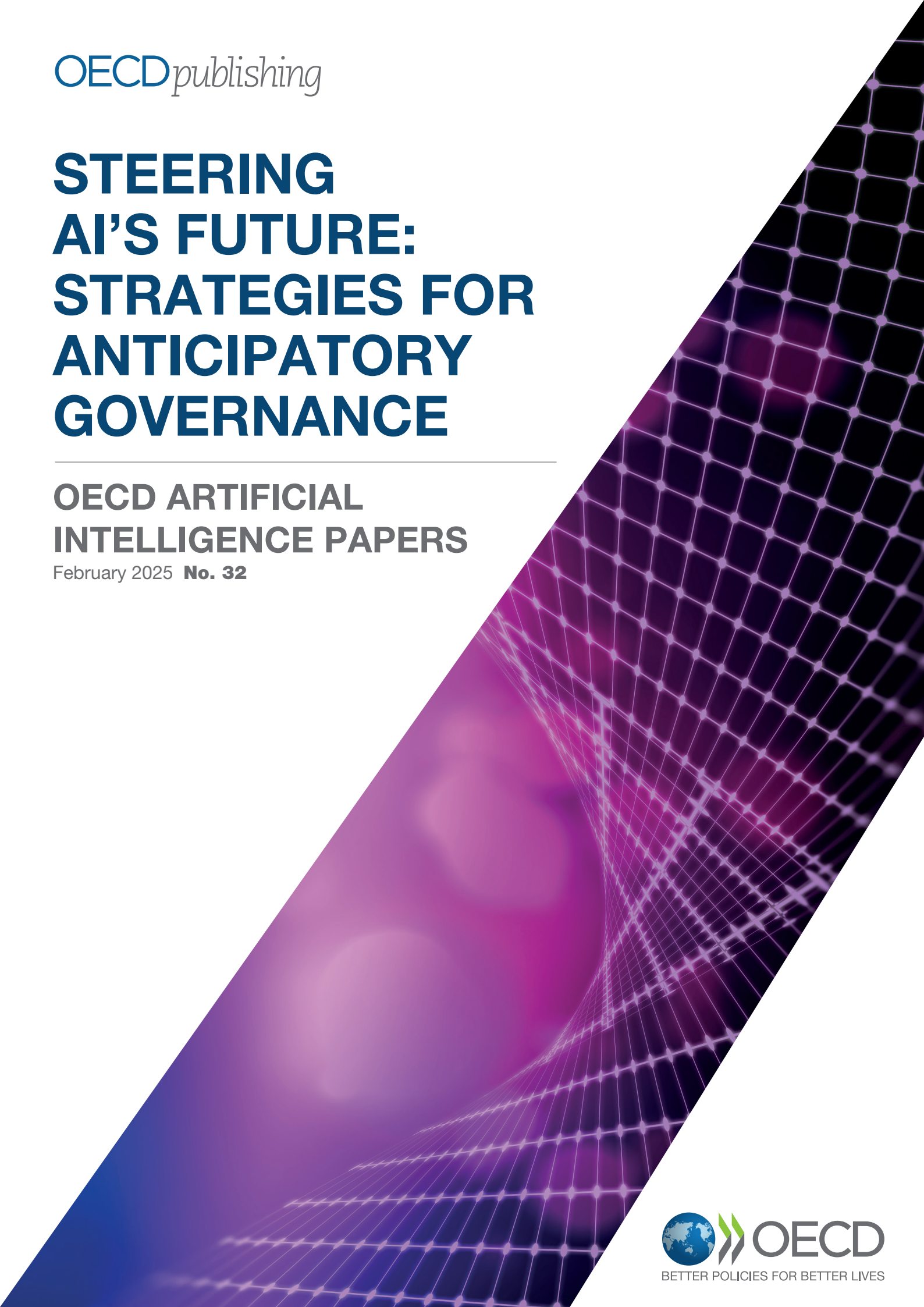
OECD publishing

STEERING AI'S FUTURE: STRATEGIES FOR ANTICIPATORY GOVERNANCE

OECD ARTIFICIAL INTELLIGENCE PAPERS
February 2025 **No. 32**



OECD
BETTER POLICIES FOR BETTER LIVES

Foreword

This report emphasises the need for adaptable governance frameworks to navigate the rapid advancements and complexities of artificial intelligence (AI). Highlighting the potential and challenges of AI, the paper advocates for the inclusion of proactive strategies that anticipate future developments and ensure responsible innovation.

The report was written by Karine Perset, Luis Aranda and Bénédicte Rispal under the supervision of Audrey Plonk, Deputy Director of the OECD Science Technology and Innovation Directorate. The report also benefitted from the inputs of delegates for the OECD Working Party on Artificial Intelligence (AIGO), the Working Party on Bio-, Nano-, and Converging Technologies (BNCT) and the Committee for Scientific and Technology Policy (CSTP). Andreia Furtado, John Tarver and Sarah Ferguson provided editorial support.

This report was developed as a contribution to the OECD's 2023-2024 horizontal project on Going Digital, Phase IV, under the Technology Governance Pillar. This paper was approved and declassified by written procedure by the OECD Digital Policy Committee (DPC) on 13 December 2024 and prepared for publication by the OECD Secretariat.

Note to Delegations:

This document is also available on O.N.E Members & Partners under the reference code:

DSTI/DPC/AIGO(2024)10/FINAL

Acknowledgements

This report was developed as a contribution to the OECD's 2023-2024 horizontal project on Going Digital, Phase IV, under the Technology Governance Pillar, in collaboration with the Committee for Scientific and Technological Policy (CSTP), the Regulatory Policy Committee (RPC), the Investment Committee (INV), the Digital Policy Committee (DPC) and the OECD Strategic Foresight Unit.

This report was presented at the Working Party on Bio-, Nano-, and Converging Technologies (BNCT) and the Committee for Scientific and Technology Policy (CSTP) and circulated to the AI Governance Working Party (AIGO), the Digital Policy Committee (DPC), the Regulatory Policy Committee (RPC), the Investment Committee (INV) and the Working Party on Responsible Business Conduct (WPRBC). As such, it benefited from feedback and suggestions from numerous country delegates and experts. The authors thank in particular the US Delegation to DPC for their invaluable insights.

The authors would also like to thank David Winickoff and Laura Kreiling for their guidance, leadership and effective co-ordination of this horizontal project. Additionally, Rashad Abelson and Jamie Berryhill provided substantial inputs for the drafting of sections 4 and 2, respectively.

Finally, the authors thank all those who have contributed to the report throughout its development. This includes Jerry Sheehan, Audrey Plonk, Hanna-Mari Kilpelainen, Alessandra Colecchia, David Winickoff, Laura Kreiling, Jamie Berryhill, Douglas K.R. Robinson, Gallia Daor and Julia Carro (Directorate for Science, Technology and Innovation); Rebecca King (Public Governance Directorate); Barbara Bijelic and Rashad Abelson (Directorate for Financial and Enterprise Affairs); and Rafal Kierzenkowski, Dexter Docherty and Hamish Hobbs (OECD Strategic Foresight Unit) for providing valuable comments and guidance. The authors would like to thank Andreia Furtado, John Tarver and Sarah Ferguson for editorial support. The overall quality of this report benefited significantly from their engagement.

Table of contents

Foreword 2

Acknowledgements 3

Abstract 6

Résumé 7

Executive summary 8

Introduction 10

1 Guiding values for AI governance 11
- Identifying shared values for AI governance 11
- Enabling deliberative AI governance processes 12
- Embedding values throughout the AI innovation process 13

2 Strategic intelligence approaches to AI governance 14
- Sentinel and real-time monitoring approaches for AI governance 14
- Strategic foresight approaches for AI governance 16

3 Stakeholder engagement for AI governance 19
- Informative engagement for AI governance 19
- Consultative engagement for AI governance 19
- Collaborative engagement for AI governance 20
- AI as tool to boost engagement in AI governance 21

4 Agile governance for AI 22
- Controlled environments and regulatory sandboxes 22
- Standards and by-design approaches for AI governance 24
- A responsible business conduct (RBC) approach to AI governance 26

5 International co-operation for AI governance 28

Conclusion 30

References 31

FIGURES

Figure 1. Five elements of emerging technology governance 10
Figure 1.1. The OECD AI Principles 12
Figure 1.2. The Catalogue of Tools and Metrics for Trustworthy AI 13
Figure 2.1. AI Incidents Monitor (AIM) 15
Figure 2.2. Weak signals and future AI threats identified by the AI Incidents Monitor 15
Figure 4.1. High-level AI risk-management interoperability framework 25

TABLES

Table 4.1. Benefits and challenges of regulatory sandboxes 23

BOXES

Box 2.1. The OECD.AI expert group on AI incidents: Defining, monitoring and reporting AI incidents 16
Box 2.2. What is strategic foresight? 17
Box 2.3. Illustrative potential future AI benefits, risks and solutions prioritised by the OECD.AI Expert Group on AI Futures 18
Box 3.1. OECD Network of Experts on AI (ONE AI) 21
Box 4.1. Sample regulatory sandboxes for AI 24
Box 4.2. The RBC approach in practice: leveraging the role of finance in AI 27

Abstract

This report emphasises the necessity for adaptable governance frameworks to manage the rapid developments and complexities of Artificial Intelligence (AI). By highlighting the potential and challenges posed by AI, the paper advocates for the incorporation of proactive strategies that anticipate future advancements and ensure responsible innovation. This report maps existing OECD and related initiatives in AI to illustrate how the five elements of anticipatory governance of the OECD Framework for Anticipatory Governance of Emerging Technologies apply to AI governance. The learnings and insights in this domain can help guide the governance of other emerging technologies.

