseases by the year 2030, especially for cancer diagnosis; leading the world in safe and ethical use of data and AI; using automation to do extreme jobs which endanger human life; and helping people develop the skills needed for the future jobs.

United States – Plan for Federal Engagement in Developing Technical Standards

This technical standards plan provides guidance regarding important characteristics of standards to help agencies in their decision-making about AI standards. It was developed with broad public and private sector input, and by specifying trustworthiness as one area of focus for AI standards, the plan speaks to all five values-based G20 AI Principles. It also contributes to shaping an enabling policy environment for AI, reinforcing the strong complementary between the values-based and policy-oriented AI Principles.

In August 2019, the National Institute of Standards and Technology (NIST) introduced U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools.[21] The plan directs U.S. government agencies to prioritise involvement in AI standards efforts that are inclusive and accessible, open and transparent, consensus-based, globally relevant, and non-discriminatory. It emphasises the need for federal agencies to be flexible in selecting AI standards for use in regulatory or procurement actions. It also calls for prioritising multidisciplinary research and expanding public-private partnerships to advance reliable, robust and trustworthy AI. The plan also highlights related tools that will be needed to support AI, including benchmarks, evaluations and challenges that could drive creative problem solving.

The plan recognises different levels of potential agency involvement, from least to most engaged: monitoring, participating, influencing and leading. It provides practical steps for agencies to take as they decide about engaging in AI standards and identifies nine areas of focus for AI standards: 1) Concepts and terminology, 2) Data and knowledge, 3) Human interactions, 4) Metrics, 5) Networking, 6) Performance testing and reporting methodology, 7) Safety, 8) Risk management, and 9) Trustworthiness. Trustworthiness standards include guidance and requirements for accuracy, explainability, resiliency, safety, reliability, objectivity, and security.

[21] https://www.nist.gov/document/report-plan-federal-engagement-developing-technical-standards-and-related-tools

United States – Guidance for Regulation of AI Applications

This draft Memorandum provides guidance to all Federal agencies to inform development of regulatory and non-regulatory approaches regarding technologies and industrial sectors that are empowered or enabled by AI. With goals of ensuring public engagement, limiting regulatory overreach and promoting trustworthy technology, the guidance is strongly aligned to the five values-based G20 AI Principles, and would contribute to shaping an enabling policy environment for AI.

Introduced by the White House's Office of Management and Budget (OMB) in January 2020, this draft memorandum[22] aims to govern the development and use of AI technologies in the private sector and detail the U.S. approach to AI regulation for innovators and entrepreneurs. Consistent with the American AI Initiative Executive Order, the OMB guidance seeks to support the U.S. approach to free markets, federalism, and good regulatory practices leading to a robust innovation ecosystem. The principles set out seek to clarify regulatory uncertainty that could hinder private sector innovation and development of AI technologies. Through these principles, the United States aims to advance emerging technology in a way that reflects its values of freedom, human rights, and civil liberties.

[22] https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf

2. Human-centered values and fairness

a) AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognized labor rights.

b) To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.

Explanation and rationale:

AI should be developed consistent with human-centred values, such as fundamental freedoms, equality, fairness, rule of law, social justice, data protection and privacy, as well as consumer rights and commercial fairness.

Some applications or uses of AI systems have implications for human rights, including risks that human rights (as defined in the Universal Declaration of Human Rights)[23] and human-centred values might be deliberately or accidently infringed. It is therefore important to promote "values-alignment" in AI systems (i.e., their design with appropriate safeguards) including capacity for human intervention and oversight, as appropriate to the context. This alignment can help ensure that AI systems' behaviours protect and promote human rights and align with human-centred values throughout their operation. Remaining true to shared democratic values will help strengthen public trust in AI and support the use of AI to protect human rights and reduce discrimination or other unfair and/or unequal outcomes.

This principle also acknowledges the role of measures such as human rights impact assessments (HRIAs) and human rights due diligence, human determination (i.e., a "human in the loop"), codes of ethical conduct, or quality labels and certifications intended to promote human-centred values and fairness.

Considerations for governments:

Striving to advance the G20 AI Principle on human-centred values and fairness can help tip the scale towards AI that protects and fulfils human rights and that uncovers and reduces bias, thereby reinforcing trust in its application, opening more opportunities for innovation, and boosting the benefits of AI. AI systems can have significant impacts on human rights, including exacerbating hate speech and incitement to violence online, and amplifying misinformation. They can also perpetuate bias in critical areas such as recruitment, justice, and personal finance. Acknowledging such potential outcomes and their chilling effect on the adoption and use of AI, and actively seeking to steer towards human-centred values and fairness, is critical.

[23] *Available at:* https://www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf

Examples of initiatives:

Australia – AI Ethics Principles

Australia's AI Ethics Principles are intended to provide organisations with signposts on how AI should be developed and used in Australia. While reflecting an important focus on fairness – noting that ensuring fairness across different groups in Australian society cuts to the heart of ethical AI – the AI Ethics Principles are also strongly aligned to the G20 AI Principles of transparency and explainability, robustness, security and safety, and accountability.

Australia's AI Ethics Principles[24] were launched in November 2019 and comprise eight voluntary principles intended to complement existing regulations. Their objective is to guide businesses and governments looking to design, develop and implement AI in Australia to:

- achieve better outcomes
- reduce the risk of negative impact
- practice the highest standards of ethical business and good governance.

The eight principles are: human, social and environmental well-being; human-centred values; fairness; privacy protection and security; reliability and safety; transparency and explainability; contestability; and accountability. They will be accompanied by practical guidance material with a strong focus on compatibility with international best practice. The main beneficiaries are envisaged as government agencies, businesses (including SMEs), data science groups and AI developers, those procuring AI systems, and corporate boards and executives with responsibility for outcomes of AI systems. Overseen by the Department of Industry, Science, Energy and Resources, the development of the AI Ethics Principles included a period of public consultation on a discussion paper, seeking input on *inter alia* the principles, the degree to which their implementation would meet the needs of customers/suppliers, and the likely support mechanisms needed to be able to implement ethical principles for AI.

Korea – Ethics Guidelines for the Intelligent Information Society

Korea's Ethics Guidelines, developed together with the National Information Society Agency, aim to achieve a human-oriented intelligent information society. Introduced in 2018, the guidelines strongly contribute to all the G20 values-based AI Principles.

Under the responsibility of the Ministry of Science and ICT (MSIT), Korea's Ethics Guidelines establish basic principles of intelligent information technology ethics with four common principles ("publicness", accountability, controllability, and transparency) and detailed guidelines for AI actors, specifically developers, suppliers and users of AI. The purpose of the guidelines is to strengthen the ethical responsibilities of developers and suppliers of intelligent information technologies and services and to provide guidance to prevent misuse of users. They are aimed at firms, academics, national government actors, and economic actors, especially workers.

24 https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework

Russian Federation – National Strategy for AI Development

Adopted in October 2019, Russia's National Strategy for AI Development will serve as the basis for development and enhancement of state programmes and projects as well as strategic documents of state-owned corporations and companies that support AI development in Russia. With ethics an important element of the Strategy, and with objectives spanning from research to regulation, the Strategy addresses a number of the values-based G20 AI Principles as well as supports fostering a digital ecosystem for AI.

Under the responsibility of the Ministry of Digital Development, Communications and Mass Media, this Strategy was adopted by the Russian President's decree of 10 October 2019, No. 490, "On the development of AI in the Russian Federation".[25] Its adoption seeks to assure expedited development of AI in the Russian Federation, scientific research, broader access to information and computing capacity for users, and improved workforce development system in this area. Development of ethical rules for human interaction with AI is an important element under the decree, as is scientific research to forecast the evolution of AI technologies as well as the social and ethical aspects of their application.

The Strategy has a number of objectives, including to:

- support scientific research to ensure expedited development of AI;
- design and develop software on the basis of AI technologies;
- improve availability and quality of data required for AI development;
- improve availability of hardware required to achieve AI-related objectives;
- improve availability of qualified human resources on the Russian AI market and increase the public awareness of potential AI applications; and
- build a comprehensive system to regulate social relations arising in connection with the development and use of AI technologies.

Switzerland – Guidelines for the implementation of AI-related measures by the federal government

In light of the rapid developments and broad discussions on AI, Switzerland's Federal Council has decided to develop strategic guidelines to orient federal agencies in their sectoral AI policies. With digital policy putting people at the forefront of an inclusive digital transformation, this initiative particularly addresses human-centred values and fairness in AI.

Building on the report "Challenges of artificial intelligence - Report of the interdepartmental working group on artificial intelligence to the Federal Council" of 13 December 2019[26], Switzerland is currently developing guidelines for the implementation of AI-related measures by the federal government and commissioned institutions. The Strategic Guidelines are intended to serve as an orientation for the responsible federal agencies in their AI sectorial policy, for identifying challenges or adapting regulations.

[25] http://www.kremlin.ru/acts/bank/44731 (in Russian); https://cset.georgetown.edu/wp-content/uploads/Decree-of-the-President-of-the-Russian-Federation-on-the-Development-of-Artificial-Intelligence-in-the-Russian-Federation-.pdf

[26] https://www.sbfi.admin.ch/sbfi/fr/home/le-sefri/numerisation/intelligence-artificielle.html

AI-relevant policies will also be included in the broader "Digital Switzerland" strategy[27] and its accompanying committees. AI is considered not as an isolated technology, but as an essential part of the ongoing digitalisation of the economy and society. Digital policy in Switzerland places people at the forefront of an inclusive democratic information and knowledge-based society. To ensure that people are integrated into the digital society, they must be involved in the processes of digital transformation. These include greater empowerment for independent action, the protection of people and their rights and more opportunities to participate actively in political and social life under digital conditions.

Turkey – National AI Strategy

Turkey's AI Strategy will aim to take a decisive step towards a human-centred, fair, transparent and explainable AI ecosystem. With a major intended output being a strategy document to lift nationwide awareness on data privacy and sharing and AI applications, the Strategy will contribute to implementing the G20 AI Principles, especially on human-centre values and fairness.

In 2020, Turkey's Presidency of Digital Transformation Office[28] intends to publish a National AI Strategy, seeking to foster interoperability between those actors holding a crucial position on technology development, notably government, academia, industry and NGOs. The strategy aims to implement data-driven active decision-making processes in governance, supporting the effective use of big data and AI applications by society, defining a framework for data access and sharing, introducing new regulations for ethics and law on AI, and proposing new action plans to advance qualified human resources. The document also has the intention to address fundamental principles such as human-centered development, fairness, transparency, trustworthiness, accountability, and commitment to ethical values.

[27] https://www.bakom.admin.ch/bakom/en/homepage/digital-switzerland-and-internet/strategie-digitale-schweiz.html

[28] www.cbddo.gov.tr

3. Transparency and explainability

AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:

i. to foster a general understanding of AI systems;

ii. to make stakeholders aware of their interactions with AI systems, including in the workplace;

iii. to enable those affected by an AI system to understand the outcome; and

iv. to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision

Explanation and rationale:

The term transparency carries multiple meanings. In the context of this principle, the focus is first on disclosing when AI is being used (in a prediction, recommendation or decision, or that the user is interacting directly with an AI-powered agent, such as a chatbot). Disclosure should be made with proportion to the importance of the interaction. The growing ubiquity of AI applications may influence the desirability, effectiveness or feasibility of disclosure in some cases.

Transparency further means enabling people to understand how an AI system is developed, trained, operates, and deployed in the relevant application domain, so that consumers, for example, can make more informed choices. Transparency also refers to the ability to provide meaningful information and clarity about what information is provided and why. Thus transparency does not in general extend to the disclosure of the source or other proprietary code or sharing of proprietary datasets, all of which may be too technically complex to be feasible or useful to understanding an outcome. Source code and datasets may also be subject to intellectual property, including trade secrets.

An additional aspect of transparency concerns facilitating public, multi-stakeholder discourse and the establishment of dedicated entities, as necessary, to foster general awareness and understanding of AI systems and increase acceptance and trust.

Explainability means enabling people affected by the outcome of an AI system to understand how it was arrived at. This entails providing easy-to-understand information to people affected by an AI system's outcome that can enable those adversely affected to challenge the outcome, notably – to the extent practicable – the factors and logic that led to an outcome. Notwithstanding, explainability can be achieved in different ways depending on the context (such as, the significance of the outcomes). For example, for some types of AI systems, requiring explainability may negatively affect the accuracy and performance of the system (as it may require reducing the solution variables to a set small enough that humans can understand, which could be suboptimal in complex, high-dimensional problems), or privacy and security. It may also increase complexity and costs, potentially putting AI actors that are micro, small and medium-sized enterprises (MSMEs) at a disproportionate disadvantage.

Therefore, when AI actors provide an explanation of an outcome, they may consider providing – in clear and simple terms, and as appropriate to the context – the main factors in a decision, the determinant factors, the data, logic or algorithm behind the specific outcome, or explaining why similar-looking

circumstances generated a different outcome. This should be done in a way that allows individuals to understand and challenge the outcome while respecting personal data protection obligations, if relevant.

Considerations for governments:

Advancing the G20 AI Principle on transparency and accountability is important not just for bolstering trust in AI systems but also for reinforcing core principles around human-centred values and fairness. Without transparency, identifying when human rights have been violated (or substantiating a claim of violation) is difficult. The same is true for seeking remedy, determining causality and accountability. Moreover, explanation is essential for situations in which fault needs to be determined in a specific instance – a situation that may grow more frequent as AI systems are deployed to make recommendations or decisions currently subject to human discretion.

Examples of initiatives:

Canada – Directive on Automated Decision-making

In the context of using AI to make, or assist in making, administrative decisions to improve service delivery, Canada has introduced a directive on automated decision-making to ensure compatibility with core administrative law principles such as transparency, accountability, legality, and procedural fairness. Accompanied by an Algorithmic Assessment Tool, this initiative addresses multiple G20 AI Principles, notably transparency and explainability, accountability, and security.

The Directive on Automated Decision-Making[29] launched in March of 2019, with compliance required by April 1, 2020. Under the responsibility of the Treasury Board of Canada Secretariat, the Directive applies to any Automated Decision System developed or procured after April 1, 2020. The objective of the Directive is to ensure that Automated Decision Systems are deployed in a manner that reduces risks to Canadians and federal institutions, and leads to more efficient, accurate, consistent, and interpretable decisions made pursuant to Canadian law.

The accompanying open source Algorithmic Impact Assessment (AIA) tool[30] was also unveiled in 2019, enabling government departments to gauge the potential risks of their use of automated systems. The questionnaire is designed to help government departments assess and mitigate the risks associated with deploying automated decision systems. It is a core component of the Directive as the publication of its outputs are required for compliance, fulfilling commitments to openness about automated decision-making within the Government of Canada.

The Directive and AIA tool enhance transparency in governance by ensuring that the use of automated decision-making in public service provision is disclosed to the public and other departments in detail. Automated decisions must also be individually transparent in that decisions must be explainable to those impacted by them in a meaningful way. They uphold the value of openness through the process in which they are developed. The AIA is open source and available on github. The Directive encourages that source code be released to the public as a default, with certain exemptions. The Directive also ensures accountability by stipulating consequences for compliance failure, and includes measures to ensure systems are robust and secure through stipulations on *inter alia* peer review, employee training requirements and monitoring.