Résumé

Ce rapport soulève la nécessité de disposer de cadres de gouvernance adaptables pour faire face aux progrès rapides et à la complexité de l'intelligence artificielle (IA). Soulignant le potentiel et les défis de l'IA, le document met en avant l'importance d'inclure des stratégies proactives qui anticipent les développements futurs et garantissent une innovation responsable. Ce rapport présente des initiatives existantes de l'OCDE et des initiatives connexes dans le domaine de l'IA afin d'illustrer comment les cinq éléments du Cadre relatif à la gouvernance anticipative des technologies émergentes s'appliquent à la gouvernance de l'IA. Les enseignements tirés dans ce domaine peuvent aider à guider la gouvernance d'autres technologies émergentes.

Executive summary

The swift advancement of Artificial Intelligence (AI) emphasises the importance of integrating anticipatory and future-oriented approaches as components of AI governance.

Artificial intelligence (AI) technologies present significant opportunities but also challenges. Nevertheless, the complexity and rapid evolution of AI make anticipating future opportunities and challenges difficult. This, in turn, poses challenges for the design of governance mechanisms and tools that are both effective and durable. In this context, agile and proactive governance approaches are necessary. Governments should seek and leverage knowledge from relevant stakeholders to identify, adjust, and develop appropriate regulatory approaches.

The OECD Framework for Anticipatory Governance of Emerging Technologies supports governments and societies to plan for and manage the challenges stemming from technologies.

By leveraging the OECD's existing work and legal instruments, including the OECD AI Principles and the framework for the classification of AI systems, the OECD Framework for Anticipatory Governance of Emerging Technologies empowers governments, innovators, and societies to foresee governance challenges and enhance their ability to shape innovation, both by design and post-deployment. It offers five key elements for governing emerging technologies and AI: guiding values, strategic intelligence, stakeholder engagement, agile regulation, and international co-operation.

Guiding values should be intentionally embedded into the AI innovation process.

Policymakers and other actors should identify foundational values to support responsible innovation, as exemplified by the OECD AI Principles. They should establish robust processes and involving experts from all stakeholder groups, including those at the forefront of AI innovation, to deliberate on proposed values and on application. They should also integrate values into every phase of the innovation cycle by utilising tools such as those found in the OECD.AI Catalogue of Tools and Metrics for Trustworthy AI.

Collecting strategic intelligence is essential for technologies with significant societal impacts but uncertain timelines, pathways, and real-world applications.

Sentinel and real-time monitoring tools and methods, like the OECD AI Incidents Monitor (AIM), are key to anticipating and managing emerging trends, risks, and opportunities. They can enable the identification of weak signals and possible future threats. Tools for technology appraisal, such as technology impact assessments and strategic foresight, should complement these approaches. Future-

focused activities, like the OECD.AI Expert Group on AI Futures, can help the global AI community understand and proactively shape AI's possible medium and long-term impacts.

Stakeholder engagement is essential for the success of AI governance initiatives.

Based on the level of involvement, stakeholder engagement can be categorised into three primary activities: informative engagement, encompassing blogs and social media updates; consultative engagement, which includes public consultations and open discussions; and collaborative engagement, exemplified by the Global Partnership on AI (GPAI) and the OECD Working Party on AI Governance (AIGO).

AI governance must remain agile to keep up with the rapid advancements in AI technology.

Adaptive approaches can help ensure that regulatory frameworks effectively balance innovation with ethical and safety considerations. Strategies that promote forward-looking and agile governance of AI systems include, amongst others, utilising controlled environments like regulatory sandboxes for more formal regulation, enabling interoperability through standards and by-design approaches, and Responsible Business Conduct (RBC) guidelines for AI. The latter underscores the role of businesses in ensuring due diligence across the AI value chain.

AI governance is a global issue that requires international co-operation to be effective.

International AI governance remains a dynamic field. Discussions are ongoing to address the rapid pace of AI development and the diverse challenges and opportunities it presents. A range of initiatives is emerging in this regard, with efforts spanning global, regional, and multilateral levels. These initiatives, led by various stakeholders like international organisations and national authorities, include promoting interoperability and facilitating shared good practices in AI governance.

Experiences in AI governance may provide valuable insights for other emerging technologies.

Governments have implemented diverse strategies and regulations, including forward-looking approaches and policies, to govern AI. While many of these efforts are still nascent, several lessons and good practices are starting to emerge. These insights could prove valuable to the governance of other emerging technologies, such as quantum and biotechnologies.

Despite progress, steering the future of AI through anticipatory governance requires additional efforts.

This paper draws primarily on the OECD's AI governance experiences, highlighting the critical role of forward-looking approaches in effectively managing and guiding AI developments. While anticipatory governance efforts at the OECD and among key stakeholders have made progress in shaping the future of AI, much more remains to be done to fully address its challenges and harness its benefits. Innovative approaches from other fields and emerging technologies could offer valuable insights to enhance AI governance.

Introduction

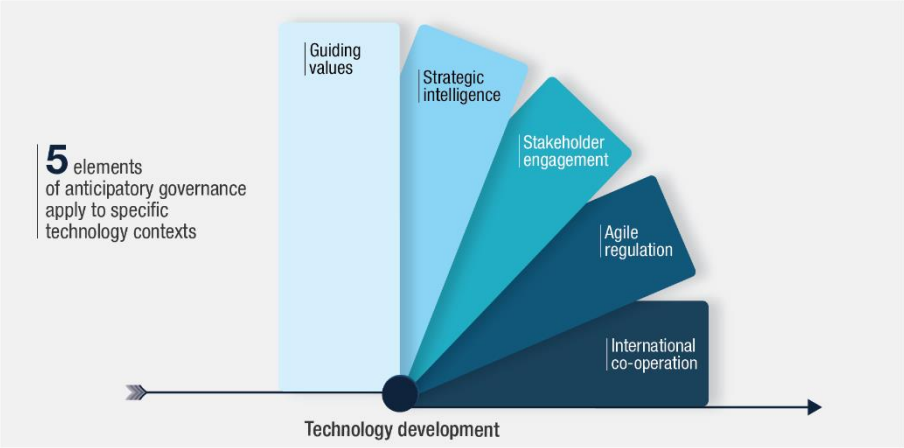
Artificial intelligence (AI) systems are increasingly shaping daily life, offering opportunities for economic growth, and tackling social, economic and environmental challenges. Yet, these systems also present growing risks, such as enabling malicious cyber activity, spreading disinformation, invasive surveillance and privacy violations (OECD, 2024[1]). As AI continues to evolve, governing these developments becomes even more crucial.

AI governance addresses the need to promote the benefits of AI systems while preventing and mitigating their risks. However, the rapid pace of AI developments makes designing governance mechanisms that can stand the test of time challenging. To help address this challenge, it is crucial to include anticipation and forward-looking approaches in AI governance.

The OECD Framework for Anticipatory Governance of Emerging Technologies equips governments, other innovation actors and societies to anticipate and get ahead of governance challenges and build longer-term capacities to shape innovation more effectively. This framework consists of five interdependent elements and associated tools for agile and anticipatory governance of emerging technologies: (1) guiding values, (2) strategic intelligence, (3) stakeholder engagement, (4) agile regulation and (5) international co-operation (Figure 1).

Agile approaches, including controlled, experimental environments and standards, can promote adaptive regulation and international co-operation. This report draws from existing OECD and related initiatives in AI to examine how the five elements of anticipatory governance figure in the AI policy landscape.

Figure 1. Five elements of emerging technology governance



Source: OECD (2024[2]), "Framework for Anticipatory Governance of Emerging Technologies", OECD Science, Technology and Industry Policy Papers, No. 165, OECD Publishing, Paris, https://doi.org/10.1787/0248ead5-en.

1 Guiding values for AI governance

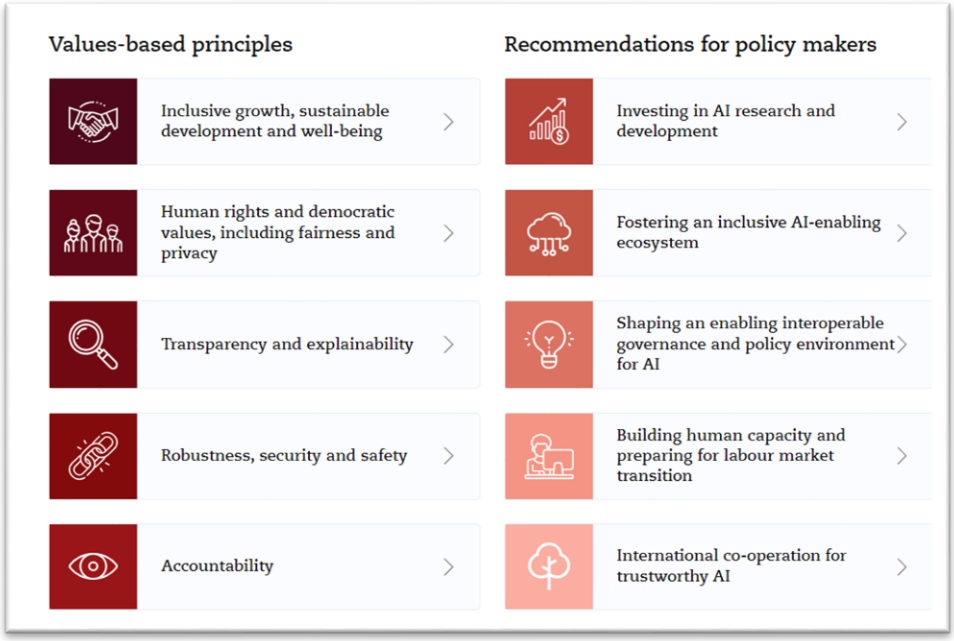
The first element of the OECD Framework for Anticipatory Governance of Emerging Technologies emphasises the importance of embedding guiding values in the innovation process (OECD, 2024[2]). This includes identifying a set of foundational values and technology-specific values to root responsible innovation; building robust processes and engaging fora in which to discuss these values and how they should be applied in particular contexts; and integrating values through different means in different phases of the innovation cycle.

Identifying shared values for AI governance

Effective AI governance requires guiding values and principles to ensure the trustworthy development and use of AI systems. The OECD Recommendation on Artificial Intelligence, also known as the "OECD AI Principles", was initially adopted in 2019 as the first intergovernmental standard on AI. The Principles were updated in May 2024 to consider new technological and policy developments, ensuring they remain robust and fit for purpose (Čorba et al., 2024[3]).

The OECD AI Principles aim to foster innovative and trustworthy AI which upholds human rights and democratic values. They guide AI actors in their efforts to develop trustworthy AI and provide policymakers with recommendations for effective AI policies. As such, they set out a framework containing ten principles – divided into five values-based principles and five recommendations to governments – to promote and implement trustworthy AI (Figure 1.1). The Principles include both foundational (e.g., respect for human rights and privacy) and AI-specific values (e.g., transparency and explainability).

Figure 1.1. The OECD AI Principles



Source: OECD (2024), OECD AI Principles, https://oecd.ai/ai-principles.

Since their adoption in 2019, these Principles have become a global reference point for trustworthy AI. Several countries around the world have leveraged the Principles to design their governance frameworks. Additionally, countries use the OECD AI Principles and related tools to shape policies and create AI risk frameworks, building a foundation for global interoperability between jurisdictions. Today, the European Union, the Council of Europe, the United States, the United Nations and other jurisdictions use the OECD's definition of an AI system and lifecycle in their legislative and regulatory frameworks and guidance. The principles, definition and lifecycle are all part of the OECD Recommendation on Artificial Intelligence (OECD, 2023[4]).

Enabling deliberative AI governance processes

Embedding values into governance processes involves creating fora at various levels of governance and across diverse stakeholder communities to share good practices and gather input for political and technological decision-making. An example might be to support technology observatories like the OECD.AI Policy Observatory (OECD.AI).

OECD.AI is a forum where countries and stakeholder groups join forces to shape trustworthy AI. Its tools, data and other AI policy resources are freely accessible to all actors and stakeholder groups in developed and developing countries. As a global hub for AI policy, OECD.AI serves as a one-stop-shop that brings stakeholders together to inform policy responses on emerging topics, from AI risk assessments to tools for implementing trustworthy AI and measuring national AI compute capacities.

OECD.AI helps governments worldwide to put guiding values for AI governance – especially the OECD AI Principles – into practice.

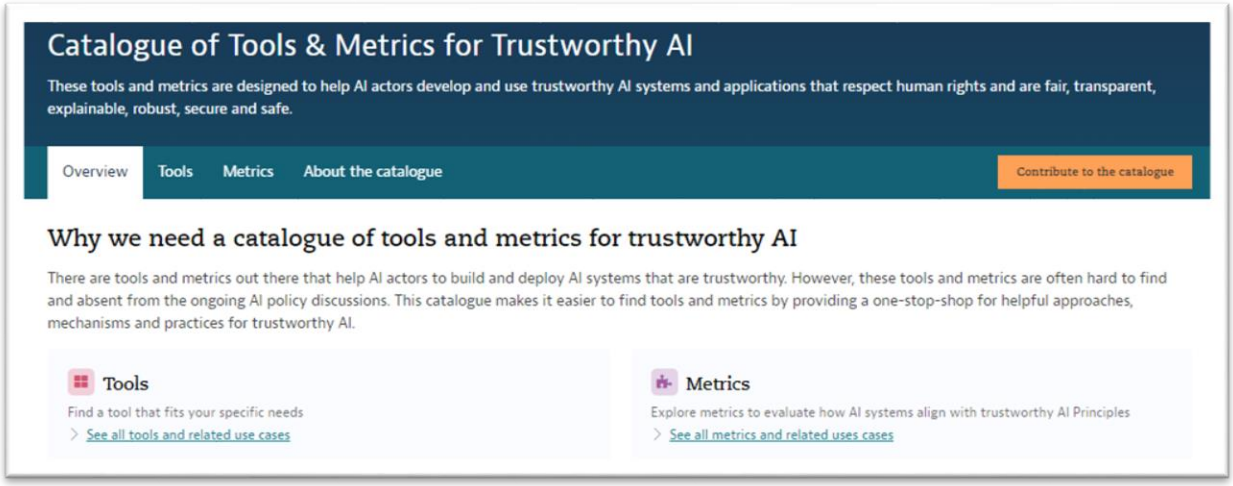
Embedding values throughout the AI innovation process

AI actors should strive to integrate values throughout the AI innovation process and the lifecycle of an AI system, not merely as an aspirational initial statement. For instance, explainability, transparency, privacy, avoiding bias and minimising disinformation are among the most critical challenges for AI practitioners, and can be particularly hard to attain for advanced AI systems. While many tools exist today to address these challenges, it is not always easy to find them and even more difficult to know which ones are the most effective.

The Catalogue of Tools & Metrics for Trustworthy AI aims to help solve this issue (Figure 1.2). It provides the much-needed space where AI practitioners from all over the world can share and compare tools and build upon each other's efforts to create global best practices and speed up the process of implementing the OECD AI Principles throughout the AI system lifecycle. Additionally, the catalogue allows users to submit their experiences as use cases, where they can give guidance, insights and a general appreciation of the tools.

Figure 1.2. The Catalogue of Tools and Metrics for Trustworthy AI

A living repository of tools and metrics to help AI actors develop and use trustworthy AI.



Source: OECD.AI (2025), Catalogue of Tools & Metrics for Trustworthy AI, www.oecd.ai/catalogue.

2 Strategic intelligence approaches to AI governance

Strategic intelligence involves collecting, processing, analysing, disseminating and utilising information to support long-term decision-making and planning (Kuosa, 2011[5]). Strategic intelligence enables forward-looking analysis of emerging technologies like AI, including potential future developments, economic impacts, ethical and societal considerations, and potential benefits and risks. While it is challenging to reach consensus on future technological risks and their likelihood, strategic intelligence mechanisms are crucial for technologies with potentially high societal impacts but uncertain timelines and pathways (OECD, 2024[2]).

This section illustrates strategic intelligence mechanisms informing AI governance today, including sentinel and real-time monitoring approaches and strategic foresight.

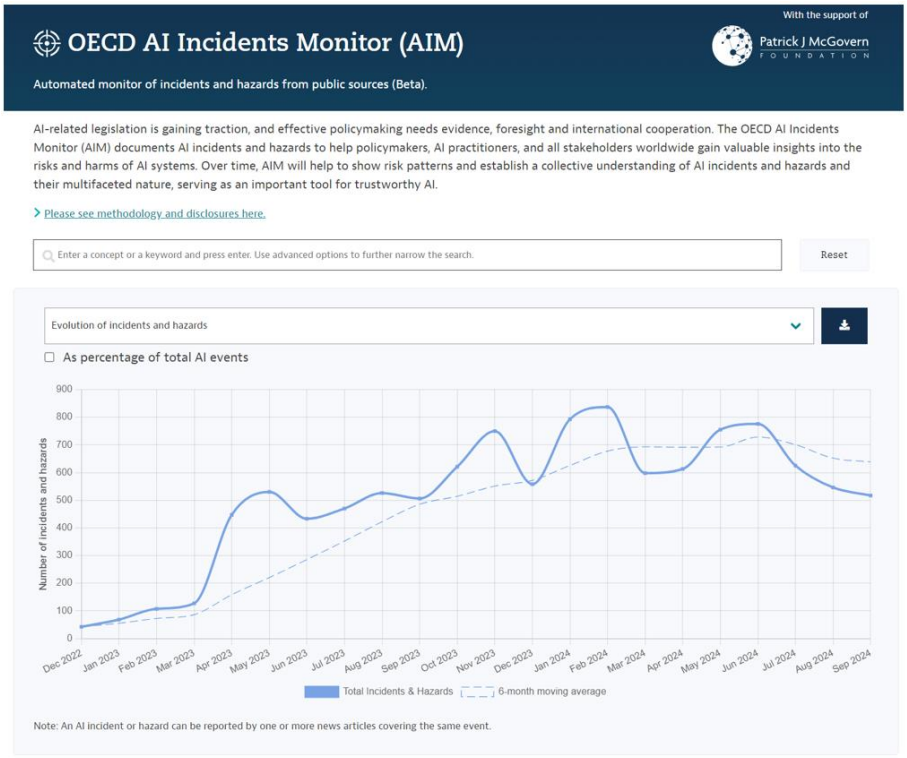
Sentinel and real-time monitoring approaches for AI governance

Sentinel and real-time monitoring are two key mechanisms in technology governance approaches to anticipate and manage emerging trends, risks, and opportunities. Sentinel monitoring involves systematically scanning the environment (e.g., news, social media, research articles) for weak signals, trends, and disruptions that could significantly impact economies or societies. Unlike sentinel monitoring, which focuses on identifying emerging trends and weak signals, real-time monitoring emphasises the continuous tracking and analysis of ongoing events and environment changes.