[29] Directive on Automated Decision Making

[30] Algorithmic Impact Assessment Tool

4. Robustness, security and safety

a) AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk.

b) To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.

c) AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.

Explanation and rationale:

In this context, robustness signifies the ability to withstand or overcome adverse conditions, including digital security risks. This principle further states that AI systems should not pose unreasonable safety risks including to physical security, in conditions of normal or foreseeable use or misuse throughout their lifecycle. Existing laws and regulations in areas such as consumer protection already identify what constitutes unreasonable safety risks. Governments, in consultation with stakeholders, must determine to what extent they apply to AI systems.

AI actors can employ a risk management approach to identify and protect against foreseeable misuse, as well as against risks associated with use of AI systems for purposes other than those for which they were originally designed. Issues of robustness, security and safety of AI are interlinked. For example, digital security can affect the safety of connected products such as automobiles and home appliances if risks are not appropriately managed.

Two important ways to maintain robust, safe and secure AI systems are: i) traceability and subsequent analysis and inquiry, and ii) applying a risk management approach.

- Traceability: like explainability, traceability can help analysis and inquiry into the outcomes of an AI system and is a way to promote accountability. Traceability differs from explainability in that the focus is on maintaining records of data characteristics, such as metadata, data sources and data cleaning, but not necessarily the data themselves. In this, traceability can help to understand outcomes, to prevent future mistakes, and to improve the trustworthiness of the AI system.
- Risk management approach: AI systems pose potential risks to human rights, bodily integrity, privacy, fairness, equality and robustness. Protecting from these risks entails costs, including by building transparency, accountability, safety and security into AI systems. Different uses of AI present different risks, and some risks require a higher standard of prevention or mitigation than others. A risk management approach, applied throughout the AI system lifecycle, can help to identify, assess, prioritise and mitigate potential risks that can adversely affect a system's behaviour and outcomes. Documenting risk management decisions made at each lifecycle phase can contribute to the implementation of the other principles of transparency and accountability.

Considerations for governments:

Addressing the safety and security challenges of complex AI systems is critical to fostering trust in AI, placing a premium on advancing the G20 AI Principle on robustness, security and safety. Issues of robustness and safety of AI are interlinked, and have an impact on everyday life as well as critical economic and social activities. For example, digital security can affect product safety if connected products such as driverless cars or AI-powered home appliances are not sufficiently secure over their lifetime; hackers could take control of them and change settings at a distance. Similar concerns can exist for AI-enabled systems for finance, energy and other utilities, among others. Even if a system's operation results in only minor harm, it may still require high robustness if this harm occurs across a large number of people or sectors with collectively significant and disruptive effects.

Examples of initiatives:

No specific initiatives were listed under this Principle, though a number address it as part of their broad strategy, or via R&D efforts.

5. Accountability

AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.

Explanation and rationale:

The terms accountability, responsibility and liability are closely related yet different, and also carry different meanings across cultures and languages. Generally speaking, "accountability" implies an ethical, moral, or other expectation (e.g., as set out in management practices or codes of conduct) that guides individuals' or organisations' actions or conduct and allows them to explain reasons for which decisions and actions were taken. In the case of a negative outcome, it also implies taking action to ensure a better outcome in the future. "Liability" generally refers to adverse legal implications arising from a person's (or an organisation's) actions or inaction. "Responsibility" can also have ethical or moral expectations and can be used in both legal and non-legal contexts to refer to a causal link between an actor and an outcome.

Given these meanings, the term "accountability" best captures the essence of this principle. In this context, "accountability" refers to the expectation that organisations or individuals will: a) ensure the proper functioning, throughout their lifecycle, of the AI systems that they design, develop, operate or deploy, in accordance with their roles and applicable regulatory frameworks; and b) will demonstrate this through their actions and decision-making process (for example, by providing documentation on key decisions throughout the AI system lifecycle or conducting or allowing auditing where justified).

Considerations for governments:

Advancing the G20 AI Principle on accountability will help to place the onus on relevant parties to ensure that AI systems function properly and respect other G20 AI Principles of inclusive growth, human values, fairness, transparency, explainability, robustness and safety. In applications or functions where the potential for substantial harm is great, a lack of accountability could result in violation of social norms and legal constraints.

Examples of initiatives:

No specific initiatives were listed under this Principle, though a number address it as part of their broad strategy, or via R&D efforts.

6. Investing in AI research and development

a) Governments should consider long-term public investment, and encourage private investment, in research and development, including interdisciplinary efforts, to spur innovation in trustworthy AI that focus on challenging technical issues and on AI-related social, legal and ethical implications and policy issues.

b) Governments should also consider public investment and encourage private investment in open datasets that are representative and respect privacy and data protection to support an environment for AI research and development that is free of inappropriate bias and to improve interoperability and use of standards.

Explanation and rationale:

Scientific breakthroughs enabled by AI could help solve societal challenges and create entirely new industries. These possibilities underscore the importance of basic research and of considering long time horizons in research policy. While the private sector has taken the lead in applied AI R&D investments in recent years, governments, at times complemented by foundations focused on public good, have an important role to play in providing sustained investment in public research with long-term horizons. This type of investment is essential to driving and shaping trustworthy AI innovation and ensuring beneficial outcomes for all, particularly areas under-served by market-driven investments. Publicly funded research can help address challenging technological issues that affect a broad range of AI actors and stakeholders. AI research includes research in: AI applications, such as natural language processing; techniques to teach AI systems, such as neural networks; optimisation notably to reduce the amount of data required for AI development, such as one-shot-learning; and research addressing societal considerations, such as transparency and explainability, as well as technologies to protect data integrity.

In addition, because AI has broad reach and pervasive implications on multiple facets of life, this recommendation calls for investment in inter-disciplinary research, on social, legal and ethical implications of AI that are relevant to public policy.

Due to the importance of data to the AI system lifecycle, a key element in ensuring further and better AI R&D is the availability of open, accessible and representative datasets that do not compromise privacy and personal and consumer data protection, intellectual property rights and other important rights. In particular, while it may be impossible to achieve a completely “bias free” environment, by providing (and providing incentives for) representative datasets that are publicly available, governments can contribute to mitigating the risks of inappropriate bias in AI systems. For example, this could lessen the incidence of AI systems using datasets that are not sufficiently representative in accordance with the system’s intended use, even without the intention to discriminate.

This recommendation complements the recommendation on fostering a digital ecosystem for AI, since long-term investment in digital technologies and infrastructure and mechanisms for sharing AI knowledge are means to fostering this digital ecosystem. In particular, investment in open, accessible and representative datasets facilitates sharing of AI knowledge.

Considerations for governments:

Investing in AI R&D can help shape AI innovation to achieve the G20 AI Principles of inclusive growth and human-centred values, reinforcing broader strategic and policy directions with practical actions to drive the development of trustworthy AI. AI R&D can also help provide technical solutions to achieve transparency, explainability, robustness and accountability in efficient and effective ways. For instance, not every use of AI presents the same risks, and requiring explanation, for example, imposes its own set of costs. AI R&D can bring practical solutions to help achieve the suite of G20 AI Principles in innovative ways.

Examples of initiatives:

Australia – AI Technology Roadmap

Australia's AI Technology Roadmap is intended to help guide future investment in AI and provide a pathway to ensure Australia captures the full potential of AI. As well as identifying three high potential areas of AI specialisation, it elaborates the foundations needed in terms of skills, data governance, trust research, infrastructure and ethics, underscoring the mutual complementarity of the G20 AI Principles.

Australia's AI Technology Roadmap[31] was released in November 2019. It aims to help Australia harness AI to boost Australian industry and the economy (notably productivity), create jobs and improve quality of life for current and future generations. It identifies three domains of AI development and application where AI could transform Australian industry, based on existing strengths and comparative advantages, opportunities to solve Australian problems, and opportunities to export solutions to the rest of the world. These domains are AI for better health, aged care and disability services; AI for better towns, cities and infrastructure (including connected and automated vehicle technology); and AI for better natural resource management (especially building on strengths related to mining and agriculture).

At the same time, the Roadmap highlights that for Australia to take full advantage of the potential benefits of AI technologies, it must build the right environment for effective adoption, adaption and development, including removing barriers to growth. Foundations for the future include developing an AI specialist workforce, career transitions and skills upgrades, data governance and access, building trust in AI, science, research and technology development, digital infrastructure and cybersecurity, and standards, interoperability and ethics. Overseen by the Department of Industry, Science, Energy and Resources and CSIRO, the development of the Roadmap was allocated AUD 360 000 in the 2018-19 Budget.

[31] https://data61.csiro.au/en/Our-Research/Our-Work/AI-Roadmap

Brazil – Applied Research Centres in AI

Brazil will establish eight applied research centres in AI, with the aim to conduct research, foster an AI eco-system and stimulate start-ups, and build human capacity in related technologies. These multidimensional aims reinforce the need for implementation across the range of G20 AI Principles.

Starting in 2020, this initiative[32] under the responsibility of Brazil's Ministry of Science, Technology, Innovation and Communications will establish eight centres hosted by established academic and research institutions. They will be selected on a competitive basis among initial project proposals in four focus areas: health, agriculture, industry and smart cities. Academic/research entities and private sector companies must jointly present the project proposals. The funding mechanism matches private sector funds with public funds, for an initial commitment of 5 years, renewable once for an additional period of 5 years. The objective of the research centers is to bring together governmental, academic and private sector entities. This initiative, combining the creation of infrastructures and providing partial direct financial support (BRL 1 million per research centre), aims to benefit the private and public sectors and the workforce.

Germany – call for R&D projects on explainability and transparency

Explainability and transparency are necessary prerequisites for a trustworthy use of AI methods. Germany is providing funding for research and development projects that aim to improve the comprehensibility, explainability and transparency of machine learning and AI, especially with regard to practical applications. This underscores the interplay between the values-based and policy-oriented elements of the G20 AI Principles as well as the importance of R&D to develop practical solutions to AI policy challenges.

In April 2019 Germany's Ministry of Education and Research launched a call for R&D projects on explainability and transparency of machine learning and AI, with the goal of developing new methods and tools to increase the transparency and traceability of algorithmic decisions of learning systems.[33] The call recognises that for a better understanding of the functioning of AI systems, a combination of new approaches from different areas such as physics, mathematics and cognitive science is necessary. The funding is aimed at collaborative projects between science and industry in an interdisciplinary composition. Ten projects are funded, at a total of EUR 11 million.

Korea – AI R&D Strategy

Korea's AI R&D Strategy was developed after a comprehensive analysis of the state of the nation's AI technology, talent and infrastructure. It aims to secure a world-class AI technology and R&D ecosystem.

The AI R&D Strategy covers the period 2018-2022 and includes a focus on domestic competitiveness. The Strategy's beneficiaries are research institutes, businesses, universities and citizens, with a budget of KRW 400 billion per year. The Strategy is overseen by the Ministry of Science and IT and incorporates elements of direct financial support, collaborative infrastructures, regulation and incentives.

32 http://www.mctic.gov.br/mctic/opencms/textogeral/Os-Centros-de-Pesquisa-Aplicada-CPA-em-Inteligencia-Artificial-IA.html

33 https://www.softwaresysteme.pt-dlr.de/de/ki-erkl-rbarkeit-und-transparenz.php

Saudi Arabia – National Center for AI (NCAI)

The Saudi Data and Artificial Intelligence Authority (SDAIA), established in 2019, has introduced the National Center for AI (NCAI) as its AI innovation and capability-building arm. Its vision is to enable Saudi Arabia's global leadership in AI, including by orchestrating AI research and solutions development.

The NCAI was established in August 2019, with its main mission to drive Saudi advancement in AI innovations by orchestrating AI research and solutions development, providing AI strategic advisory to the government, and promoting AI education and awareness. NCAI aims to be a sustainable world-class center of excellence leveraging collaborations with academia, and the public and private sectors. It has four objectives:

- Drive national AI strategy: by detailing the national AI strategy in collaboration with other actors and promoting its implementation across the nation.
- Orchestrate AI research: by advancing basic and applied AI knowledge, becoming KSA's reference point for AI research & innovation.
- Develop AI solutions: by generating tangible AI applications for different government entities directly or through partnerships and incubators.
- Build AI expertise: by offering training to workforce and supporting AI education.

Connected to this, several national initiatives have been implemented to build national capabilities in data science and AI. For instance, KSA organised a national 15-week upskilling bootcamp – "Tuwaiq Data Scientists bootcamp" – aiming to provide the market with much needed data-science and AI specialists. KSA is also ensuring inclusive AI awareness by organising multidisciplinary events bringing together AI experts and specialists in other fields. KSA has organised an innovative sport AI hackathon in which teams developed algorithms to predict the national football league results. KSA also organised the world's first AI art Artathon (hackathon), in which artists and data scientists worked together to co-create pieces of art with AI. The Artathon attracted over 2 000 applicants from over 50 countries.

Saudi Arabia – National Centre of Data Analytics and AI

The King Abdulaziz City for Science and Technology (KACST), which encompasses national research institutes, has established a national centre for AI R&D. This effort is complemented by the establishment of university programmes in AI and data analytics and contributes to Saudi Arabia's broader goal of harnessing Industry 4.0 technologies to transform the country to an industrial and logistics hub.

Singapore – Research Programme on the Governance of AI and Data Use

This Research Programme aims to develop international thought leadership and advance scholarship and discourse in legal and ethical issues arising from the use of AI and data. Intended to benefit both industry and government, the initiative can inform implementation of all five values-based G20 AI Principles.

The Research Programme was established in June 2018 following the award of a SGD 4.5 million research grant from the National Research Foundation and the Infocomm Media Development Authority to the

Singapore Management University's School of Law. Housed in the Centre for AI and Data Governance, the Research Programme produces research publications and stakeholder (local and international) engagement events and activities, aimed at providing guidance on long-term and complex issues with regard to AI. The Research Programme's various streams of work address and engage with all five principles of the G20 AI Principles. The Research Programme's work is intended to help both industry (through industry-relevant research) and the Government (through research on long-term and complex issues relating to AI).

The Research Programme is one of three interlinked initiatives under Singapore's National AI Strategy.

Turkey – Industry and Technology Strategy

Turkey's Industry and Technology Strategy 2023 includes sectoral and R&D competency mapping on AI technology and AI and machine learning, with a view to strengthening Turkey's capacity of scientific research and product development. With plans to establish an AI Research Institute bridging the private and public sectors, this initiative also contributes to shaping an enabling policy environment for AI.

In 2019 Turkey's Ministry of Industry and Technology published the 2023 Industry and Technology Strategy[34], taking a holistic approach to the fields of industry and technology, and aiming to ensure wide participation and to mobilise society. The Strategy consists of five main pillars: "High Technology and Innovation", "Digital Transformation and Industry Move", "Entrepreneurship", "Human Capital" and "Infrastructure". The Strategy Paper is in concert with Turkey's 2023 Goals.

Specific to AI, the Strategy will see preparation of an R&D competency map to analyse the present state of in AI technology (in particular AI and machine learning) in Turkey, notably the capacity of scientific research and product development. In 2020 a sectoral road map will be developed in cooperation with the sectoral actors, consisting of product-oriented targets in areas such as software, aviation and space. In addition, studies are ongoing for the establishment of an AI Research Institute that would act as a bridge between the public and private sectors by developing key technologies and providing academic results to industry for innovative implementation. The institute held a stakeholder workshop in February 2020.

United Arab Emirates – Synthetic Data

This R&D experiment with an AI business is exploring the potential of advanced machine learning approaches to create "synthetic data" that may enable wider use in AI applications of sensitive data. This R&D investment also contributes to fostering a digital ecosystem for AI.

This project on unlocking the power of sensitive data for use in AI is running from December 2019 to March 2020 and is budgeted at AED 1 million. It is aiming to test (through attack scenarios) the degree to which synthetic data can retain the privacy and other sensitive aspects of classified data, whilst also retaining the underlying relationships between variables. It also aims to test the technical performance of AI models trained on synthetic data alongside models using real data, and to establish the usefulness of synthetic data in a range of use cases (focused on highly sophisticated analytics and AI product development). Finally, it is exploring the potential of creating a "synthetic city data store" to boost the supply of high quality, high value data for AI.

[34] https://www.sanayi.gov.tr/strateji2023/sts-ktp.pdf

The intended beneficiaries of this initiative include government entities, who are able to liberate more data for use in sophisticated AI, AI developers, who are able to use this data in sophisticated AI, and citizens and businesses, who may benefit from the downstream creation of economic and social value. The initiative is overseen by Smart Dubai Data and Faculty AI.

United States – American AI Initiative

The American AI Initiative is a concerted whole-of-government effort to promote and protect national AI technology and innovation. While including significant budget expenditures on AI R&D, the Initiative promotes a multi-pronged approach to advance AI, tackling all policy-oriented elements of the G20 AI Principles.

Established via Executive Order (E.O.) in February 2019, the American AI Initiative[35] implements a whole-of-government strategy in collaboration and engagement with the private sector, academia, the public, and like-minded international partners. The Initiative emphasises that AI should be used thoughtfully, as noted in the President's E.O.: "The United States must foster public trust and confidence in AI technologies and protect civil liberties, privacy, and American values in their application in order to fully realize the potential of AI technologies for the American people."

The AI Initiative directs federal agencies to pursue a multipronged approach to advance AI, with important emphases in areas of AI research and development, data and computational resources, technical standards, regulatory and non-regulatory approaches to the use of AI, education and workforce, and international engagement. Some of the items, such as prioritising AI R&D, have budget expenditures. The President's FY2021 Budget Request committed to doubling government-wide non-defense investments on AI R&D by FY 2022, compared to the FY2020 level of nearly USD 1 billion. Information on progress and impact can be found at AI.gov.

United States – National AI R&D Strategic Plan

The 2019 update to the National AI R&D Strategic Plan defines priority areas for Federal investments in AI R&D and stands in support of the American AI Initiative. It was developed by leading AI researchers and research administrators from across the Federal Government, with input from the broader civil society, including from many of America's leading academic research institutions, nonprofit organisations, and private sector technology companies.

Following the first National AI R&D Strategic Plan released in 2016, this update[36] is focused on addressing the R&D priorities associated with advancing AI technologies. It accounts for new research, technical innovations, and other considerations that have emerged since 2016, and specifically identifies public-private partnerships for AI R&D as a new priority. The Plan establishes a set of objectives for federally funded AI research, identifying eight strategic priorities:

1. Make long-term investments in AI research,

[35] https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/; https://www.whitehouse.gov/ai/

[36] https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf and AI.gov

2. Develop effective methods for human-AI collaboration. Increase understanding of how to create AI systems that effectively complement and augment human capabilities,
3. Understand and address the ethical, legal, and societal implications of AI. Research AI systems that incorporate ethical, legal, and societal concerns through technical mechanisms,
4. Ensure the safety and security of AI systems. Advance knowledge of how to design AI systems that are reliable, dependable, safe, and trustworthy,
5. Develop shared public datasets and environments for AI training and testing. Develop and enable access to high-quality datasets and environments, as well as to testing and training resources,
6. Measure and evaluate AI technologies through standards and benchmarks. Develop a broad spectrum of evaluative techniques for AI, including technical standards and benchmarks,
7. Better understand the national AI R&D workforce needs. Improve opportunities for R&D workforce development to strategically foster an AI-ready workforce, and
8. Expand public-private partnerships to accelerate advances in AI. Promote opportunities for sustained investment in AI R&D and for transitioning advances into practical capabilities, in collaboration with academia, industry, international partners, and other non-Federal entities.

R&D performers, Federal agencies and the American people are considered the beneficiaries of this policy action. A progress report for the 2016-19 period is available at https://www.whitehouse.gov/wp-content/uploads/2019/11/AI-Research-and-Development-Progress-Report-2016-2019.pdf.

Other examples

Other policies aiming at investment in AI R&D include:

- The AI-Russia Alliance, a cross-sectoral initiative established in 2019. The Alliance includes Russian leading companies Sberbank, Yandex, Mail.ru Group, Gazprom Neft, MTS as well as the Russian Direct Investment Fund. The alliance aims to invest in R&D, develop digital infrastructure, provide education programs as well as monitor the implementation of the National Strategy for AI Development.

7. Fostering a digital ecosystem for AI

Governments should foster the development of, and access to, a digital ecosystem for trustworthy AI. Such an ecosystem includes in particular digital technologies and infrastructure, and mechanisms for sharing AI knowledge, as appropriate. In this regard, governments should consider promoting mechanisms, such as data trusts, to support the safe, fair, legal and ethical sharing of data.

Explanation and rationale:

The development of trustworthy AI requires an enabling ecosystem. This recommendation calls on governments – engaging with the private sector as appropriate – to work towards providing, or promoting the provision of, the infrastructure and digital technologies for AI and the mechanisms for AI knowledge sharing, taking into account their national frameworks.

The necessary digital technologies and infrastructure include access to affordable high-speed broadband networks and services, computing power and data storage – as well as supporting data-generating technologies such as the Internet-of-Things (IoT). For example, recent AI advances can be attributed, in part, to the exponential increase in computational speed including with graphics-processing unit resources. Appropriate mechanisms for sharing AI knowledge, including data, code, algorithms, models, research, and know-how, are also required to understand and participate in the AI system lifecycle. Such mechanisms must respect privacy, intellectual property and other rights. Open source tools and high-quality training datasets for managing and using AI, which allow for the diffusion of AI technology and crowdsourcing solutions to software bugs play a key role in AI development.

When developing means of data sharing, such as data trusts or trusted third parties, it is important to pay attention to risks related to data access and sharing. Risks to individuals (including consumers), organisations and countries of sharing data may include confidentiality and privacy breaches, risks to intellectual property rights, data protection, competition and commercial interests, as well as potential national security and digital security risks. Promoting and utilising datasets that are as inclusive, diverse and representative as possible is key.

This recommendation draws special attention to policies for micro, small and medium-sized enterprises (MSMEs): to facilitate MSME access to data, AI technologies and relevant infrastructure (such as connectivity, computing capacities and cloud platforms) in order to foster digital entrepreneurship, competition and innovation through the adoption of AI.

Considerations for governments:

Progress in the development of trustworthy AI, and its responsible stewardship, may be accelerated or hindered by the state of the digital ecosystem, underscoring the importance of fostering such an ecosystem in G20 economies. Similar to the digital transformation more broadly, establishing basic building blocks of digital infrastructure and access to quality data is fundamental for enabling firms, governments and individuals to reap the benefits of AI.

Examples of initiatives:

Korea – Data and AI-Driven Economy Promotion Plan

This initiative aims to promote active use of the data value chain throughout its life cycle and foster a world-class AI innovation ecosystem in Korea.

Established in 2019 and under the responsibility of the Ministry of Science and ICT, Korea's Data and AI-Driven Economy Promotion Plan is a 5-year implementation strategy for fostering data and AI and promoting their convergence. By promoting active use of the data value chain, it seeks to establish Korea as a data and AI powerhouse. The plan incorporates elements of direct financial support and collaborative infrastructures and aims to benefit research institutes, firms, academics and government entities.

Mexico – Program for the Development of the Software Industry (PROSOFT) and Innovation

This initiative led by the Mexican Ministry of Economy has supported creation of two Industrial Innovation Centres to promote AI adoption, the generation and specialisation of human resources and the transfer of knowledge. The span of activities and its regional dimensions means this initiative not only contributes to fostering a digital ecosystem and human capacity for AI, but also to inclusive growth, sustainable development and well-being, thereby fostering implementation of several G20 AI Principles.

PROSOFT and Innovation[37] was originally introduced to support the Mexican software industry. In 2016 it widened its scope to support the adoption of Industry 4.0 technologies across all industry sectors and promote the creation of regional innovation ecosystems. It has led to the creation of industrial innovation centres that promote development and adoption of AI in two different Mexican regions:

- The "Centre for Industrial Innovation in Artificial Intelligence of the Monterrey IT Cluster", which seeks to facilitate the development of human capital for the adoption of Industry 4.0 models, through the availability of specialised technologies based on AI, in addition to having a Development Laboratory with a focus on AI technology.
- The "Centre for Industrial Innovation in Artificial Intelligence Engineering in Yucatan", which seeks to contribute to the academic training of specialised human talent and the generation of goods and services articulated to the industry's demands, under a collaborative and participatory governance model. It also seeks to create the first master degree in AI and Industry 4.0 in Latin America.

The centres receive direct financial support (of between USD 1.2 and 1.5 million each in 2018), with medium and large enterprises the main beneficiaries.

[37] http://prosoft.economia.gob.mx/

Russian Federation – National Programme on "Digital Economy"

This overarching programme aims to achieve national goals in digital development by creating framework conditions for digitalisation and the development of cross-cutting technologies such as AI. Including a focus on infrastructure, the programme contributes strongly to the digital ecosystem in Russia.

Under the responsibility of the Ministry of Digital Development, Communications and Mass Media, the Digital Economy programme[38] aims at goals set out in the President's Decree No. 204 (7 May 2018) "On the National Development Goals and Strategic Objectives of the Russian Federation in the Period until 2024". The programme is aimed at creating framework conditions for digitalisation and penetration of digital-based technologies in most the sectors of the Russian economy. Among other things, the programme is targeted at supporting the development of such cross-cutting technologies as: big data, neurotechnologies and artificial intelligence, blockchain, quantum technologies, advanced manufacturing technologies, industrial Internet, robotics, wireless technologies, virtual and augmented reality. It aims at three primary objectives:

- An increase in domestic expenditure on the development of the digital economy from all sources (as a share of GDP) by at least 3 times compared to 2017.
- To create a stable and secure information and telecommunications infrastructure for high-speed transmission, processing and storage of large amounts of data, accessible to all organisations and households.
- The use of predominantly domestic software by government bodies, local governments and organisations.

The programme entails annual budget expenditure of over RUB 500 million.

Saudi Arabia – Saudi Data and Artificial Intelligence Authority

The Saudi Data and Artificial Intelligence Authority (SDAIA) was established in 2019 as the owner of Saudi Arabia's national and artificial intelligence agenda, mandated with unlocking the value of data and AI to elevate Saudi Arabia as a pioneering nation among the elite league of data-driven economies. Addressing data, innovation and capability-building, SDAIA serves to foster the digital ecosystem while also supporting the suite of values-based G20 AI Principles.

The key role of SDAIA is to help unlock the value of AI for the nation and for the good of humanity in general. By building key AI pillars, including data usage and sharing regulations and AI ethics, data operations and AI-powered insights generation, and AI innovation and capability building, in a holistic manner, SDAIA helps achieve consistency, transparency and fairness across sectors. SDAIA operates through three arms: the National Data Management Office (NDMO), which is the main data regulator and enabler of national data and AI policies and best practices; the National Information Center (NIC) which is the main operator of government data infrastructure, advanced analytics and AI-powered insight platforms, and G-cloud services; and the National Center for AI (NCAI) which drives implementation of the national AI strategy, AI innovation and capability-building.

[38] https://oecd.ai/dashboards/policy-initiatives/2019%2Fdata%2FpolicyInitiatives%2F24621; http://government.ru/info/35568/

SDAIA's Board of Directors reports to the Prime Minister and is chaired by the Deputy Prime Minister. The body contributes to implementation of the values-based G20 AI Principles in multiple ways:

- **Inclusive growth, sustainable development and well-being:** NDMO under SDAIA will develop regulations and policies, including open data policies, to provide a fair playing field for all segments of society to create opportunities with data and AI. NCAI will enable society through different capability building programs targeting AI literacy to AI deep expertise. NCAI will also focus R&D activities around national priorities including inclusive growth. NIC is operating a national whole-of-government analytics and AI platform, which generates AI-powered scenarios to support Saudi Arabia's decision makers. These insights focus typically on socioeconomic priorities such as welfare, social protection, employment, and education.
- **Human-centred values and fairness:** SDAIA is developing a national data and AI strategy focusing on the needs of the nation and prioritising sectors according to those needs. NDMO will define, enable and enforce a comprehensive data and AI framework to ensure data and AI are used and applied ethically and fairly. SDAIA is also planning to lead the global dialogue on human-centric AI by organising an annual thought leadership event "The Global AI Summit" under the theme "AI for the Good of humanity".
- **Transparency and explainability:** NDMO will regulate data and AI in a way that would minimise negative effect on the society.
- **Robustness, security and safety, and accountability:** NDMO regulations will have special focus on protecting data, and organising the way it is used while enabling value-add applications and opportunities. NIC is running the government's data infrastructure and the national data bank, leveraging best-in-class standards and technologies to protect the nation's data assets and ensure their availability and stability. SDAIA is accountable for the national data and AI agenda, responsible for data and AI regulations, operations and innovation.

Spain – Plan for Advancement of Language Technologies

Spain's Plan for Advancement of Language Technologies aims to develop the natural language processing industry, machine translation and conversational systems in Spain, especially in Spanish and co-official languages. As well as contributing to the digital ecosystem for AI, the practical outcomes of this initiative are well-aligned with the values-based G20 AI Principle of inclusive growth, sustainable development and well-being.

The Spanish Plan for Advancement of Language Technologies[39], running over 2015-2020 and under the responsibility of the Secretary of State for Digitisation and AI, aims to:

- Drive development of linguistic infrastructures, adoption of technical standards for interoperability and promotion of methods of automatic resource generation.
- Boost the language technology industry by improving its visibility, supporting internationalisation and cooperating with the Latin American community.

39 https://www.plantl.gob.es/tecnologias-lenguaje/PTL/Bibliotecaimpulsotecnologiaslenguaje/Detalle%20del%20Plan/Plan-Advancement-Language-Technology.pdf

- Establish the public administration as a driver of the language industry, with the creation of common platforms for natural language processing and machine translation and the development of resources for the reuse of public sector information.