As AI use grows, so do occurrences of AI incidents and hazards. An example of both a sentinel and real-time monitoring approach to AI governance is the OECD AI Incidents Monitor (AIM). Using machine learning models, the AIM identifies and classifies AI incidents and hazards reported in reputable international media globally. It helps policymakers, AI practitioners, and all stakeholders worldwide gain valuable insights into AI risks and harms (Figure 2.1). Over time, and with its continuous enhancement, the AIM will help to show patterns and establish a collective understanding of AI incidents and their multifaceted nature and serve as an important tool for trustworthy AI.

Figure 2.1. AI Incidents Monitor (AIM)

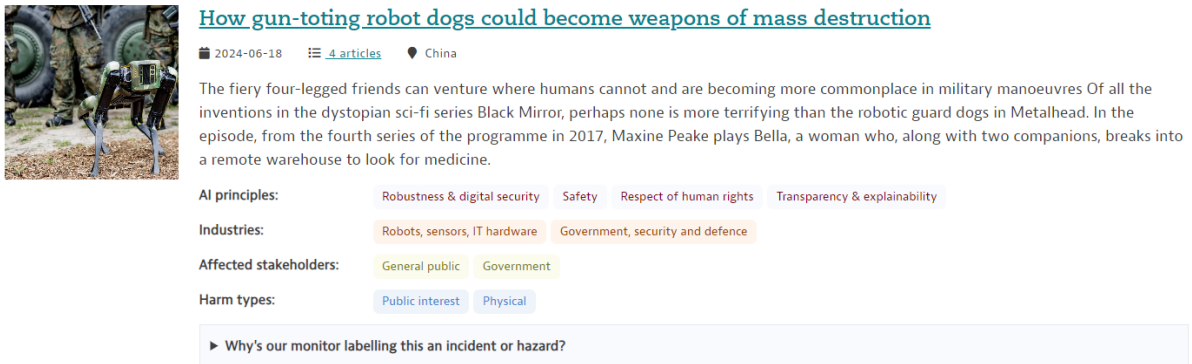
Tracking of AI incidents in real-time to inform AI governance discussions.



Note: Methodology and disclosures available at https://oecd.ai/incidents-methodology.
Source: OECD.AI (2025), OECD AI Incidents Monitor (AIM), www.oecd.ai/incidents.

Besides providing a mechanism to monitor the risks and harms of AI systems in real-time, the AIM also enables strategic intelligence insights by helping to identify weak signals and possible future threats (Figure 2.2).

Figure 2.2. Weak signals and future AI threats identified by the AI Incidents Monitor



Note: The AIM methodology and disclosures are available at https://oecd.ai/incidents-methodology.
Source: OECD.AI (2024[6]), OECD AI Incidents Monitor (AIM), https://oecd.ai/en/incidents.

The AIM was created to serve as a resource to track and comprehend the evolution of AI-related challenges. However, it currently relies solely on incidents reported by the media, which most likely

results in incomplete and possibly non-representative information. An open submission and review process allowing stakeholders to submit new reports or contribute to existing ones is being developed to overcome this limitation and promote transparency, accountability and participation. Also, it is planned that incident and hazard data from news articles be complemented with court rulings and decisions from public oversight bodies wherever available.

To ensure trustworthy AI, relevant AI actors and stakeholders need to handle risks in a coordinated and coherent fashion. This involves having a shared understanding of AI incidents. Reporting and monitoring of incidents should be interoperable worldwide, allowing AI practitioners and policymakers to learn from global experiences. To achieve this interoperability, the OECD has defined AI incidents and hazards and is developing a common framework for reporting them (Box 2.1).

Box 2.1. The OECD.AI expert group on AI incidents: Defining, monitoring and reporting AI incidents

The OECD.AI expert group on AI incidents operates through two streams of work. The first is focused on defining AI incidents and hazards and developing a common reporting framework. The second leverages the theoretical work on definitions and frameworks to monitor AI incidents and hazards via the AI incidents monitor (AIM), launched at the Paris Peace Forum in November 2023.

Combined, these two streams of work establish an evidence-based monitoring system of AI incidents and hazards, enabling governments to prevent, address and learn from these events.

Source: OECD.AI (2024[6]), OECD (2023[7]; 2024[8]) and OECD.AI (2024[9])

While implementing sentinel real-time monitoring approaches is expected to benefit AI governance, it can also pose challenges, including ensuring the quality and accuracy of the collected data, extrapolating weak signals in the present into potential future pathways, and integrating these systems into existing AI governance frameworks. Other tools for technology appraisal, such as technology assessments and strategic foresight should thus complement these approaches.

Strategic foresight approaches for AI governance

The long-term impacts of rapidly advancing AI technologies are still largely unknown and intensely debated. Experts highlight potential risks as well as significant benefits from AI. Future-focused activities are essential to understand AI's possible long-term impacts and to shape them proactively.

The difficulty of identifying and agreeing on future opportunities and challenges of AI is compounded by its complexity and rapid evolution. Various approaches drawn from a range of disciplines are used to assess AI progress and anticipate future impacts. These approaches and methodologies often complement each other. By combining diverse perspectives and developing a nuanced understanding of various assumptions and uncertainties, strategic foresight can help navigate these challenges (Box 2.2).

Box 2.2. What is strategic foresight?

Strategic foresight is a structured and systematic way of using ideas about the future to anticipate and better prepare for change. It involves exploring different plausible futures that could arise, and the opportunities and challenges they might present. Strategic foresight achieves this by uncovering implicit assumptions, challenging dominant perspectives, and engaging with surprising and significant disruptions that might otherwise be dismissed or ignored. It employs a range of methodologies that include scanning the horizon for emerging changes, analysing weak signals and megatrends, and developing multiple scenarios, to reveal and discuss ideas about the future.

Foresight can support government policymaking in the following ways:

- better anticipation of changes that could emerge in the future;
- policy innovation to reveal options for experimentation through innovative approaches; and
- future-proofing to stress-test existing or proposed strategies and policies.

Strategic foresight does not attempt to offer definitive answers about what society's future will hold. It understands the future as an emerging entity that is only partially visible in the present, not a predetermined destiny that can be fully known in advance. There are no hard facts about the future and the evidence base is always incomplete. The objective is not to 'get the future right', but to broaden and reframe the range of plausible developments to consider.

Source: OECD (2025), Strategic Foresight, www.oecd.org/strategic-foresight.

1. Strategic foresight approaches used with regards to AI include:

- **Expert surveys and consultations**: This involves tapping into the "wisdom of crowds" to generate forecasts in the form of timelines for potential developments of AI and accompanying milestones (Tetlock and Scoblic, 2020[10]). Such approaches are particularly useful for identifying areas of consensus and disagreement among experts and highlighting areas of uncertainty for further research.
- **Scenario planning and road mapping:** This involves creating and exploring hypothetical future narratives based on different assumptions and variables. This process helps identify potential risks and opportunities by capturing uncertainties, interrelationships and possible time frames that may influence AI development. For example, alternative scenarios can show how AI might create opportunities or challenges for various stakeholders (Ringland et al., 2020[11]). Road mapping offers a framework to align strategies and actions required to achieve specific AI goals or milestones over time (Ricard, 2011[12]).
- **Trends and data analysis:** This involves collecting, processing and interpreting data on the current state of AI to identify patterns, correlations, insights and opportunities. These findings can be extrapolated to predict future landscapes of AI (Eckersley and Nasser, 2019[13]; Constantin, 2019[14]; Martínez-Plumed et al., 2019[15]).
- **Horizon scanning**: This involves a broad look at emerging developments that could alter expected trajectories in AI, such as weak or early signals, trends in adjacent domains and wild cards (i.e., low-probability, high-impact events that can significantly alter the course of future developments) (OECD, 2017[16]).
- **Other methodologies** include literature reviews to provide a comprehensive overview of existing knowledge and identify gaps or controversies (Snyder, 2019[17]); public engagement through consultations and polls to gauge public awareness, trust and expectations regarding

AI; empirical evidence to validate and enhance theoretical models and scenarios; and philosophical analysis to help interpret current AI advancements (Honorof, 2023[18]).

The OECD.AI Expert Group on AI Futures uses foresight methods to spot potential AI trajectories and suggest actions

Since July 2023, the OECD.AI expert group on AI futures has been using foresight to help governments make more effective, forward-looking policies on AI (OECD.AI, 2024[19]).

For example, the group conducted a thorough literature review, held discussions among experts, and engaged the public to identify 21 potential future AI benefits, 38 potential future AI risks, and 66 potential policy solutions and AI governance approaches (OECD.AI, 2024[20]). These were prioritised to focus on key government actions (Box 2.3).