With budget expenditure of EUR 5 million p.a., the Plan has four action areas. Axis 1 comprises support for the development of language infrastructures (i.e. processors and linguistic resources that serve as fuel for the development of the natural language processing). Axis 2 aims for promotion of the language technologies industry (especially supporting the transfer of knowledge between the research sector and industry, as well as the internationalisation of the companies and institutions that make up the sector). Axis 3 aims to create common language processing, conversational and machine translation platforms for Public Administrations (and, in addition, leverage the reuse of public sector information as a channel for the development of linguistic resources). Axis 4 is aimed at promoting lighthouse projects for the application of natural language technologies, initially undertaken by Public Administrations in strategic sectors (e.g. Health, Justice and Education), and open to future public-private initiatives, which are intended to serve as a demonstration of their capabilities and benefits, generate industry and create resources that can be reused in other projects.

Turkey – Open Data Project and Open Government Partnership

Turkey's Open Data Project will establish an open data portal so citizens, researchers, public institutions and organisations, and state affiliates can leverage data produced by public resources. As a platform for the datasets needed for developing AI technologies and applications, the initiative will contribute to fostering a digital ecosystem for AI.

The Open Data Project will be launched under the responsibility of Turkey's Presidency of Digital Transformation Office in 2020. The main focus is to establish an open data portal but the project will also manage the regulatory and legislative steps for participation in the Open Government Partnership. The Open Data Project will provide a distributed and scaleable data management system for AI research that requires cleaned, labelled, classified and stored datasets. Initial action will establish the infrastructure required for the open data portal and raise awareness among the institutions and organisations within the digital ecosystem. In the next phase, open data collected from all public institutions and organisations will be uploaded to the portal in appropriate formats and the portal will be made available for usage. Organisations seeking to take part in the Open Government Partnership will need to meet certain regulatory and legislative requirements. The initiative also aims to provide effective coordination in preparing the labour market for digital transformation.

United Kingdom – Centre for Data Ethics and Innovation

The Centre for Data Ethics and Innovation will have an explicit and unique mandate to advise the UK government on the measures which are needed to ensure safe and ethical innovation in data and AI. This initiative serves to further foster a digital ecosystem for AI in the UK, and contributes to the values-based G20 AI Principles, especially human-centred values and fairness.

Established in 2018, the Centre for Data Ethics and Innovation will convene, connect and build upon the best evidence, insights and practices available and translate these into direct, real-world actions that enhance the way in which data and AI are used. It will identify the measures needed to strengthen and

improve the way data and AI are used and regulated. This will include articulating best practice and advising on how to address potential gaps in regulation. The Centre will not, itself, regulate the use of data and AI - its role will be to help ensure that those who govern and regulate the use of data across sectors do so effectively.

Across its work, the Centre will seek to deliver the best possible outcomes for society from the use of data and AI. This includes supporting innovative and ethical uses of data and AI. These objectives will be mutually reinforcing: by ensuring data and AI are used ethically, the Centre will promote trust in these technologies, which will in turn help to drive the growth of responsible innovation and strengthen the UK's position as a trusted place for data-driven businesses to invest in.

8. Shaping an enabling policy environment for AI

a) Governments should promote a policy environment that supports an agile transition from the research and development stage to the deployment and operation stage for trustworthy AI systems. To this effect, they should consider using experimentation to provide a controlled environment in which AI systems can be tested, and scaled-up, as appropriate.

b) Governments should review and adapt, as appropriate, their policy and regulatory frameworks and assessment mechanisms as they apply to AI systems to encourage innovation and competition for trustworthy AI.

Explanation and rationale:

This recommendation focuses on the policy environment that enables AI innovation – in other words, the institutional, policy and legal framework. It complements the previous recommendation on the necessary physical and technological infrastructure and similarly highlights the need to pay particular attention to MSMEs.

Considering the fast pace of AI developments, setting a policy environment that is flexible enough to keep up with developments and promote innovation yet remains safe and provides legal certainty is a significant challenge. This recommendation seeks to address this challenge by identifying means for improving the adaptability, reactivity, versatility and enforcement of policy instruments, in order to responsibly accelerate the transition from development to deployment and where relevant, commercialisation. In fostering the use of AI, a human-centred approach should be taken, in accordance with the earlier Principles on inclusive growth and human-centred values.

The recommendation highlights the role of experimentation as a means to provide controlled and transparent environments in which AI systems can be tested, and in which AI-based business models that could promote solutions to global challenges can flourish. Policy experiments can operate in “start-up mode” whereby experiments are deployed, evaluated and modified, and then scaled up or down, or abandoned, depending on the test outcomes.

The recommendation acknowledges the importance of oversight and assessment mechanisms to complement policy frameworks and experimentation. In this respect, there may be room for encouraging AI actors to develop self-regulatory mechanisms, such as codes of conduct, voluntary standards and best practices. Such initiatives can help guide AI actors through the AI lifecycle, including for monitoring, reporting, assessing and addressing harmful effects or misuse of AI systems. To the extent possible and relevant, these mechanisms should be transparent and public.

Considerations for governments:

An enabling policy environment for AI is critical for fostering research and preserving economic incentives to innovate, while also providing guardrails for ensuring human rights, data protection, privacy, and other critical elements of trustworthy AI. Without it, AI actors will struggle to bring beneficial AI applications to economies and societies, and the ambition to reach other G20 AI Principles will be curtailed.

Examples of initiatives:

Canada – Government of Canada Advisory Council on AI

Canada's Advisory Council on AI is mandated to build on Canada's strengths in AI, identify new opportunities in the AI sector and make recommendations to ensure Canadians benefit from the growth of the AI sector in terms of jobs and growth. Its wide scope will help it to shape an enabling policy environment for AI and to contribute to the values-based G20 AI Principles of human-centred values and inclusive growth.

The Advisory Council on AI[40] was announced in May 2019, under the responsibility of the Department of Innovation Science and Economic Development Canada. Its objectives are to create more jobs for Canadians, further Canada's position as a global leader in AI R&D, better support entrepreneurs and scale-ups and ensure Canadians have the education and skills they need to succeed in a changing economy. It is also tasked with conducting a dialogue with Canadians to ultimately foster trust in AI to better ground Canadian discourse in a measured understanding of the technology, its potential uses, and its associated risks. Finally the Council is advising the Canadian Government on its international positions on AI at the G7/G20, the OECD, and the World Economic Forum. It will also provide recommendations regarding the establishment of the Global Partnership on Artificial Intelligence, which aims to support and guide the responsible adoption of AI that is human centric and grounded in human rights, inclusion, diversity, innovation and economic growth.

China – Trustworthy AI Operational Guidelines

The China Artificial Intelligence Industry Alliance (AIIA) is developing norms and principles to apply to product R&D and operational services of enterprises, with the aim of facilitating healthy development and self-regulation of the AI industry while sharing enterprises' best practices. This initiative, to be released in July 2020, shows the relevance of the G20 AI Principles to all AI stakeholders.

The objective of the Trustworthy AI Operational Guidelines is to create an enabling and trustworthy environment for a safe, reliable and controllable AI industry; establish an assessment and evaluation framework for AI products; constantly monitor, support and promote the quality, intelligence, reliability and safety of major products and services; and facilitate the in-depth integration of AI and the real economy for the accelerated development of the AI industry.

The initiative builds on the AIIA's *Artificial Intelligence Industry Self-Regulatory Convention* released in May 2019 and can be seen in the context of Chinese policymakers' emphasis on the relevance of R&D on AI-related legal, ethical and social issues to ensure safety, reliability and controllability of AI. Both the *New Generation Artificial Intelligence Development Plan* issued by the State Council, and the *Three-Year Action Plan for Developing New Generation Artificial Intelligence Industry* (2018-2020) issued by the Ministry of Industry and Information Technology, stipulate relevant requirements and goals, including to enhance risk evaluation and prevention; safeguard public interests and national security; develop and improve laws and regulation, institutional systems and ethics in support of healthy AI development; and build capacity in AI safety assessment and governance.

[40] Advisory Council Terms of Reference

European Union – EU Strategy on AI: AI for Europe

This initiative sets out Europe's ambition to boost its technological and industrial capacity and AI uptake across the economy, prepare for socio-economic changes and ensure an appropriate ethical and legal framework. It aims to build and strengthen the environment for Europe to become a leader in the AI revolution.

In April 2018, the European Commission's Communication on Artificial Intelligence for Europe[41] set out the need for a European framework to chart a course amid fierce global competition in AI. The initiative aims to:

- Boost the EU's technological and industrial capacity and AI uptake across the economy, both by the private and public sectors. This includes investments in research and innovation and better access to data.
- Prepare for socio-economic changes brought about by AI by encouraging the modernisation of education and training systems, nurturing talent, anticipating changes in the labour market, supporting labour market transitions and adaptation of social protection systems.
- Ensure an appropriate ethical and legal framework, based on the Union's values and in line with the Charter of Fundamental Rights of the EU. This includes guidance on existing product liability rules, a detailed analysis of emerging challenges, and cooperation with stakeholders, through a European AI Alliance, for the development of AI ethics guidelines.

The approach to AI described in the Strategy shows the way forward and highlights the need to join forces at European level, to ensure that all Europeans are part of the digital transformation, that adequate resources are devoted to AI and that the Union's values and fundamental rights are at the forefront of the AI landscape. The Strategy is complemented by the EU's Co-ordinated Plan on AI (discussed further below, under the Principle of international co-operation for trustworthy AI).

Singapore – Advisory Council on the Ethical Use of AI and Data

This industry-led Advisory Council brings together leaders of international and local companies with advocates of social and consumer interests, to advise on legal and ethical issues from the use of data-driven technologies. Its activities particularly contribute to building human capacity, and its work addresses all five values-based G20 AI Principles.

The Advisory Council on the Ethical Use of AI and Data[42] was established in 2018 under the guidance of Singapore's Infocomm Media Development Authority to: (i) provide guidance businesses to minimise the legal/ethical and sustainability risks, and to mitigate adverse impact on consumers, from the use of data-driven technologies; and (ii) advise the Government on ethical and legal issues arising from the use of such technologies in the private sector. Amongst its activities, the Advisory Council assists the government in engaging stakeholders on issues that support the development of AI governance capabilities and frameworks (e.g., the Model Framework). The 11 members of the Council include both developers and

[41] https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe

[42] https://www.imda.gov.sg/news-and-events/Media-Room/Media-Releases/2018/composition-of-the-advisory-council-on-the-ethical-use-of-ai-and-data and https://www.imda.gov.sg/-/media/Imda/Files/About/Media-Releases/2018/2018-06-05-Fact-Sheet-for-AI-Govt.pdf

users of AI technologies, who provide recommendations to the industry and government. The work of the Advisory Council engages and addresses all five principles of the G20 AI Principles such as by advising and engaging stakeholders on the Model Framework, which touches on ethical considerations aligned with the G20 AI Principles.

The Advisory Council forms one of three interlinked initiatives under Singapore's National AI Strategy.

Other examples

Other policies seeking to shape an enabling policy environment for AI include:

- Australia's AI Standards Roadmap, currently under development and intended to identify priority areas for AI standards development and a pathway for Australian leadership on international standardisation activities for AI. This work is led by Standards Australia and the Department of Industry, Innovation and Science.
- Korea's Regulatory Sandbox, introduced in 2019, grants a regulatory waiver for a set amount of time to businesses, allowing them to test innovative products, services and business models in the market. The sandbox seeks to allow the government to improve related regulations based on real-life data. The initiative is overseen by the Ministry of Science and ICT (MSIT), Ministry of Trade, Industry and Energy (MOTIE), and Ministry of SMEs and Startups (MSS).
- Russia's establishment of a Sub-Commission on AI Development[43], a multistakeholder body responsible for coordinating AI development in Russia and the implementation of the National Strategy for AI Development. The Sub-Commission was created in 2019 under the *Government Commission on Digital Development and the Use of ICT to Improve the Quality of Life and Doing-Business Conditions*. The Ministry of Economic Development and Ministry of Digital Development, Communication and Mass Media hold responsibility for the initiative.
- Russia's Digital Rights Law[44], which came into force in October 2019 and introduced a number of new legal concepts into Russian legislation, including digital rights, e-transactions, smart contracts, and big data. The law aimed to enable the development of an efficient legal framework of digital economy in Russia, reflecting current digital technologies and challenges including big data and AI.
- Russia's draft conceptual framework for Legal Regulation of AI Technologies and Robotics[45], which aims to establish a legal framework for the further development of AI technologies and robotics in Russia and eliminate excessive legal barriers impeding the use of AI-systems for economic and social purposes. The initiative aims to give guidance for regulators and is under the responsibility of the Ministry of Economic Development.

[43] http://publication.pravo.gov.ru/Document/View/0001201911260012; http://publication.pravo.gov.ru/Document/View/0001201911260029?index=0&rangeSize=1

[44] http://publication.pravo.gov.ru/Document/View/0001201903180027

[45] http://sk.ru/foundation/legal/p/11.aspx;https://economy.gov.ru/material/directions/gosudarstvennoe_upravlenie/normativnoe_regulirovanie_cifrovoy_sredy/regulirovanie_primeneniya_tehnologiy_iskusstvennogo_intellekta/

- A suite of standards under the responsibility of Russia's Federal Technic Regulation and Metrology Agency. These include:
    - A forthcoming state standard on "Means of Monitoring Human Behaviour and Predicting People's Intentions. Terms and Definitions (GOST R 58776-2019), which aims to enable the development of intelligent systems which predict human behavior.
    - A forthcoming state standard on "Air Transport. Airports. Technical Means of Inspections. Methodology for Determining Indicators of the Quality of Recognition of Illegal Items on X-Ray Images (GOST R 58777-2019), which aims to establish uniform requirements for the systems and algorithms of recognition of illegal items in baggage and hand luggage on X-ray images, and increase reliability of their test results.
    - Development of six state standards for "AI-Systems for Clinical Medicine", to enable the deployment of AI systems in healthcare.
- Russia's establishment in 2019 of a Technical Committee on Standardisation "Artificial Intelligence" (TC164)[46], on the order of the Federal Agency on Technical Regulating and Metrology (Rosstandart) to increase the efficiency of development of the regulatory and technical base in the field of AI. Its objectives are to develop standards for AI, ensure the quality of AI systems, develop AI technology applications and implement AI technologies in education.
- Switzerland's general technology-neutral legislative and regulatory approach, which largely refrains from promoting specific technologies and tries to avoid technology-specific regulations wherever possible. This openness towards new technologies aims to allow the potential of new ideas and innovations to be optimally exploited.[47] This also includes the principles of legal certainty and efficient regulation.

[46] https://oecd.ai/dashboards/policy-initiatives/2019%2Fdata%2FpolicyInitiatives%2F24906

[47] https://www.seco.admin.ch/dam/seco/en/dokumente/Wirtschaft/Wirtschaftspolitik/digitalisierung/Executive_Summary.pdf.download.pdf/Executive_Summary_en.pdf

9. Building human capacity and preparing for labour market transformation

a) Governments should work closely with stakeholders to prepare for the transformation of the world of work and of society. They should empower people to effectively use and interact with AI systems across the breadth of applications, including by equipping them with the necessary skills.

b) Governments should take steps, including through social dialogue, to ensure a fair transition for workers as AI is deployed, such as through training programmes along the working life, support for those affected by displacement, and access to new opportunities in the labour market.

c) Governments should also work closely with stakeholders to promote the responsible use of AI at work, to enhance the safety of workers and the quality of jobs, to foster entrepreneurship and productivity, and aim to ensure that the benefits of AI are broadly and fairly shared.

Explanation and rationale:

AI is broadly expected to change the nature of many aspects of life, as it diffuses across sectors. This is particularly true in the context of labour, employment and the workplace – where AI will complement humans in some tasks, replace them in others and generate new types of jobs and work organisation. If poorly managed, these labour market transformations could have significant economic and social costs. In managing these transitions and ensuring they are fair, policy makers – together with stakeholders such as social partners, employer organisations and trade unions – will need to consider questions around social protection, education programmes, skills development, labour market regulation, public employment services, industrial policy, and taxation – as well as the financing of transitions.

Managing fair transitions requires policies for life-long learning, skills development and training that would allow people, and workers (in different contractual contexts) in particular, to interact with AI systems, adapt to AI-generated changes and access new opportunities in the labour market. This includes the skills required of AI practitioners (which are currently in shortage) and those needed for other workers (such as doctors or lawyers) to be able to leverage AI in their areas of expertise, so that AI augments human capabilities. In parallel, skills development policies will need to focus on the distinctly human aspects necessary to complement AI systems, such as judgement, creative and critical thinking and interpersonal communication.

Additionally, AI-led changes to the labour market may require adapting or adopting labour standards and agreements between management and workers, where applicable, to reflect these changes, address possible challenges to equality, diversity and fairness (posed, for example, by data collection and processing), and to reinforce reliable, safe and productive workplaces. This may be achieved through a combination of regulation, social dialogue and collective bargaining. It is important to allow for flexibility at the workplace while safeguarding workers' autonomy and job quality.

Considerations for governments:

The labour market consequences of AI have been perhaps the most visible and most feared amongst people to date. Building human capacity and preparing for labour market transformation is therefore one

of the most important challenges governments face in order to advance and harness trustworthy AI. It is also one of the most important opportunities. Digital-intensive activities have provided many new jobs in the past decade, and while the net impact of AI on work remains uncertain, there is considerable scope for changing the nature of work towards more creative and flexible tasks. At the same time, depending on their demographic profile, G20 economies may face significant challenges in preparing their youth populations or reorienting their middle-aged populations towards a future where AI-specific and generic skills increase in importance. They must also ensure that women play their full role in developing, using and harnessing AI, so that the digital gender divide is not widened.

Examples of initiatives:

Argentina – EDUCAR

Argentina is taking action to build AI-relevant human capacities from the earliest stages in education, targeting children from ages 6 to 18. It highlights the importance of building skills that can enable countries to shape AI in a trustworthy way.

Through its programme EDUCAR (running since 2018), the Ministry of Education incentivises digital literacy through Internet connection in schools, the delivery of digital tools, the development of software and virtual training platforms and the creation of spaces of technological innovation. Together with Priority Learning Hubs (NAPs in Spanish) and the National Plan Learning Connected, Argentina is aiming to prepare young generations with "future skills", including AI-related capabilities such as statistical thinking, mathematics, logic, computer sciences, programming, computational thinking, and comprehension of AI implications (e.g. relationships between people and technologies, privacy issues, use of data, critical thinking). It also aims to spur development of skills that differentiate people from machines, including advanced cognitive skills and socio-emotional skills (e.g. thinking capacity, reflection, making conclusions, analysis, empathy with others, curiosity, teamwork and creativity). Other initiatives are outlined in Argentina's National Plan for AI (Plan Nacional de Inteligencia Artificial).

Canada – Pan-Canadian AI Strategy

Launched in 2017 and implemented in partnership with established AI institutes, this initiative seeks to increase the number of outstanding AI researchers and skilled graduates in Canada and support a national research community on AI, thereby contributing to scientific excellence, economic growth and enhanced consideration of the societal and political impacts of AI.

Through a funding contribution of CAN 125 million over 5 years to the Canadian Institute for Advanced Research (CIFAR), Canada has established a Pan-Canadian AI Strategy[48] in partnership with Vector Institute (Toronto), the Montreal Institute for Learning Algorithms – Mila (Montreal), and the Alberta Machine Intelligence Institute – Amii (Edmonton). The Provincial governments of Ontario, Alberta, and Quebec have collectively committed an additional CAD 280 million over 5 years to supplement the Federal funding. The objectives of the initiative are:

- To increase the number of outstanding AI researchers and skilled graduates in Canada.

[48] CIFAR Pan-Canadian Artificial Intelligence Strategy

- To establish interconnected nodes of scientific excellence in Canada's three major centres for AI in Edmonton, Montréal and Toronto.
- To develop global thought leadership on the economic, ethical, policy and legal implications of advances in AI; and,
- To support a national research community on AI.

The benefits of the initiative are diverse, including:

- Researchers and students in Canada benefit from increased access to funding.
- Canadian universities benefit from better incentives to retain and attract top AI researchers.
- Students in Canada benefit from inclusive, pro-social educational initiatives and opportunities.
- Canadian firms benefit from an ever-stronger talent base to draw from.
- The Canadian government benefits through CIFAR's expert-led support for existing policy initiatives, which inform the development of ethical, inclusive, and sustainable regulatory measures for AI.
- Canadian citizens and residents are benefitting from the economic growth catalysed by the strategy as well as enhanced consideration of the societal and political impacts of AI though CIFAR's support for policy initiatives and AI & Society programming.
- Talent abroad benefits from growth in the Canadian labour market and increased opportunities to study in Canada.
- Canada's international partners will benefit from the research and policy innovations and talent produced through the strategy.

In their 2019 report, CIFAR notes that Canada saw a 27% increase in new Canadian start-ups and a 51% increase in venture capital funding for Canadian AI firms year over year in 2017-18. CIFAR also reported that its AI Research Chairs program retained 24 Canadian researchers and attracted 22 world-class researchers to Canada in 2017-2018.

Germany – Research on the impact of AI on work and employees

This research project between IBM and the German labour union ver.di looks at the effect of AI on work and employees. By helping to understand labour market transformation through the use of AI, the research aims to provide insights to help develop guidelines for the use of AI in the world of work.

Supported by the German Federal Ministry of Labour and Social Affairs, this research[49] examines the use of the Watson AI program designed by IBM in companies and organisations. Central questions included:

- What effects does the use of AI in the workplace have on employees and work?
- How does AI affect the quality of work and experience of employees?

[49] https://www.ibm.com/de-de/blogs/think/2019/09/17/watson-ki-studie/ (in German only).

- What kind of new jobs are created, what kind of jobs will possibly be replaced by AI, and how do competence profiles of existing jobs change?
- What opportunities, perspectives and necessities arise from this?

As well as increasing understanding of labour market transformation, the project seeks to create a human-centred design of the digitalised working world and discuss guidelines. It particularly aims to benefit businesses (especially SMEs), labour unions, employees and researchers, and the results fertilise the newly launched German AI Observatory and other activities of the Federal Ministry. First results of using AI in the form of a chatbot at a big company show that so far, there has been no substitution of human work by the use of AI. Employees rated the use of the program as very positive in terms of work quality and job satisfaction (new tasks were perceived as interesting and demanding) and viewed it as a useful assistance system. Two further case studies will investigate the effects of the chatbot on job and job profiles (including new jobs), and to what extent training and up- or re-skilling of employees have taken place successfully.

Japan – Boosting skills for young people

Japan has launched two new initiatives to strengthen skills for AI amongst its youth population. Aiming to improve the ICT environment in high schools and to promote acquisition of mathematics, data science and AI skills, Japan is making cross-government efforts to boost human capacity and prepare for labour market transformation.

In 2020, Japan introduced the concept of GIGA (Global and Innovation Gateway for All) Schools. This initiative aims to improve the ICT environment in elementary and junior high schools, particularly through developing terminals for each student and maintaining high-speed large-capacity communication networks. Through this commitment of financial support and infrastructure development, GIGA Schools contributes to learning that fosters creativity as well as individualised learning, ensuring inclusion of the diversity of children, including those who require special support. This initiative is overseen by the Cabinet Secretariat, Ministry of Education, Culture, Sports, Science and Technology, Ministry of Economy, Trade and Industry, and Ministry of Internal Affairs and Communication.

Also in 2020, Japan introduced a certification system for mathematical science, data science and AI education programmes in universities and technical colleges. The objective is for all colleges and colleges of technology graduates (approximately 500 000 graduates / year), regardless of majors, to acquire elementary level mathematics, data science and AI skills. This initiative, based on standards and reference models, enables the government to certify outstanding educational programmes recognised as graduation credits for these establishments. It is overseen by the Cabinet Office, Ministry of Education, Culture, Sports, Science and Technology, and Ministry of Economy, Trade and Industry.

United Kingdom – Centres for Doctoral Training

The United Kingdom's Engineering and Physical Sciences Research Council has introduced 16 new centres for doctoral training focused on AI, with the aim of increasing capacity in this relatively new discipline.

Centres for Doctoral Training in the United Kingdom provide training for cohorts of PhD students within focused research areas, often defined strategically by the Research Council funder(s). The model enables students to learn from each other's experience and the funding body to provide training to support the

student's overall development. The Centres aim to manage and assist PhDs funded by the Research Councils using a cohort-based approach and with the involvement of UK universities. Typical UK PhD students take three years to complete their doctoral research (or four years in cohort training programmes) under the guidance of an academic supervisor or small supervisory team, and tend to be located within an existing research group. Initially, CDTs were regarded as a strategic mechanism for increasing capacity in interdisciplinary research activities such as the life sciences interface and complexity science, areas that were difficult to locate within a traditional University's departmental organisation. CDTs are being extended to many new disciplines in four year doctorate level cohort based training via a critical mass of supervisors.

10. International co-operation for trustworthy AI

a) Governments, including developing countries and with stakeholders, should actively co-operate to advance these principles and to progress on responsible stewardship of trustworthy AI.

b) Governments should work together in the OECD and other global and regional fora to foster the sharing of AI knowledge, as appropriate. They should encourage international, cross-sectoral and open multi-stakeholder initiatives to garner long-term expertise on AI.

c) Governments should promote the development of multi-stakeholder, consensus-driven global technical standards for interoperable and trustworthy AI.

d) Governments should also encourage the development, and their own use, of internationally comparable metrics to measure AI research, development and deployment, and gather the evidence base to assess progress in the implementation of these principles.

Explanation and rationale:

This recommendation calls for international co-operation among governments and with stakeholders, to address the global opportunities and challenges of AI. Such co-operation includes advancing the implementation and dissemination of these principles and policies across countries, including developing countries and least developed countries, and with stakeholders. It is in this exact spirit that the 2020 G20 DETF discussions on AI, including the sector-specific dialogue on AI, are taking place.

International co-operation can leverage multiple international and regional fora to share AI knowledge in order to build long-term expertise on AI; to develop technical standards for interoperable and trustworthy AI; and to develop, disseminate and use metrics to assess the performance of AI systems, such as accuracy, efficiency, advancement of societal goals, fairness and robustness. It can also contribute to transborder flows of data with trust, that safeguard security, privacy, intellectual property, human rights and democratic values and that are key to AI innovation.

Considerations for governments:

The rise of AI is affecting all G20 economies, with shared challenges, the risk of deepening divides, and significant scope for mutual learning. In some instances, only collective action and dialogue can bring progress. Implementing international co-operation for trustworthy AI can help ensure that this promising technology is steered towards the greatest global good, with appropriate safeguards to ensure that the risks are managed and minimised, and that trustworthy AI becomes a reality.

Examples of initiatives:

Canada and France – Global Partnership on AI

Realising the full potential of AI that benefits all citizens requires international collaboration and coordination. Canada and France are working with the international community to create the Global Partnership on Artificial Intelligence (GPAI) to support and guide the responsible development of artificial intelligence that is grounded in human rights, inclusion, diversity, innovation, and economic growth. This initiative seeks to establish a global reference point on AI, fostering international collaboration and coordination on AI policy development among like-minded partners. Its work will contribute to implementing the full range of values-based G20 AI Principles.

GPAI is expected to launch in early 2020, with the first GPAI Multistakeholder Experts Group Plenary to take place in Canada in fall 2020. GPAI will be a long-term organisation that is expertise-based and dedicated to AI, working on identified topics. It will be global and inclusive of emerging and developing countries, convening experts from a wide range of sectors and with membership informed by shared values set out in a GPAI Declaration whose elements are strongly aligned with the G20 AI Principles. Its objectives are multiple, including to:

- Promote and protect a human-centric and ethical approach to AI, grounded in human rights
- Support a multi-stakeholder approach to AI
- Stimulate innovation, growth and well-being through AI
- Align efforts on AI with the principles of sustainable development and the achievement of the 2030 Agenda for Sustainable Development
- Strengthen diversity and inclusion through AI
- Foster transparency and openness of AI systems
- Foster trust and accountability in AI
- Promote and protect democratic values, processes and institutions
- Bridge digital divides
- Promote international scientific collaboration on AI.

In September 2019, Canada's Minister Bains announced up to CAN 10 million over 5 years to establish an International Centre of Expertise in Montreal, in collaboration with the Government of Quebec, which is contributing CAN 5 million.[50]

[50] Press Release: Centre for Expertise

European Union – Co-ordinated Plan on AI

To ensure the success of the European Union's AI Strategy, EU Member states and the European Commission are collaborating in a Co-ordinated Plan on AI – "Made in Europe". At its core, it aims to encourage synergies and co-operation across the EU, in a clear example of international co-operation for trustworthy AI.

In December 2018, the European Commission delivered a Communication to the European Parliament and other bodies, setting out a Co-ordinated Plan on AI accompanied by a series of actions for the 2019-2020 period.[51] The Plan provides a strategic framework for national AI strategies (including their investment and implementation measures), with the aim of fostering development of trusted AI that corresponds to European ethical values and citizens' aspirations. The Plan sets out a set of actions at EU, national and regional level to:

- Boost investment and reinforce excellence in AI technologies which are trustworthy and "ethical and secure by design";
- Develop shared agendas for industry-academia collaborative R&D and innovation;
- Adapt learning and skilling programmes and systems to prepare Europe's society for AI;
- Build essential capacities in Europe underpinning AI, such as data spaces;
- Make public administrations front-runners in using AI;
- Implementing ethical guidelines for the development of AI with a view to setting global ethical standards;
- Reviewing the national and European legal framework where needed, to adapt to specific challenges.

In addition to European co-operation, the Plan also calls for EU Member states to pool their efforts for a responsible development of AI at the global level.

France – Joint French-German AI Network

As part of efforts to strengthen general cooperation and integration, France and Germany have developed a roadmap for a research and innovation network on AI, with objectives including the creation of a common AI ecosystem. As well as exemplifying international cooperation for trustworthy AI, this nascent initiative also involves investment in AI R&D.