Box 2.3. Illustrative potential future AI benefits, risks and solutions prioritised by the OECD.AI Expert Group on AI Futures

Potential future AI benefits: Accelerated scientific progress; better economic growth, productivity gains and living standards; reduced inequality and poverty; better approaches to urgent and complex issues, including mitigating climate change and advancing other Sustainable Development Goals (SDGs); better decision-making, sense-making and forecasting; improved information production and distribution; better healthcare and education services; improved job quality; empowered citizens, civil society and social partners; improved institutional transparency and governance, instigating monitoring and evaluation.

Potential future AI risks: Facilitation of increasingly sophisticated malicious cyber activity; manipulation, disinformation, fraud and resulting harms to democracy and social cohesion; races to develop and deploy AI systems cause harms due to a lack of sufficient investment in AI safety and trustworthiness; unexpected harms result from inadequate methods to align AI system objectives with human stakeholders' preferences and values; power is concentrated in a small number of companies or countries; minor to serious AI incidents and disasters occur in critical systems; invasive surveillance and privacy infringements; governance mechanisms and institutions unable to keep up with rapid AI evolutions; AI systems lacking sufficient explainability and interpretability erode accountability; exacerbated inequality or poverty within or between countries.

Potential future AI policy solutions: Establish clearer rules, including on liability, for AI harms; consider approaches to restrict or prevent certain "red line" AI uses; require or promote the disclosure of key information about some types of AI systems; ensure risk management procedures are followed throughout the lifecycle of AI systems that may pose a high risk; mitigate competitive race dynamics in AI development and deployment that could limit fair competition and result in harms; invest in research on AI safety and trustworthiness approaches, including AI alignment, capability evaluations, interpretability, explainability and transparency; facilitate educational, retraining and reskilling opportunities to help address labour market disruptions and the growing need for AI skills; empower stakeholders and society to help build trust and reinforce democracy; mitigate excessive power concentration; targeted actions to advance specific future AI benefits.

Source: OECD (2024[1]), "Assessing potential future artificial intelligence risks, benefits and policy imperatives", OECD Artificial Intelligence Papers, No. 27, OECD Publishing, Paris, https://doi.org/10.1787/3f4e3dfb-en.

3 Stakeholder engagement for AI governance

Engaging stakeholders and society in policymaking is essential, as it brings diverse perspectives and enriches understanding. Involving stakeholders, including citizens, in public decision making, helps anticipate public concerns, improves communication and is key to enhancing public trust in government institutions (OECD, 2020[21]). However, public consultations should be meaningful, only involving citizens when there is room for influence in decision-making, genuine commitment from leadership to consider inputs, adequate resources, and sufficient time within the decision-making cycle to collect and incorporate inputs (OECD, 2022[22]). In emerging technology policy, soliciting input early in technology development through "anticipatory engagement" is crucial to enable equity and inclusion. Including stakeholders' involvement in processes from the outset, from agenda-setting to governance design, helps to align science and technology with societal needs and goals (OECD, 2024[2]).

Stakeholder engagement is key to the success of AI governance initiatives. It involves interacting with individuals or groups who have an interest or stake in the AI ecosystem – be it suppliers of AI knowledge or resources, actors in the AI system lifecycle, users and affected stakeholders (OECD, 2023[23]). Based on the degree of involvement, stakeholder engagement could be divided it into three main types: informative, consultative and collaborative engagement (OECD, 2017[24]).

Informative engagement for AI governance

Informative engagement provides stakeholders with information to understand a project or initiative. Examples include websites, newsletters, webinars and reports. One of the primary goals of informative engagement is keeping stakeholders up to date with accurate and timely information (OECD, 2017[24]).

As a global hub for AI policy, one of the essential missions of the OECD.AI Policy Observatory is to keep stakeholders informed on AI developments in a timely manner. It achieves this through various means:

- **Data and evidence:** OECD.AI's live data section utilises alternative data sources to display timely trends in AI development and usage.
- **Blog:** Weekly blog posts by experts on the AI Wonk provide insights into the latest trends and developments in AI.
- **Social media**: Leveraging social media, OECD.AI conducts outreach and effectively engages the wider AI community, notably through its LinkedIn community of over 30 000 members.

Consultative engagement for AI governance

Consultative engagement is a collaborative approach that involves actively seeking input, feedback, and perspectives from stakeholders. The goal is to foster trust, inclusivity, and effective decision-making by considering and acting on the insights and concerns of relevant parties. Consultative engagement is

more interactive than informative engagement, yet the final decision-making power usually resides with the organisation conducting the consultation (OECD, 2017[24]).

Part of the OECD's mission to raise awareness around implementing the OECD AI Principles is to ensure that open dialogue shapes trustworthy AI. Some examples of consultative engagement in AI governance include:

- **Public consultations:** Several public consultations have been conducted to inform OECD.AI work on priority AI governance topics, including the classification of AI systems according to their potential impact on individuals, society and the planet (OECD, 2022[25]) and the measurement of national AI compute capacity (OECD, 2023[26]).
- **Calls for submissions or volunteers:** OECD.AI leverages calls for submissions or volunteers to consult and gather inputs from stakeholders on specific projects or initiatives, such as to develop and grow the Catalogue of Tools and Metrics for Trustworthy AI.
- **Open discussions:** Forums or platforms where stakeholders can openly share their perspectives, exchange ideas and collaborate have been leveraged to inform AI governance discussions, including on potential AI futures.

Collaborative engagement for AI governance

Collaborative engagement is a participatory approach where stakeholders work together in partnership to address common challenges, share resources, and co-create solutions, emphasising co-operation and shared decision-making (OECD, 2017[24]).

Three main initiatives exist at the OECD to foster collaborative engagement for AI governance:

- **Global Partnership on AI**: The OECD and the Global Partnership on Artificial Intelligence (GPAI) joined forces in July 2024 to advance an ambitious agenda for implementing human-centric, safe, secure and trustworthy AI embodied in the principles of the OECD Recommendation on AI. GPAI capitalises on the extensive, multi-disciplinary and multi-stakeholder expertise of an AI expert community, pooling the networks and groups currently contributing to the GPAI and OECD.
- **OECD Working Party on AI Governance (AIGO):** The OECD's Digital Policy Committee (DPC) has a Working Party on Artificial Intelligence Governance (AIGO) to oversee its work on AI policy. Working party members are nominated by OECD member governments and are primarily national officials responsible for AI policies in their respective countries. Representatives from all stakeholder groups also participate in AIGO.
- **OECD Network of Experts on AI (ONE AI):** The network of experts works with AIGO as an informal group of AI experts from government, business, academia, trade unions and civil society. The network provides AI-specific policy advice for the OECD's work on AI policy and contributes to the OECD.AI Policy Observatory (Box 3.1).

Box 3.1. OECD Network of Experts on AI (ONE AI)

The network – now expanded to include GPAI experts – provides the OECD with an "on the ground" perspective on AI. It is a forum where the OECD collaborates and shares information with other international initiatives, organisations and experts on pressing AI governance issues.

The network has seven expert groups on: AI, data and privacy; building an AI index; AI risk and accountability; AI futures; AI incidents; compute and climate; and AI and health. It provides a space for the international AI community to have in-depth discussions about shared AI policy opportunities and challenges.

The network brings together over 400 AI experts from many sectors and backgrounds, including:

- **AI policy experts** from national governments, international organisations, civil society, trade unions, technical organisations and the private sector. Network members from national governments are often AI policy experts in charge of coordinating, designing and implementing national AI strategies.
- **AI technical experts**, such as professors, researchers, computer scientists, and engineers.
- **Experts from social sciences and humanities**, such as experts in AI-related legal and ethical issues.

ONE AI and the GPAI expert networks are in the process of being merged to combine their respective expertise and capabilities and expand their collective reach.

Source: https://oecd.ai/network-of-experts

AI as tool to boost engagement in AI governance

AI systems can serve as powerful tools for engagement, including for AI governance. Governments increasingly use AI applications, as enablers of larger and more meaningful citizen participation into policymaking. It can serve as a sense-maker (processing consultation data), content moderator or facilitator, virtual assistant or chatbot, or translator. Such tools may enable a larger and more meaningful citizen participation into policymaking (OECD, 2024[27]).

AI-based tools like the open-source platform Polis have been used for participatory dialogue to inform referendums, public policy development and increase youth outreach, among others. Such tools have been adopted across OECD member countries, including in Colombia, where the virtual agent Chatico serves as an official channel for citizen participation and communication. This demonstrates how AI systems can inform policy by enhancing communication and engagement with the broader public. Challenges remain in using AI systems for citizen participation, particularly in ensuring the quality of AI-generated information and addressing access barriers for non-technologically savvy users (OECD, 2024[27]).

4 Agile governance for AI

AI governance needs to be agile and anticipatory to keep pace with the rapid advancements and evolving nature of AI technology. This adaptability ensures that regulatory frameworks can effectively enable innovation while addressing ethical and safety considerations. Agile governance helps improve rules and regulations by promoting more flexible and adaptable regulatory approaches. For instance, creating standards and guidelines early in the development of new technologies can lead to stronger regulations later on as these technologies mature and their impacts become clearer (OECD, 2021[28]).