On 22 January 2019 France and Germany signed the Aachen Treaty to strengthen French-German cooperation and integration. The Treaty includes cooperation in the field of AI, with one priority project being the establishment of a 'Research and Innovation Network on Artificial Intelligence' (also called 'Joint AI Network').[52] As objectives for the Joint AI Network, which commenced in October 2019, the French and German Governments will strive to strengthen ties between existing structures, create a common AI ecosystem to bring about new cooperation projects, and share common positions on AI policy at EU level and promote coherent European action at the international level.

[51] https://ec.europa.eu/digital-single-market/en/news/coordinated-plan-artificial-intelligence

[52] https://www.entreprises.gouv.fr/files/files/directions_services/numerique/grands-dossiers/intelligence-artificielle/Feuille_de_route_franco-allemandepourIntelligenceartificielle-Toulouse16102019_EN.pdf

To achieve these objectives, the governments will:

- initiate, support and promote common events and projects in order to foster dialogue and cooperation between research institutions and industrial partners in the field of AI,
- establish sectoral focus groups to intensify concrete AI cooperation projects, including a possible bilateral funding scheme and a sector-specific data sharing initiative, and
- provide joint input and common recommendations for the EU's initiatives in the field of AI.

Other examples

Other examples of international co-operation for trustworthy AI include:

- Brazil's hosting of the UNESCO Regional Forum on AI in Latin America and the Caribbean (São Paulo, 12-13 December 2019).
- The Freedom Online Coalition (FOC), which promotes human rights online, bringing together over 30 countries (including several G20 countries), and a multi-stakeholder Advisory Network that encompasses academia, civil society, and the tech industry. In 2020 the FOC will focus on the global governance of AI, among other files, with the launch of a Taskforce on AI and Human Rights (T-FAIR). Chaired by Canada, T-FAIR will operate from April 2020 until late 2021 and provide a hub for FOC member countries and the Advisory Network to promote human rights-respecting AI by sharing and disseminating information and collaborating on joint initiatives. T-FAIR will bring together a network of diverse stakeholders representing countries, academia, civil society, and industry and through monthly meetings and learning calls on AI, will increase the capacity of members to collaborate on relevant policy issues and strategise towards shaping international norms on human rights-respecting AI.
- The ITU's "AI For Good Global Summit", held annually since 2017 and which supports the ITU's efforts to identify practical applications of AI to accelerate progress towards the SDGs. The 2018 and 2019 events spurred the launch of several projects, including an "AI for Health" focus group led by the ITU and WHO, an "AI for Autonomous and Assisted Driving" focus group led by the ITU, and an open framework for collaboration in the "AI Commons". The 2020 event will focus on connecting AI innovators with "problem owners" to solve global challenges.
- The OECD's AI Policy Observatory (OECD.AI) launched on 27 February 2020, which aims to help countries enable, nurture and monitor the responsible development of trustworthy AI systems for the benefit of society. It combines resources from across the OECD with those of stakeholder groups and partners to provide a comprehensive database of AI policies from around the world, present multidisciplinary evidence-based policy analysis, and facilitate multistakeholder dialogue. Many G20 countries have contributed to the development of the Observatory and all can benefit from the resources available on the platform.
- Switzerland's active involvement in relevant international organisations and processes, such as the Council of Europe, the ITU's "AI for Good Summit" and UNESCO.[53] Switzerland was also a member of the OECD expert group on AI that developed the OECD principles on AI adopted in 2019. Switzerland is also supporting the implementation of recommendation 3c on AI by the UN

[53] https://www.bakom.admin.ch/bakom/en/homepage/ofcom/international-activities.html

SG's High-Level Panel on Digital Cooperation. As a small, highly developed and networked country, Switzerland considers it essential to actively shape the debate on the global governance of AI. Particularly important for Switzerland is to ensure that fundamental and established values and norms such as human rights are respected and that all relevant stakeholders are involved in decision-making.

- UNESCO's efforts to develop a global standard-setting instrument on the ethics of AI, which will leverage its expertise on ethics in science and technology and bioethics and will result in a normative instrument raising awareness of the ethical impact of AI on social, cultural and scientific aspects of society. This initiative follows the decision of UNESCO's General Conference at its 40th session in November 2019, and builds on UNESCO's broader activities on AI, including a Forum on AI in Africa (Morocco, December 2018), an International Conference on AI and Education (Beijing, May 2019), and a Regional Forum on AI in Latin America and the Caribbean (Sao Paulo, December 2019).
- The UN's High Level Panel on Digital Cooperation, which in July 2019 published its report on how the international community can work together to optimise the use of digital technologies and mitigate the risks. A High Level Panel Follow-up Roundtable on AI has been established to take forward the Panel's recommendation (3C) to enhance digital cooperation to think through the design and application of standards and principles such as transparency and non-bias in AI systems in different settings.

Box 2. Spotlight on trustworthy AI in health

The potential of AI in health is profound, given the growing volume of electronic data as well as the inherent complexity of the sector, its reliance on information to solve problems, and the variability and complexity of how disease interacts with individuals and populations. But the risks of unintended and negative consequences associated with AI are commensurately high, especially at scale. As part of the G20 Dialogue on AI in 2020, G20 members and guest countries were invited to provide examples of actions and initiatives they are taking to improve trustworthy AI in health. The following examples were collected:

- **Argentina:** For some years now, the National Council for Scientific and Technical Research (CONICET) has included AI as well as reliability and computer security as strategic issues within their calls to enter the Scientific Researcher Career. This policy seeks to encourage the formation of resources highly skilled people in such areas.
- **Australia:** The Therapeutic Goods Administration (TGA) is a member of the Standards Australia Artificial Intelligence Committee that is developing AI standards as part of an International Organization for Standardization (ISO) standards-development effort. In addition, the National Health and Medical Research Council (NHMRC) has a remit to consider areas of ethics in medicine and research, through its Australian Health Ethics Committee. This work could include areas of emerging technology such as AI. The Australian Government also, in 2019, announced an Applied Artificial Intelligence (AI) Research in Health (AAIRH) Grant Opportunity, through the Medical Research Future Fund. The AAIRH provides funding to health and medical researchers to transform the future of health care using AI ideas, technologies and methodologies. The objectives of the AAIRH are to:
  - Translate or implement innovative AI technologies into health applications that benefit multiple health disciplines/areas;
  - Involve consumers in the research journey to ensure the research is applicable to the needs of the Australian community;
  - Increase AI workforce capacity and capability, particularly in relation to health, through cross-sector and interdisciplinary collaboration.

  The intended outcome of the AAIRH is to promote the application of novel AI technologies and methodologies to cross-sectoral and interdisciplinary health research that will transform healthcare and outcomes through improved preventive, diagnostic and treatment approaches.
- **Brazil:** Brazil is in the process of establishing regulation in the area of privacy and personal data protection in health systems, consistent with existing legislation (Personal Data Protection Law – LGPD, Law n. 13.709/2018). To this end, the country is currently developing a national electronic health records system, which aims to provide a robust database for current medical use, as well as for technology development and innovation.
- **China:** The National Medical Products Administration, together with 14 institutions including the China Academy of Information and Communications Technology, established the Artificial Intelligent Medical Device Innovation and Cooperation Platform (http://www.aimd.org.cn). The Platform aims to build an open, collaborative innovation system for AI medical devices, and to facilitate the regulation, technological innovation and product translation of AI medical devices.
- **Germany:** The German Ministry of Health is financing several pilots and projects in order to showcase the benefits of AI in health. The projects also tackle questions of capacity building in the areas of digital skills of the workforce, digital infrastructure, and conditions to develop AI

applications in R&D. It is important that the particular conditions in the health sector are taken into account from the earliest stages of development. Based on these projects, there is a constant analysis in order to assess the current regulatory framework. In this context, the Ministry holds a constant dialogue with stakeholders of the system in order to assess further steps for AI in health. Examples of work in co-operation with developing countries include:

  - The openIMIS initiative, operated by the "Deutsche Gesellschaft für Internationale Zusammenarbeit" (GIZ). This initiative supports the development of open source software for managing health financing mechanisms, including health insurance. openIMIS supports the management of beneficiary data and allows health facilities to digitally submit claims for services provided. openIMIS allows the financing scheme operator to digitally review claims and reimburse health facilities, ensuring efficient and reliable financial flows in the health system. The software is used in five countries (Nepal, Tanzania, Cameroon, DR Congo, Chad) and reaches more than 3 million people. In the near future applications of AI could be envisaged in this project, e.g. with increases in the submission of claims from health facilities, additional checks for securing the medical validity of the claims (currently undertaken by highly qualified medical professionals) could be facilitated through the use of AI.
  - The Artificial Intelligence for All - FAIR Forward initiative, partnering with countries in Africa and Asia, focuses on promoting AI innovations and open language. At the same time, the initiative also aims at supporting partner governments in the area of data protection and regulation as well as to support local technical know-how.

- **Germany**: The "Plattform Lernende Systeme", the German Federal Ministry of Education and Research's Platform for AI, brings together leading experts in self-learning systems and AI from science, industry, politics and civic organisations. In specialised focus groups, they discuss the opportunities, challenges and parameters for developing self-learning systems and using them responsibly. One of these groups focuses on "Health Care, Medical Technology, Care". It derives scenarios and recommendations for the responsible use of AI in the health sector.
- **Indonesia**: Indonesia considers the availability of an integrated trustworthy health data system as a key challenge for operationalising trustworthy AI in the health sector. As well as the forthcoming National Strategy on AI, which will affect the actions of all Ministries, the Indonesian government is facilitating the development of public cloud services that will provide AI services for the wider public and stakeholders. The services will also provide shared infrastructures and platforms through which digital companies can distribute metadata, data examples, computing and learning services that are free to use by AI developers. Indonesia is also fostering a quadruple helix collaboration in AI research and innovation initiatives.
- **Italy:** In the healthcare domain, Italy has an increasing number of applications and AI technologies, leveraging the increasing amount of data coming from the research sector, hospital medical records, reports and laboratory tests. Recent relevant initiatives supported by the Italian government include:
  - Collation of a list of active projects in the domain of AI and health, by the Italian National Research Council (CNR) and the Laboratory for AI and Autonomous Systems (AIIS Lab) of the National Inter-university Consortium for Informatics. Structured initiatives for human-centric trustworthy AI in health are being pursued within the EU ICT-48 network of centers of excellent AI research "Humane-AI-Net" and "TAILOR - Foundations of Trustworthy AI" and "ELISE", in collaboration with specific national and European projects, such as the ERC Grant "XAI – Science and technology for the explanation of AI decision-making". Italian Labs participate in the DeepHealth EU project for a federated infrastructure of deep learning for medical imaging.

- Launch of a National AI Doctoral Programme, by Italy's Ministry for University and Research, which aims at recruiting around 200 doctoral candidates all over the country.
- The "Rome Call for AI Ethics" aimed at increasing awareness of the role of ethics in AI. The document was signed in February 2020 by the Pontificia Accademia per la Vita, Microsoft, IBM, FAO and the Italian Government and proposes a more human-centric approach to AI.
- Signature of an MoU between the Minister of Technological Innovation and Digitization and Fondazione Leonardo to shape the framework and boundaries for AI adoption in Public Administration.
- Exploration of a specific platform to improve the level of citizen education on AI matters, with a view to fostering idea generation for future adoption and ensuring a better understanding of trustworthiness on use cases where AI is used.

- **Korea:** The Korean government made amendments to the country's three main data privacy laws, including the Personal Information Protection Act, to improve personal data protection while encouraging the development of data-related industry through the promotion of data use. In addition, the government has plans to develop AI Ethics Principles in order for all sectors including AI in health to observe and implement trustworthy AI principles including transparency, safety and accountability, as one of the follow-up measures to the National Strategy for Artificial Intelligence released in December 2019.
- **Saudi Arabia:** At the level of governance and facilitation, the creation of the Saudi Data and Artificial Intelligence Authority contributes to trustworthy AI in health, notably through including a healthcare focus within its activities to set the national data and AI strategy and oversee its execution through harmonised data policies, data and insights capabilities and continuous AI and data innovation. Saudi Arabia is also working on data quality, and since 2015 has focused on promoting and exchanging health care data across the country with the development of the Health Information Exchange Policy and standards, based on international standards refined by collaborative work between more than 18 Saudi health and non-health organisations. This included the development of 9 Health Information Exchange policies (e.g. security, secondary data usage) and the development of 15 health information exchange specifications (including patient and provider identification, laboratory tests, imaging, medications, discharge summaries and referral) along with associated terminologies and value lists and the establishment of a testing environment for the first use cases. Finally, all clinical AI used within the Saudi Ministry of Health requires extensive clinical testing and signoff by an agreed implementation process prior to adoption.
- **Spain:** Health is a high priority objective in the National AI Strategy and several actions and pilot projects are included that address specific aspects of AI in Health. The strategy is now completed though its publication has been postponed due to the Covid-19 crisis. In addition, the Plan for the Advancement of Language Technologies aims to develop the natural language processing industry, machine translation and conversational systems in Spain, and especially in Spanish and co-official languages. Some of these methods are being successfully applied to the analysis of healthcare data (both text, such as clinical notes, and structured data from electronic health records).
- **Switzerland:** Challenges related to AI and health issues such as data protection, the advantages/disadvantages of using AI, and validation of the effectiveness and liability of new AI methods, are likely to increasingly occupy the Swiss health system and health policy community. In this context, the responsible authorities in Switzerland will monitor the impact of AI on medicine and health care and, if necessary, submit proposals for adapting the relevant federal legal bases. Several research institutions are already active in various fields, e.g. in the context of SPHN (Swiss Personalized Health Network).

- **Turkey:** Efforts are underway to implement an e-Triage platform ("What I Have") that operates within the Central Physician Appointment System and works with the Machine Learning (ML) method. This platform will ask the patient about their discomfort before the appointment, and screen and present information to the patient according to the answers given. In addition to this study, infrastructure development studies have been initiated for platforms that will enable the processing of radiology images with AI, and that will enable the processing of Electronic Health Records with AI.
- **United Arab Emirates:** Dubai Health Authority has developed AI use cases and taken into production AI systems using the Smart Dubai/IBM AI Lab. For proofs of concept like the Patient Deterioration Detection Tool, developers are required to assess the system against guidelines to establish the decision classification of the system as well as its overall performance against fairness, accountability, trust and explainability.
- **ITU and WHO** jointly established a Focus Group on Artificial Intelligence for Health (FG-AI4H) (https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx) in June 2018, which aims to build a benchmarking database and assessment system for assessing and supervising healthcare AI products.

Note: For further reading on policy issues arising from the use of AI in the health sector, see OECD (2020), *Trustworthy AI in Health*, a background paper prepared for the G20 Saudi Presidency for the G20 Dialogue on AI.

2 Observations from existing policy approaches

The compilation of strategy and policy examples detailed in the previous chapter highlight that much activity and experimentation is taking place in G20 countries to build and support trustworthy AI ecosystems. Drawing on the previous chapter, a stylised (non-exhaustive) mapping of key AI policies against the G20 AI Principles is contained in Table A below.