Agile governance of emerging technologies can leverage various tools, including iterative and adaptive regulatory cycles that address stakeholder and public concerns while coordinating across different regulatory areas. It can leverage regulatory experimentation tools, such as testbeds and sandboxes, and apply outcome-based approaches that set clear goals and performance indicators while allowing flexibility in how these goals are met. Agile governance can benefit from non-binding methods like principles, guidelines, and codes of conduct to complement formal regulations. Engaging the private sector early with approaches like "ethics-by-design" and Responsible Business Conduct also promotes agile governance responsible innovation (OECD, 2024[2]).

This section explores three approaches that facilitate a forward-looking and agile governance of AI systems: using controlled environments like regulatory sandboxes; enabling interoperability through standards and by-design approaches; and developing Responsible Business Conduct (RBC) guidelines for AI.

Controlled environments and regulatory sandboxes

Regulatory experimentation tools are used to test new economic, institutional and technological approaches and legal provisions in an adapted regulatory environment. These include innovation testbeds, living labs, policy prototypes and regulatory sandboxes (OECD, 2024[29]). This section highlights regulatory sandboxes as one of the most commonly used experimentation tools for AI governance, allowing authorities to collaborate with firms to test innovative products or services that challenge existing legal frameworks. Participating firms receive waivers from certain legal provisions or compliance processes to encourage innovation (OECD, 2023[30]).

Regulatory sandboxes can differ in their approaches but generally share key features: they are temporary, typically lasting around six months; they involve collaboration between regulators and firms; they offer waivers for certain legal provisions and provide customised legal support for specific projects, often through a trial-and-error process. The technical and market data collected helps regulatory authorities determine if existing legal frameworks are adequate or need adjustment (OECD, 2023[30]).

Regulatory sandboxes must strike a delicate balance. On one hand, they aim to foster innovation by allowing experimentation without strict regulatory constraints. However, they also need to consider potential risks to stakeholders, society, and the environment. Regulatory sandboxes present both benefits and challenges to regulators, firms and consumers (Table 4.1).

Table 4.1. Benefits and challenges of regulatory sandboxes

	Benefits	Challenges
To regulators	• Inform long-term policy making through learning and experimentation • Signal commitment to innovation and learning • Engage and communicate collaboratively with market participants • Adjust regulations that may restrain innovation or result in safety risks	• Insufficient technical expertise within regulatory bodies • Multi-disciplinary scope of AI products calls for collaboration of multiple stakeholders • Defining the evaluation methods to determine the eligible participating firms
To firms	• Reduce time to market by streamlining the approval process • Reduce regulatory uncertainty by providing clarity on prohibited technologies • Gather feedback on regulatory requirements or risks • Improve access to capital • Facilitate market entry for companies, especially SMEs and start-ups, by providing accessible information about legal frameworks	• Difficulty in understanding the eligibility and selection criteria to participate • Capacity constraints may limit the number of firms that can participate
To consumers	• Promote introduction of innovative and potentially safer products • Increase access to AI products and services	• Inefficient implementation of regulatory sandboxes and their associated safeguards can lead to negative impacts on consumers

Source: OECD (2023[30]), The Norwegian Data Protection Authority (2023[31]) and Attrey, Lesher and Lomax (2020[32])

Countries like Norway and Singapore have implemented sandbox initiatives to foster the development of trustworthy AI. While their goals and structure may vary, these initiatives are designed to support innovation and ensure compliance with ethical and regulatory standards (Box 4.1).

Box 4.1. Sample regulatory sandboxes for AI

The Norwegian regulatory sandbox for responsible AI, launched in autumn 2020 by the Data Protection Authority of Norway, focuses on AI projects involving the use of personal data. Its primary objective is to promote good data protection practices in the development and use of AI systems.

In October 2023, Singapore's Infocomm Media Development Authority (IMDA) developed a sandbox dedicated to innovations leveraging generative AI. This sandbox aims at informing global standards and policy recommendations for generative AI and large language models (LLMs).

Several other AI sandboxes have been set up to align with future regulations and promote international co-operation. For example, the Spanish government's AI regulatory sandbox aims to clarify EU AI Act requirements and provide guidelines to help AI system providers and users prepare for the forthcoming regulatory changes. It also invites EU member states to participate, fostering international collaboration.

Source: The Norwegian Data Protection Authority (2023[31]), Infocomm Media Development Authority (2023[33]) and European Commission (2022[34]).

Standards and by-design approaches for AI governance

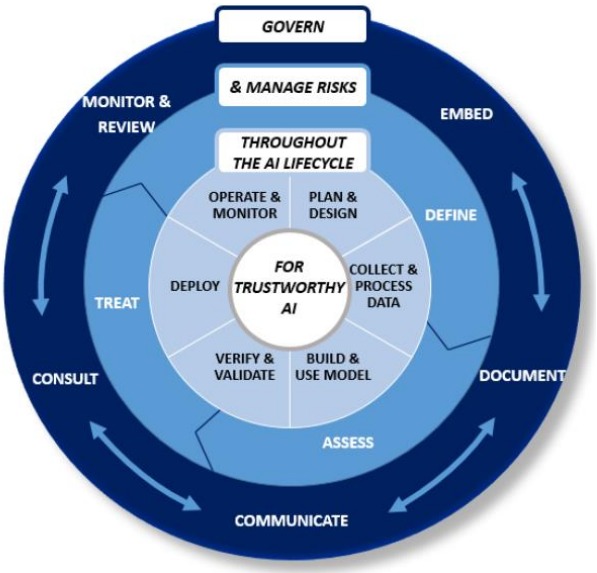
Standards and by-design approaches are essential to create an AI governance ecosystem that is agile and interoperable. Standards provide consistent norms and guidelines to facilitate compliance and accountability, including on fairness, privacy, safety, transparency and interoperability. By-design approaches, such as privacy, ethics, security, and transparency by design, integrate these principles into AI systems from the outset. Together, these strategies help to enable trust in AI development and use and facilitate co-operation across countries and stakeholder groups.

Integrating risk management processes into each stage of the AI system lifecycle – including planning and design, data collection and processing, model building and validation, deployment, operation and monitoring – is crucial to promote accountability in AI. Processes, tools and by-design approaches already exist to implement values-based principles and manage risks at each stage of the AI lifecycle (OECD, 2023[23]).

Interoperability in risk management facilitates agile AI governance by enabling consistent risk assessment and response to emerging threats. It promotes collaboration among stakeholders and simplifies compliance and reporting processes. Based on a study of burgeoning AI risk management frameworks and standards, the high-level AI risk-management interoperability framework identifies four common steps to manage AI risks while ensuring respect for human rights and democratic values: (1) **Define** scope, context, actors and criteria; (2) **Assess** the risks at individual, aggregate, and societal levels; (3) **Treat** risks in ways commensurate to cease, prevent or mitigate adverse impacts; and (4) **Govern** the risk management process (Figure 4.1) (OECD, 2023[23]).

Figure 4.1. High-level AI risk-management interoperability framework

Governing and managing risks throughout the lifecycle for trustworthy AI.



Source: OECD (2023[23]), "Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI", *OECD Digital Economy Papers*, No. 349, OECD Publishing, Paris, https://doi.org/10.1787/2448f04b-en..

A recent OECD study of leading AI risk management frameworks and standards looked at the commonalities and differences among their scope, target audiences and requirements (OECD, 2023[35]). The report found that while the order of operations, target audience, risk scope, and specific terminology used may differ, all the frameworks generally seek to achieve the same outcomes (responsible, ethical, trustworthy AI) through roughly the same risk management process. Importantly, most of the differences between frameworks relate to the governance of the risk management process (the 'Govern' function).

Several national and international voluntary industry standards and initiatives could help pave the way for a more agile and interoperable AI governance ecosystem. Examples include the United States National Institute of Standards and Technology AI Risk Management Framework (NIST RMF) (NIST, 2023[36]), the ISO/IEC 23894:2023 Risk Management Guidelines and other related ISO/IEC standards on AI (ISO, 2023[37]), and the Institute of Electrical and Electronics Engineers 7000-21 Standard Model Process for Addressing Ethical Concerns during System Design (IEEE 7000-21) (IEEE, 2021[38]).

The G7 Hiroshima AI Process provides a voluntary code of conduct which includes elements of risk management (G7, 2023[39]). The Code of Conduct sets out a commitment to identify, address and report on certain risks linked to what it refers to as "advanced" AI systems. In Europe, the European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CENELEC) are working together to create harmonised standards that will provide a clear and recognised path towards compliance with the EU AI Act. These standards aim to establish consistent guidelines and requirements for AI systems, ensuring they meet the regulatory expectations set forth by the EU.

While standards and by-design approaches aim to improve interoperability and agility in AI governance, they also present challenges. The proliferation of incompatible standards can hinder interoperability and add complexity, increasing compliance costs. Additionally, compliance with some technical standards depends on developing specific metrics for aspects like explainability, transparency, privacy, and safety,

which may not yet exist or be agreed upon. Certifications of conformity with different standards must also be interoperable to ensure a cohesive and functional AI ecosystem.

A responsible business conduct (RBC) approach to AI governance

Since its inception, the OECD has been committed to leveraging the power of competitive markets, science and new technology as a driving force for sustainable economic and social development. In 1976, the OECD developed the first and most comprehensive international standard on responsible business behaviour, with the adoption of the OECD Guidelines for Multinational Enterprises on Responsible Business Conduct (the MNE Guidelines) (OECD, 2023[40]). The MNE Guidelines acknowledge and encourage the positive contributions that business can make to economic, environmental and social development, and also recognise that business activities can result in adverse impacts related to workers, human rights, the environment, rule of law, consumers and corporate governance.