With the caveat that the strategies and policies described are only a sample of the full range of countries' policy activity on AI, a number of broad observations can nevertheless be noted from the examples provided by G20 and guest countries:

- Most strategies and policies aimed at or having the effect of advancing the G20 AI Principles are very recent or in the process of being developed. Very few policies have been operating for long enough to conduct evaluations. This suggests that there is significant scope for sharing experiences amongst peers, including through multi-stakeholder dialogue, to facilitate learning from good (and less effective) approaches. It also underscores the importance of building in evaluation or review mechanisms to ensure that governments reflect in due course on the efficacy and efficiency of their policy and strategy choices, in terms of impact on desired goals and on achievement of the core values-based G20 AI Principles.
- Related to this, several countries are actively leveraging information on use cases and practical examples that shed light on implementation at the firm level. At this early stage of experimentation with AI policies, this type of activity may be another important source of policy intelligence.
- Many strategies and policies address, either explicitly or implicitly, multiple G20 AI Principles at once. This makes it more difficult to categorise approaches, but is consistent with the intent of the Principles to be complementary and considered as a whole. It is also consistent with the Principles being mutually reinforcing – for instance, improving transparency may enable addressing important issues of fairness and bias. By taking the Principles as a package, G20 countries may make faster progress towards the overarching goal of advancing trustworthy AI for the benefit of economies and societies. At the same time, it appears that few policies place a primary focus on the Principles of robustness, security and safety, and accountability, compared to those of inclusive growth or human-centred values, raising the question of whether there is an opportunity to bring more emphasis to these important issues.
- Many policies and strategic approaches to achieving trustworthy AI leverage policy tools around R&D, fostering a digital ecosystem, shaping an enabling environment, building human capacity and supporting international cooperation for trustworthy AI. In other words, the policy recommendations noted by the G20 at the time of welcoming the G20 AI Principles are highly relevant to the achievement of trustworthy AI.
- A significant number of policies are oriented around R&D for AI, highlighting that countries consider that much more progress can be achieved with the technology itself, and its application to various economic and social questions. Countries are increasing public investment in R&D to complement the strong private investment taking place, offering potential for steering towards more socially

oriented applications and issues. The G20 AI Principles open the scope for possible regulatory action or legislation, but also provide complementary avenues for policy makers seeking to make progress on issues such as accountability, explainability, fairness and transparency. Investing in AI research can bring greater understanding of the challenges and is one potential source of new solutions. Similarly, investing in development of technical standards can provide critical input to policy makers by allowing more accurate and informative measurement of AI systems and their attributes.

Overall, the policy intelligence gathered in this exercise suggest that a mix of policies are needed when building trustworthy and human-centric AI – the goal of advancing the G20 AI Principles. As the G20 considers further actions toward advancing the G20 AI Principles, it will be valuable to reflect on policies around infrastructure, data access, the AI ecosystem, and human capacity.

Table A: Illustrative Actions taken by G20 and Guest countries to implement the G20 AI Principles

	Argentina	Australia	Brazil	Canada	China	European Union*	France	Germany
Principles for Responsible Stewardship of Trustworthy AI								
1. Inclusive growth, sustainable development and well-being	**National Plan for AI**	*Principles focus on human, social & environmental well-being*	**National AI Strategy**	*Focus of GPAI and Advisory Council*	**Governance Principles for the New Generation of AI**	**Ethics Guidelines on Artificial Intelligence**	**French AI Strategy: Economic Action Plan**	**AI Strategy**
2. Human-centred values and fairness	*Plan is human-centred*	**AI Ethics Principles**	*National AI Strategy includes regulation and ethical use of AI*	*Focus of GPAI and Advisory Council*	*Human friendly, fairness, justice and privacy included in Principles*	*Addressed in Ethics Guidelines on AI*	*Addressed by French AI Strategy and GPAI*	*Addressed in AI Strategy*
3. Transparency and explainability	*Plan seeks to align with ethical and legal principles*	*Included in AI Ethics Principles*	*National AI Strategy includes regulation and ethical use of AI*	**Directive on Automated Decision-Making**	*Safety and controllability included in Principles*	*Addressed in Ethics Guidelines on AI*	*Addressed by French AI Strategy and GPAI*	*R&D project on explainability & transparency*
4. Robustness, security and safety	*Plan seeks to align with ethical and legal principles*	*Included in AI Ethics Principles*	*National AI Strategy includes regulation and ethical use of AI*	*Addressed in Directive*	*Shared responsibility included in Principles*	*Addressed in Ethics Guidelines on AI*	*Addressed by French AI Strategy and GPAI*	*Addressed in AI Strategy*
5. Accountability	*Plan seeks to align with ethical and legal principles*	*Included in AI Ethics Principles*	*National AI Strategy includes regulation and ethical use of AI*	*Addressed in Directive*	*Agile governance included in Principles*	*Addressed in Ethics Guidelines on AI*	*Addressed by French AI Strategy and GPAI*	*Addressed in AI Strategy*
National Policies and International Cooperation for Trustworthy AI								
6. Investing in AI Research and Development	*Plan will foster R&D initiatives*	**AI Technology Roadmap**	**Applied Research Centres in AI**	*Included in Pan-Canadian AI Strategy*	*Three-Year Action Plan for Developing New Generation AI Industry*	*European Network for AI Excellence Centres*	*Addressed in French AI Strategy*	**R&D Project on Explainability and Transparency**
7. Fostering a Digital Ecosystem for AI	*Plan focused on conditions for development of AI*	**AI Standards Roadmap**	*Included in National AI Strategy*	*Included in Pan-Canadian AI Strategy*	*New Generation Artificial Intelligence Development Plan*	*Open Data Directive; EUROHPC Initiative*	*Addressed in French AI Strategy*	*Addressed in AI Strategy*
8. Fostering an Enabling Policy Environment for AI	*Plan focused on conditions for development of AI*	*AI Technology Roadmap aims to develop national AI capabilities*	*Included in National AI Strategy*	**Advisory Council on AI**	**Trustworthy AI Operational Guidelines**	**EU Strategy on AI**	*Addressed in French AI Strategy*	*Addressed in AI Strategy*
9. Building Human Capacity and Preparing for Labour Market Transformation	**EDUCAR**	*Skills development included in AI Technology Roadmap*	*Skills development included in National AI Strategy*	**Pan-Canadian AI Strategy**		*Policy and Investment Recommendations for Trustworthy AI*	*Addressed in French AI Strategy*	**Research on the Impact of AI on work and employees**
10. International Co-operation for Trustworthy AI	*Plain includes international cooperation*		*Included in National AI Strategy*	**Global Partnership on AI;**		**Coordinated Plan on Artificial Intelligence**	**Global Partnership on AI; French-German AI Network**	*Addressed in AI Strategy*

Note: The table shows key actions from G20 and guest countries, as submitted to the Presidency. Key implementation examples are highlighted in bold, other elements in Italics. Actions shown from the European Union, India and the Russian Federation draw on the OECD.AI Policy Observatory; no information was available for South Africa and Jordan.

	India*	Indonesia	Italy	Japan	Korea	Mexico	Russia*	Saudi Arabia
Principles for Responsible Stewardship of Trustworthy AI								
1. Inclusive growth, sustainable development and well-being	**National Strategy on AI - AI for All**	**Development of a National Strategy on AI**	**Development of a National AI Strategy**	**Human-Centric AI Principles and AI Strategy**	**National Strategy for AI**		**National AI Development Strategy**	
2. Human-centred values and fairness	*Addressed in AI Strategy*	*To be addressed in National AI Strategy*		*Addressed in AI Principles*	*Addressed in AI Strategy*		*Addressed in AI Strategy*	
3. Transparency and explainability	*Addressed in AI Strategy*			*Addressed in AI Principles*	*Addressed in AI Strategy*		*Addressed in AI Strategy*	*Addressed by National Data & Management Authority*
4. Robustness, security and safety	*Addressed in AI Strategy*	*To be addressed in National AI Strategy*	*To be addressed in National AI Strategy*	*Addressed in AI Principles*	*Addressed in AI Strategy*		*Addressed in AI Strategy*	*Addressed by National Data & Management Authority*
5. Accountability				*Addressed in AI Principles*	*Addressed in AI Strategy*		*Addressed in AI Strategy*	*Addressed by Saudi Data and AI Authority*
National Policies and International Cooperation for Trustworthy AI								
6. Investing in AI Research and Development	*Addressed in AI Strategy*	*To be addressed in National AI Strategy*	*To be addressed in National AI Strategy*	*Addressed in AI Strategy*	**AI R&D Strategy**	*Addressed in PROSOFT and Innovation*	*Addressed in AI Strategy*	**National Centre for AI**
7. Fostering a Digital Ecosystem for AI	*Addressed in AI Strategy*	*To be addressed in National AI Strategy*	*To be addressed in National AI Strategy*	*Addressed in AI Strategy*	*Data and AI-driven Economy Promotion Plan*	**PROSOFT and Innovation**	*Addressed in AI Strategy*	**Saudi Data and AI Authority**
8. Fostering an Enabling Policy Environment for AI	*Addressed in AI Strategy*	*To be addressed in National AI Strategy*	*To be addressed in National AI Strategy*	*Addressed in AI Strategy*	*Addressed in AI Strategy*		**National Strategy for AI**	*Addressed by National Centre for AI*
9. Building Human Capacity and Preparing for Labour Market Transformation	*Addressed in AI Strategy*	*To be addressed in National AI Strategy*	*To be addressed in National AI Strategy*	**Boosting Skills for Young People**	*Addressed in AI Strategy*	*Addressed by Centres for Industrial Innovation in AI*	*Addressed in AI Strategy*	*Addressed by National Centre for AI*
10. International Co-operation for Trustworthy AI		*To be addressed in National AI Strategy*			*Addressed in AI Strategy*			

Note: The table shows key actions from G20 and guest countries, as submitted to the Presidency. Key implementation examples are highlighted in bold, other elements in Italics. Actions shown from the European Union, India and the Russian Federation draw on the OECD.AI Policy Observatory; no information was available for South Africa and Jordan.

	South Africa	Turkey	United Kingdom	United States	Jordan	Singapore	Spain	Switzerland	United Arab Emirates (Dubai)
Principles for Responsible Stewardship of Trustworthy AI									
1. Inclusive growth, sustainable development and well-being		**National AI Strategy & AI Technology Roadmap**	**Response to Independent AI Review**	**Guidance for Regulation of AI Applications**		**Model AI Governance Framework**	**Development of a National AI Strategy**		**AI Principles and Guidelines for the Emirate of Dubai**
2. Human-centred values and fairness		**Development of a National AI Strategy**	*Addressed by Centre for Data Ethics & Innovation*	*Addressed in Guidance for Regulation of AI Applications*		*Addressed in model AI Governance Framework*	*To be addressed in National AI Strategy*	**Guidelines for AI-related measures by federal government**	*Addressed in AI Principles and Guidelines*
3. Transparency and explainability		*To be addressed in National AI Strategy*	*Addressed by Centre for Data Ethics & Innovation*	*Addressed in Guidance for Regulation of AI Applications*		*Addressed in model AI Governance Framework*	*To be addressed in National AI Strategy*	*To be addressed in guidelines*	*Addressed in AI Principles and Guidelines*
4. Robustness, security and safety		*To be addressed in National AI Strategy*	*Addressed by Centre for Data Ethics & Innovation*	*Addressed in Guidance for Regulation of AI Applications*		*Addressed in model AI Governance Framework*	*To be addressed in National AI Strategy*	*To be addressed in guidelines*	*Addressed in AI Principles and Guidelines*
5. Accountability		*To be addressed in National AI Strategy*	*Addressed by Centre for Data Ethics & Innovation*	*Addressed in Guidance for Regulation of AI Applications*		*Addressed in model AI Governance Framework*	*To be addressed in National AI Strategy*	*To be addressed in guidelines*	*Addressed in AI Principles and Guidelines*
National Policies and International Cooperation for Trustworthy AI									
6. Investing in AI Research and Development		**AI Technology Roadmap**	**AI UK Sector Deal**	**National AI R&D Strategic Plan & American AI Initiative**		**National AI Strategy (Ecosystem enabler 4)**	**Strategy on R&D and Innovation in AI**	*General frameworks: digital strategy, growth, research, innovation*	**Synthetic Data**
7. Fostering a Digital Ecosystem for AI		**AI Technology Roadmap**	**Centre for Data Ethics and Innovation**	**American AI Initiative Executive Order**		**National AI Strategy (Ecosystem enabler 1)**	**Plan for Advancement of Language Technologies**	*General frameworks: digital strategy, growth, research, innovation*	*Addressed in Synthetic Data*
8. Fostering an Enabling Policy Environment for AI		*To be addressed in National AI Strategy*	*Addressed in AI UK Sector Deal*	**Technical Standards plan for AI**		**National AI Strategy (Ecosystem enabler 4)**	*To be addressed in National AI Strategy*	**Technology-neutral legislative approach**	**Ethical AI Self-Assesment Tool**
9. Building Human Capacity and Preparing for Labour Market Transformation		*To be addressed in National AI Strategy*	**Centres for Doctoral Training**	**American AI Initiative Executive Order**		**National AI Strategy (Ecosystem enabler 2)**	*Addressed in Strategy on R&DI in AI*	*General frameworks: digital strategy, growth, research, innovation*	
10. International Co-operation for Trustworthy AI				**American AI Initiative Executive Order**		**National AI Strategy (Ecosystem enabler 5)**		**International cooperation for trustworthy AI**	

Note: The table shows key actions from G20 and guest countries, as submitted to the Presidency. Key implementation examples are highlighted in bold, other elements in Italics. Actions shown from the European Union, India and the Russian Federation draw on the OECD.AI Policy Observatory; no information was available for South Africa and Jordan.

Annex: The G20 AI Principles and key terms

The G20 AI Principles welcomed by G20 Leaders in 2019 aimed to create an enabling environment for human-centred AI that promotes innovation and investment. They include five principles for the responsible stewardship of trustworthy AI:

- Inclusive growth, sustainable development and well-being
- Human-centered values and fairness
- Transparency and explainability
- Robustness, security and safety
- Accountability

The G20 AI Principles also offer guidance for consideration by policy makers, with five (non-binding) recommendations on national policies and international cooperation for trustworthy AI:

- Investing in AI research and development
- Fostering a digital ecosystem for AI
- Shaping an enabling policy environment for AI
- Building human capacity and preparing for labour market transformation
- International cooperation for trustworthy AI

The G20 AI Principles are aimed at all stakeholders in AI systems and are intended to be applied throughout the entire AI lifecycle. For policymaking, and especially given that AI policy issues frequently transcend national borders, this wide scope of application makes it essential that policy makers have a common understanding of key terms and concepts. Such a common framing may also facilitate sharing of relevant experiences of national strategies and policies, and developing practical national approaches to implementing specific principles. To give one example, having a common understanding of the general component parts of an AI system can help policymakers identify where bias may occur, and therefore where actions related to transparency, robustness and accountability might be of especially high value.

This anex sets out definitions and concepts of AI, AI systems and AI system lifecycles – terms used in the G20 AI Principles – with the aim of supporting a common base for the G20's ongoing discussions on implementing the G20 AI Principles.

What is AI?

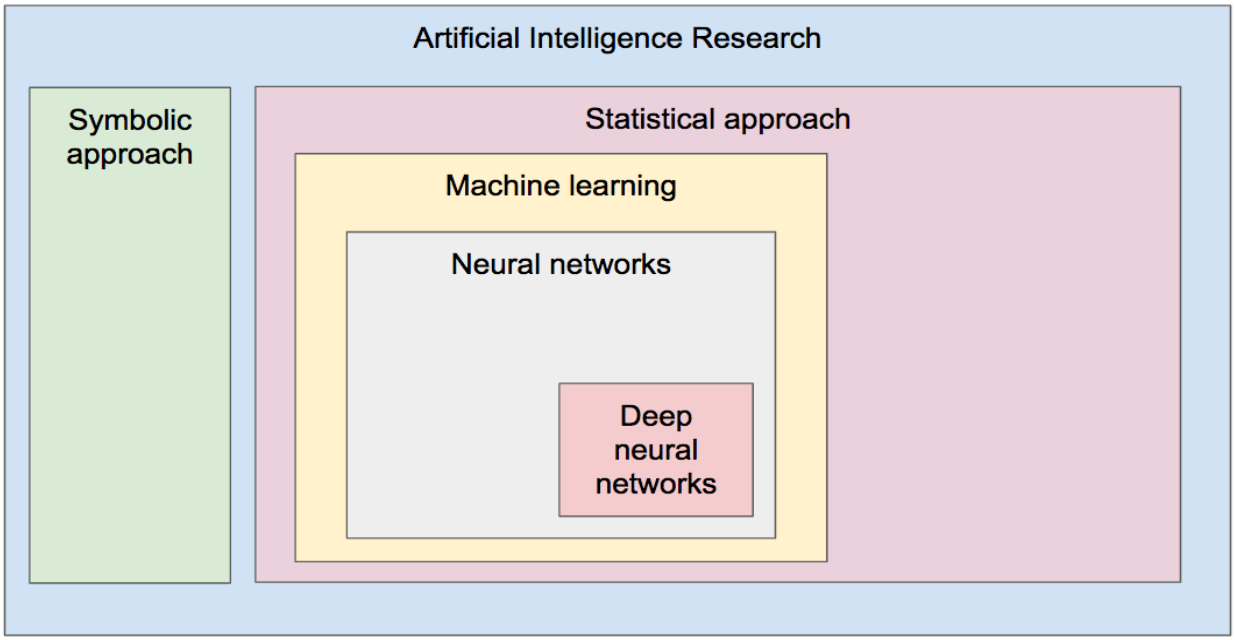
There is no agreed definition of AI, but there is a general understanding that it consists of simulating certain learning processes of human intelligence, to learn from it and replicate it. AI rests on two key elements: data and algorithms (and computing power to bring the two together). Today's AI is "narrow" and designed to accomplish a specific problem-solving or reasoning task. Even the most advanced AI systems available today, such as IBM's Watson or Google's AlphaGo, are still "narrow". This kind of AI is in contrast to a

(hypothetical) Artificial General Intelligence (AGI) in which autonomous machines would become capable of general intelligent action, like a human being, including generalising and abstracting learning across different cognitive functions.

Research has historically distinguished symbolic AI from statistical AI. Symbolic AI uses logical representations to deduce a conclusion from a set of constraints. It requires researchers to build detailed and human-understandable decision structures to translate real-world complexity and help machines arrive at human-like decisions. Symbolic AI is still in widespread use, e.g. for optimisation and planning tools. Statistical AI, where machines induce a trend from a set of patterns, has seen increasing uptake recently. A number of applications combine symbolic and statistical approaches. For example, natural language processing (NLP) algorithms often combine statistical approaches (that build on large amounts of data) and symbolic approaches (that consider issues such as grammar rules). Combining models built on data and human expertise is a promising path to help address the limitations of both approaches.

A subset of statistical AI is machine learning (ML) – a set of techniques to allow machines to learn in an automated manner through patterns and inferences rather than through explicit instructions from a human. ML approaches often teach machines to reach an outcome by showing them many examples of correct outcomes. However, they can also define a set of rules and let the machine learn by trial and error. ML can be used to build, adjust or interpret a model's results. It includes numerous techniques that have been long used by economists, researchers and technologists, such as regressions, decision trees and principle component analysis. However, "neural networks" – a sophisticated statistical modelling technique – is behind the current wave of ML applications, enabled by growing computational power and the availability of massive datasets ("big data"). Neural networks essentially modify their own code to find and optimise links between inputs and outputs. Deep learning is a phrase that refers to particularly large neural networks; there is no defined threshold as to when a neural net becomes "deep". A depiction of these concepts is presented in Figure 1 below.

Figure 1. The relationship between AI and ML

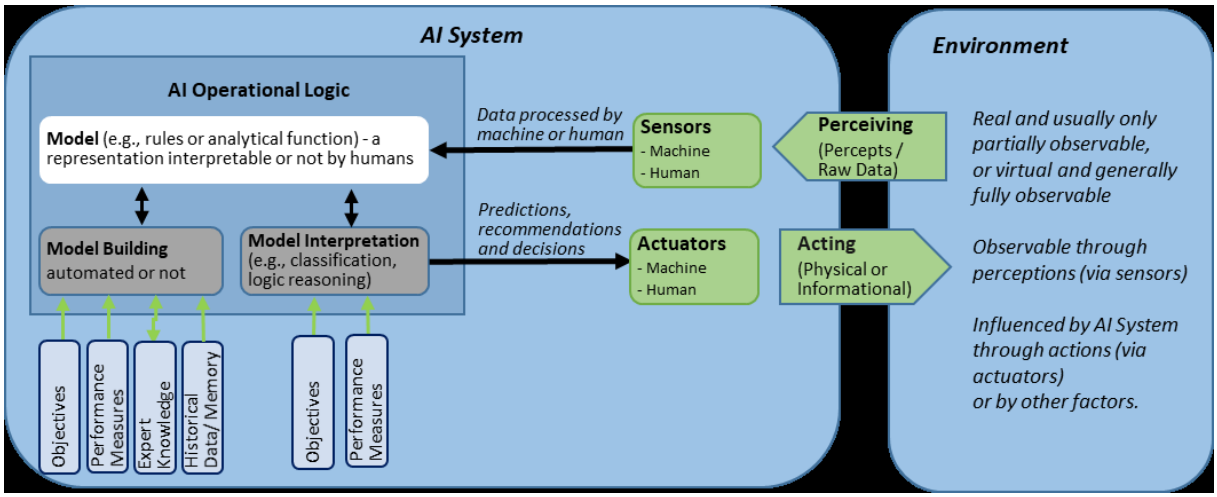


Source: Provided by the Massachusetts Institute of Technology (MIT)'s Internet Policy Research Initiative (IPRI), in OECD (2019a).

What is an AI system?

An AI system may be thought of as a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. An AI system consists of three main elements: sensors, operational logic and actuators. Sensors (which may be machines or people) collect raw data from the environment, while actuators (again, either machines or people) take actions to change the state of the environment. The key power of an AI system resides in its operational logic, which, for a given set of objectives and based on input data from sensors, provides output for the actuators – as recommendations, predictions or decisions – that are capable of influencing the state of the environment. This conceptual view of an AI system is depicted in Figure 2.

Figure 2. Conceptual view of an AI system



Source: OECD (2019b).

The environment may be real (e.g. existing physically, socially or mentally) and usually only partially observable, or it may be virtual (e.g. a board game) and generally fully observable. An AI system influences the environment by utilising machine and/or human-based inputs/data to: *i)* perceive real and/or virtual environments; *ii)* abstract such perceptions into models manually or automatically; and *iii)* use model interpretations to formulate options for outcomes.

Inside the operational logic, a model is an actionable representation of all or part of the external environment of an AI system. It represents the core of an AI system and can be based on data and/or expert knowledge, by humans and/or by automated tools like machine learning algorithms. Model interpretation is the process of deriving an outcome from a model, while objectives and performance measures guide the execution. In some cases (*e.g.*, deterministic rules), a model can offers a single recommendation, while in other cases (*e.g.*, probabilistic models), a model can offer a variety of recommendations associated with different levels of, for instance, performance measures like level of confidence, robustness or risk. In some cases, during the interpretation process, it is possible to explain why specific recommendations are made, while in other cases, explanation is almost impossible.

One practical illustration of an AI system is agricultural robots. Traditionally, agriculture has relied on the eyes and hands of experienced farmers to identify the right crops to pick. “Harvesting” robots equipped with AI technologies and data from cameras and sensors can now make this decision in real time. The AI system influences its environment by making recommendations (whether and how to pick) for a given set of objectives (selecting produce at the optimal level of ripeness/maturity, with least damage possible to the crop). It does so by utilising machine and/or human-based inputs (such as a large tagged database of

images) to: perceive its environment (a camera will capture an image and send to an application); abstract the perception into models (object recognition algorithms that can recognise produce under different light conditions); and use model interpretation to form a recommendation of options for outcomes (selection of produce to pick and in what order).[54]

What is an AI system lifecycle?

The AI system lifecycle typically involves the following four phases: i) 'design, data and models'; ii) 'verification and validation'; iii) 'deployment'; and iv) 'operation and monitoring'. Within the design, data and models phase, sub-phases include (in varying order, depending on the system):

- Planning and design of the AI system, including articulating the system's concepts and objectives, underlying assumptions, context and requirements, and potentially building a prototype;
- Data collection and processing, including gathering and cleaning data, performing checks for completeness and quality, and documenting the characteristics of the dataset.
- Model building and interpretation, including the creation or selection of models/algorithms, their calibration and/or training and interpretation.

The verification and validation phase involves executing and tuning models, with tests to assess performance. Deployment involves piloting, checking compatibility with legacy systems, ensuring regulatory compliance, managing organisational change and evaluating user experiences. Finally, the operation and monitoring phase of an AI system involves operating the AI system and continuously assessing its recommendations and impacts (both intended and unintended) in light of objectives and ethical considerations. In this phase, problems are identified and adjustments made or, if necessary, an AI system may be retired.

Who are the stakeholders and actors in AI systems?

Stakeholders encompass all public and private sector organisations and individuals involved in, or affected by, AI systems, directly or indirectly. They include, *inter alia*, civil society, the technical and academic communities, industry, governments, labour representatives and trade unions as well as individuals as workers or data subjects.

AI actors are those who play an active role in the AI system lifecycle, and are a subset of stakeholders. They include public and private sector organisations or individuals who acquire AI systems to deploy or operate them. It encompasses technology developers, systems integrators and service and data providers. Experts themselves may be drawn from multiple fields, and include *inter alia* data scientists, data engineers, governance experts, domain experts, model engineers, software engineers, and testers.

Linking AI systems, lifecycles, stakeholders and actors to the AI Principles

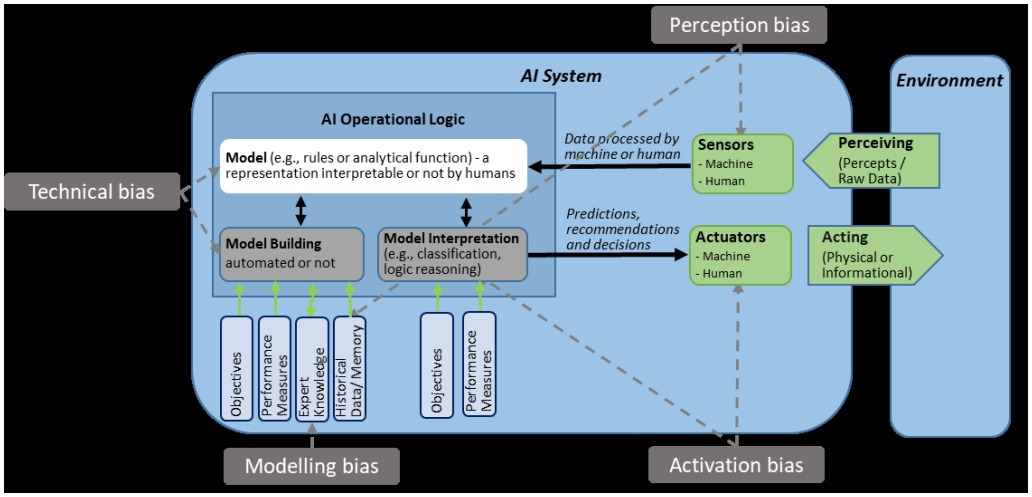
As noted earlier, a common understanding of concepts such as the AI system lifecycle can help policy makers have more informed and targeted discussions on pathways to implementing the AI Principles. As one concrete example, mitigating or avoiding bias would contribute to achieving the principle of "human

[54] For instance, Williams et al. (2019) discuss robotic kiwifruit harvesting using machine vision, neural networks and robotic arms. https://doi.org/10.1016/j.biosystemseng.2019.03.007

centred values and fairness". Figure 3 below depicts how four different types of bias can appear throughout the AI system:

- Perception bias occurs when the data collected over- or under-represents a certain population and causes the system to work better (or worse) for that population compared to others.
- Technical bias occurs when the technology itself introduces bias or inaccuracies due to, for instance, algorithms that perform better with certain AI system variables or features being introduced into an AI system with different variables or features.
- Modelling bias occurs when a manual design of a model by experts does not take into account some aspects of the environment, be it consciously or unconsciously.
- Activation bias occurs when the outputs of the AI system are used in the environment in a biased way.

Figure 3. Areas of the AI system in which biases can appear



Source: OECD (2019b).

Knowing that issues such as bias can appear at multiple points within an AI system highlights the general importance of risk management processes. Risk management can be implemented throughout the AI system lifecycle, to identify, assess, prioritise and treat potential risks that could adversely affect the behaviour of a system. The use of risk management and documentation of decisions made at each lifecycle stage can contribute to a system's general level of transparency and an organisation's accountability for the system – two key elements of the G20 AI Principles.

With respect to the example of bias outlined above, relevant AI actors can use risk management strategies to ensure they seek to mitigate or avoid biases throughout the lifecycle. Asking questions such as "Which stakeholders would be affected by selection bias?", "What level of selection bias would be acceptable in view of the benefits of the system?", and "How can actors ensure risk stays at an acceptable level?", could help manage and mitigate the risk of data selection over-representing or under-representing certain populations and the resulting disadvantage to certain groups. Notably, for policy makers, challenges are very different depending on the area of application of the AI system. For instance, AI systems related to sales present less serious consequences than those related to e.g. decisions related to justice, finance or health.

Resources:

OECD (2019a), *Artificial Intelligence in Society*, OECD Publishing, Paris, https://doi.org/10.1787/eedfee77-en

OECD (2019b), "Scoping the OECD AI principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)", *OECD Digital Economy Papers*, No. 291, OECD Publishing, Paris, https://doi.org/10.1787/d62f618a-en

OECD AI Policy Observatory (OECD.AI)

STIP Compass International Database on Science, Technology and Innovation Policy

www.oecd.ai

@OECDinnovation

STI.contact@oecd.org