The MNE Guidelines go beyond the traditional, philanthropic Corporate Social Responsibility (CSR) approach by placing the expectations on business to proactively address potential harms. The MNE Guidelines specifically recommend that companies carry out due diligence of their own operations, products and services, and also of their business relationships, to address potential and actual adverse impacts they might cause, contribute to or could be directly linked to. The concept of maximising positive potential by first addressing negative impacts forms the basis for Responsible Business Conduct (RBC).

RBC due diligence is defined as the process through which companies can identify, prevent, mitigate and account for how they address their actual and potential adverse impacts as an integral part of business decision-making and risk management systems. The OECD has developed a Due Diligence Guidance for Responsible Business Conduct (the RBC Guidance) and sector-specific guidance for carrying out due diligence in minerals, garment & footwear, agriculture, as well as for institutional investors (OECD, 2018[41]). The OECD Working Party on AI Governance (AIGO), together with the Working Party on RBC (WPRBC), is currently considering how to tailor the RBC approach for companies in the AI value chain (OECD, 2023[35]).

Risks of adverse impacts can manifest themselves in multiple ways in the development and use of AI systems. While RBC due diligence should cover all stages of the AI system lifecycle, companies have the greatest opportunity to anticipate and address risks during the earliest stages of the innovation process. By applying an RBC-by-design strategy developers can prevent and mitigate potential risks of technologies at every step of development.

Concretely, RBC due diligence early in the innovation process could involve:

- **Setting out internal policies** on the objectives and limits of AI systems the company seeks to develop.
- **Identifying and prioritising significant impacts** of the AI system's potential use or misuse, including by understanding the nature of the AI system (e.g., facial recognition technology, surveillance technology, autonomous weapons); the type of user (e.g., national security agencies, healthcare providers, judiciary bodies); the geographical scope (e.g., regions with high corruption, labour mistreatment, and human rights abuses); and the system's performance (e.g., risks of poor performance in critical tasks).
- **Designing appropriate safeguards** to prevent adverse impacts.
- **Meaningfully engaging with relevant stakeholders** – particularly representatives of impacted communities – to help ensure that risks are identified and prevention and mitigation measures are adequate.

AI companies are already making efforts to conduct due diligence and implement responsible business conduct practices. Over the course of the last few years, companies have been conducting and publishing human rights risk assessments and policies on how they will develop and use trustworthy AI. Notable examples include Microsoft's impact assessment on technologies licensed to US law enforcement; Google's celebrity recognition API impact assessment; and Meta's responsible use guide for open-source generative AI.

A key characteristic of the RBC approach is that it encompasses the entire AI value chain, not just the organisations developing and deploying AI systems. RBC expectations also extend to actors directly linked to those technologies, such as actors supplying AI knowledge and resources as well as users of AI products and services. This could include content creators, data curators, digital infrastructure providers, data labellers, hardware manufacturers, and investors. It also includes actors outside the technology sector that rely on digital products and services such as in healthcare, retail, agriculture, human resources and manufacturing.

The MNE Guidelines specify that all these value chain actors are expected to build and use leverage (including collective leverage) to influence the entity causing the adverse impact, in order prevent, mitigate, or remedy the impact (Box 4.2). This makes the MNE Guidelines a unique and especially powerful agile governance tool in interconnected digital economies.

Box 4.2. The RBC approach in practice: leveraging the role of finance in AI

Venture capital (VC) plays a key role in the development of AI systems, and AI now plays an increasingly significant role in venture capital (growing from 3% of all VC investments in 2012 to 20% of all VC investments in 2021) (OECD, 2021[42]). Finance organisations are often in an ideal position to exert leverage as they are supporting clients to shape and define a project's outlook. Financial institutions can play a major role in leading clients to identify and address actual or potential adverse RBC impacts.

For example, in 2023 Norway's USD 1.4 trillion wealth fund announced its measures to promote RBC in AI companies in its portfolio (Reuters, 2023[43]). Specifically, companies will be asked to demonstrate accountability, risk management and due diligence according to the OECD AI Principles and MNE Guidelines (NBIM, 2023[44]).

There have already been high profile cases and backlash against large banks for their role in financing certain technology used in human rights abuses. For example, a Swiss bank is the subject of a complaint due to alleged failure to observe the MNE Guidelines regarding human rights due diligence with regards to its relationship with a Chinese surveillance technology firm (OECD, 2021[45]). Similarly, a large financial institution was reported to have sold its loan to a surveillance firm at a loss following reports that the firm's technology was used to spy on journalists and human rights defenders (Smith, 2019[46]).

While different standardisation approaches can be valid, a complex and divergent ecosystem can challenge governments and organisations in promoting and implementing responsible business practices. Convergence in the AI governance ecosystem could be possible and desirable, as seen in the gold sector. The London Bullion Market Association, Dubai Multicommodity Centre, Borsa Istanbul, and the India National Stock Exchange all require demonstrating implementation of OECD RBC standards to trade gold. This industry-wide convergence around a single, government-backed standard has promoted global uptake of RBC practices across the gold value chain (OECD, 2022[47]).

5 International co-operation for AI governance

AI is a global issue; its development, use, and impacts extend beyond national borders. To be effective, anticipatory AI governance necessitates both national and international governance systems to interoperate. International co-operation in AI governance should be values-driven, based on forward-looking evidence, involve stakeholder engagement, and remain agile in adapting norms and institutions (OECD, 2024[2]).

International AI governance today remains a dynamic field, with ongoing discussions to address the rapid pace of AI development and the diverse challenges and opportunities it presents. A range of initiatives is emerging in this regard, with efforts spanning global, regional, and multilateral levels:

- **On the global stage**, organisations such as the United Nations (UN), the World Bank, and UNESCO are undertaking efforts to create frameworks and guidelines for safe and trustworthy AI. For example, the UN convened a multi-stakeholder High-level Advisory Body on AI (HLAB) to undertake analysis and advance recommendations for the international governance of AI. The body consisted of 39 experts in various disciplines coming from government, the private sector and civil society from around the world. In September 2024, HLAB published a report providing recommendations for the UN, governments, and stakeholders to improve existing AI governance frameworks, manage AI-related risks and harness its global potential (UN HLAB, 2024[48]).
- **At regional level**, entities like the European Union, the Inter-American Development Bank (IDB) and the African Union are playing crucial roles. These organisations are creating harmonised AI policies, standards and initiatives that reflect the unique socio-economic contexts of their member states. For example, the European regulation on AI (the EU AI Act) and the African Union's proposed continental strategy on AI demonstrate significant strides toward cohesive regional AI governance frameworks.
- **Multilateral initiatives** bridge the gap between global and regional efforts in AI governance. Examples include GPAI, the OECD, the Council of Europe Convention on AI, NATO's AI Strategy, the EU-US Trade and Technology Council (TTC), Globalpolicy.AI, G7 and G20 initiatives, AI safety and governance summits, and the AI safety institutes network. These fora facilitate dialogue and collaboration among countries and regions, promoting an integrated approach to AI governance. In 2023, the G7 launched the "Hiroshima AI Process" in Japan, marking a milestone in AI governance. The same year, the first AI Safety Summit in the UK resulted in the Bletchley Declaration, signed by 28 countries, focusing on the safety of advanced AI systems. Globalpolicy.AI is an online platform developed to strengthen co-operation between intergovernmental organisations with complementary mandates on AI.

The OECD has played a crucial role in AI governance since 2019, when 42 countries adopted the OECD AI Principles as the first intergovernmental standard for AI. It has acted as a bridge among global, regional, and multilateral AI initiatives, leading efforts to inform policy by defining AI systems and their lifecycle, classifying them by policy implications, and defining and monitoring AI incidents and hazards,

among others. In July 2024, the OECD and GPAI formed an integrated partnership to promote an ambitious agenda focused on implementing human-centric, safe, secure, and trustworthy AI, as outlined in the OECD AI Principles.

There is growing momentum to create synergies among AI governance stakeholders and initiatives, fostering collaboration and coherence in AI governance efforts. For example, the UN and the OECD recently announced a partnership to collaborate on regular science- and evidence-based assessments of AI risks and opportunities (UN-OECD, 2024[49]). Collectively, these efforts underscore the importance of international co-operation and the need for adaptive, forward-looking and inclusive governance structures to navigate the complexities of AI.

Conclusion

The rapid advancement of AI necessitates robust governance mechanisms that can adapt to its evolving landscape. Anticipatory governance, which focuses on proactive and forward-looking approaches, is important for effectively managing the complexities and potential impacts of AI technologies. By incorporating anticipatory and future-oriented strategies, governments can better identify and tackle emerging challenges while harnessing AI's opportunities.

This report utilises the OECD Framework for Anticipatory Governance of Emerging Technologies to examine recent AI governance experiences from the OECD and beyond. Emphasising guiding values, strategic intelligence, stakeholder engagement, agile regulation, and international co-operation remains crucial in navigating the dynamic landscape of emerging technology governance.

While significant progress has been made, particularly through initiatives like the OECD AI Principles and the OECD.AI Policy Observatory, continued collaboration and innovation are imperative to fully realise AI's potential benefits while mitigating associated risks. By integrating lessons learned from other fields and emerging technologies, policymakers can further refine anticipatory governance frameworks to foster an environment conducive to safe, ethical, and innovative AI development globally.

References

Attrey, A., M. Lesher and C. Lomax (2020), "The role of sandboxes in promoting flexibility and innovation in the digital age", *OECD Going Digital Toolkit Notes*, No. 2, OECD Publishing, Paris, https://doi.org/10.1787/cdf5ed45-en. [32]

Constantin (2019), *Performance Trends in AI*, https://srconstantin.wordpress.com/2017/01/28/performance-trends-in-ai/. [14]

Čorba, J. et al. (2024), *Evolving with innovation: The 2024 OECD AI Principles update*, https://oecd.ai/en/wonk/evolving-with-innovation-the-2024-oecd-ai-principles-update. [3]

Eckersley, P. and Y. Nasser (2019), *Measuring the Progress of AI Research*, Electronic Frontier Foundation. [13]

European Commission (2022), *First regulatory sandbox on Artificial Intelligence presented*, https://digital-strategy.ec.europa.eu/en/news/first-regulatory-sandbox-artificial-intelligence-presented (accessed on 29 March 2024). [34]

G7 (2023), *Hiroshima Process International Code of Conduct for Advanced AI Systems*, https://www.mofa.go.jp/ecm/ec/page5e_000076.html. [39]

Honorof, M. (2023), *The future of AI could hinge on two philosophical concepts*, https://www.tomsguide.com/features/ai-philosophy-solipsism-blockhead. [18]

IEEE (2021), *7000-2021 - IEEE Standard Model Process for Addressing Ethical Concerns during System Design*, https://doi.org/10.1109/IEEESTD.2021.9536679. [38]

Infocomm Media Development Authority (2023), *First of its kind Generative AI Evaluation Sandbox for Trusted AI by AI Verify Foundation and IMDA*, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox. [33]

ISO (2023), *ISO/IEC 23894:2023*, https://www.iso.org/standard/77304.html. [37]

Kuosa, T. (2011), "Different approaches of pattern management and strategic intelligence", *Technological Forecasting and Social Change*, Vol. 78/3, pp. 458-467, https://doi.org/10.1016/j.techfore.2010.06.004. [5]

Martínez-Plumed, F. et al. (2019), *Accounting for the Neglected Dimensions of AI Progress*. [15]

NBIM (2023), *Responsible artificial intelligence*, https://www.nbim.no/en/publications/our-views/2023/responsible-artificial-intelligence/. [44]

NIST (2023), *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, https://doi.org/10.6028/NIST.AI.100-1. [36]

OECD (2024), "Assessing potential future artificial intelligence risks, benefits and policy imperatives", *OECD Artificial Intelligence Papers*, No. 27, OECD Publishing, Paris, https://doi.org/10.1787/3f4e3dfb-en. [1]

OECD (2024), "Defining AI incidents and related terms", *OECD Artificial Intelligence Papers*, No. 16, OECD Publishing, Paris, https://doi.org/10.1787/d1a8d965-en. [8]

OECD (2024), "Framework for Anticipatory Governance of Emerging Technologies", *OECD Science, Technology and Industry Policy Papers*, No. 165, OECD Publishing, Paris, https://doi.org/10.1787/0248ead5-en. [2]

OECD (2024), "Governing with Artificial Intelligence: Are governments ready?", *OECD Artificial Intelligence Papers*, No. 20, OECD Publishing, Paris, https://doi.org/10.1787/26324bc2-en. [27]

OECD (2024), "Regulatory experimentation: Moving ahead on the agile regulatory governance agenda", *OECD Public Governance Policy Papers*, No. 47, OECD Publishing, Paris, https://doi.org/10.1787/f193910c-en. [29]

OECD (2023), "A blueprint for building national compute capacity for artificial intelligence", *OECD Digital Economy Papers*, No. 350, OECD Publishing, Paris, https://doi.org/10.1787/876367e3-en. [26]

OECD (2023), "Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI", *OECD Digital Economy Papers*, No. 349, OECD Publishing, Paris, https://doi.org/10.1787/2448f04b-en. [23]

OECD (2023), *Common guideposts to promote interoperability in AI risk management*, https://doi.org/10.1787/ba602d18-en. [35]

OECD (2023), *OECD Guidelines for Multinational Enterprises on Responsible Business Conduct*, https://www.oecd-ilibrary.org/docserver/81f92357-en.pdf?expires=1694803478&id=id&accname=ocid84004878&checksum=E6287EE6E4D53B54875A8257FEE19961. [40]

OECD (2023), "Regulatory sandboxes in artificial intelligence", *OECD Digital Economy Papers*, No. 356, OECD Publishing, Paris, https://doi.org/10.1787/8f80a0e6-en. [30]

OECD (2023), "Stocktaking for the development of an AI incident definition", *OECD Artificial Intelligence Papers*, No. 4, OECD Publishing, Paris, https://doi.org/10.1787/c323ac71-en. [7]

OECD (2023), "The state of implementation of the OECD AI Principles four years on", *OECD Artificial Intelligence Papers*, No. 3, OECD Publishing, Paris, https://doi.org/10.1787/835641c9-en. [4]

OECD (2022), "OECD Framework for the Classification of AI systems", *OECD Digital Economy Papers*, No. 323, OECD Publishing, Paris, https://doi.org/10.1787/cb6d9eca-en. [25]

OECD (2022), *OECD Guidelines for Citizen Participation Processes*, OECD Public Governance Reviews, OECD Publishing, Paris, https://doi.org/10.1787/f765caf6-en. [22]

OECD (2022), *Report on the Implementation of the Recommendation of the OECD Council on Due Diligence Guidance for Responsible Supply Chains of Minerals from Conflict-Affected and High-Risk Areas*, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0386. [47]

OECD (2021), *Recommendation of the Council for Agile Regulatory Governance to Harness Innovation*, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0464. [28]

OECD (2021), "Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems"*, OECD Digital Economy Papers*, No. 312, OECD Publishing, Paris, https://dx.doi.org/10.1787/008232ec-en. [45]

OECD (2021), *Venture capital investments in artificial intelligence*, https://www.oecd.org/digital/venture-capital-investments-in-artificial-intelligence-f97beae7-en.htm. [42]

OECD (2020), *Innovative Citizen Participation and New Democratic Institutions: Catching the Deliberative Wave*, OECD Publishing, Paris, https://doi.org/10.1787/339306da-en. [21]

OECD (2018), *OECD Due Diligence Guidance for Responsible Business Conduct*, http://mneguidelines.oecd.org/OECD-Due-Diligence-Guidance-for-Responsible-Business-Conduct.pdf. [41]

OECD (2017), *Recommendation of the Council on Open Government*, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0438. [24]

OECD (2017), *The Next Production Revolution: Implications for Governments and Business*, OECD Publishing, Paris, https://doi.org/10.1787/9789264271036-en. [16]

OECD.AI (2024), *AI Futures*, https://oecd.ai/en/site/ai-futures (accessed on 3 July 2024). [19]

OECD.AI (2024), *AI Futures, What do you see as the as the most significant potential benefits and risks of AI 10+ years from now?*, https://oecd.ai/en/site/ai-futures/discussions/future-benefits-risks (accessed on 3 July 2024). [20]

OECD.AI (2024), *OECD AI Incidents Monitor (AIM) - How gun-toting robot dogs could become weapons of mass destruction*, https://oecd.ai/en/incidents/90734 (accessed on 3 July 2024). [6]

OECD.AI (2024), *OECD AI Incidents Monitor, methodology and disclosures*, https://oecd.ai/incidents-methodology (accessed on 15 February 2024). [9]

Reuters (2023), *Norway wealth fund to firms: use AI, but do it responsibly*, https://www.reuters.com/business/finance/norway-wealth-fund-firms-use-ai-do-it-responsibly-2023-08-15/. [43]

Ricard, L. (2011), *From Future Scenarios to Roadmapping: A practical guide to explore innovation and strategy.*, Joint Research Centre. [12]

Ringland, G. et al. (2020), *Scenarios and roadmapping - how to navigate an uncertain future*, University of Cambridge. [11]

Smith, R. (2019), "Jefferies and Credit Suisse set to lose on Israeli cyber security deal", https://www.ft.com/content/e390685a-5a10-11e9-939a-341f5ada9d40. [46]

Snyder, H. (2019), "Literature review as a research methodology: An overview and guidelines", *Journal of Business Research*, pp. 333-339. [17]

Tetlock, P. and J. Scoblic (2020), "A Better Crystal Ball: The Right Way to Think", *Foreign Affairs*, pp. 10-19. [10]

The Norwegian Data Protection Authority (2023), *Evaluation of the Norwegian Data Protection Authority's Regulatory*, https://www.datatilsynet.no/contentassets/41e268e72f7c48d6b0a177156a815c5b/agenda-kaupang-evaluation-sandbox_english_ao.pdf. [31]

UN HLAB (2024), *Governing AI for Humanity: Final Report*, UN, https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf. [48]

UN-OECD (2024), *OECD and UN announce next steps in collaboration on Artificial Intelligence: Press release*, https://www.oecd.org/en/about/news/press-releases/2024/09/oecd-and-un-announce-next-steps-in-collaboration-on-artificial-intelligence.html#:~:text=Our%20joint%20efforts%20will%20help,AI%20risk%20and%20opportunity%20assessments. [49]