OECDpublishing

ASSESSING POTENTIAL FUTURE ARTIFICIAL INTELLIGENCE RISKS, BENEFITS AND POLICY IMPERATIVES

OECD ARTIFICIAL INTELLIGENCE PAPERS

November 2024 **No. 27**

OECD
BETTER POLICIES FOR BETTER LIVES

Foreword

This report reviews research and expert perspectives on potential future AI benefits, risks and policy actions. It features contributions from members of the OECD Expert Group on AI Futures ("Expert Group"), which is jointly supported by the OECD AI and Emerging Digital Technologies division (AIEDT) and Strategic Foresight Unit (SFU), with regard to which items should be considered high priority by policymakers. It also considers existing public policy and governance efforts and remaining gaps.

The Expert Group is co-chaired by Stuart Russell (University of California, Berkeley; Centre for Human-Compatible AI), Francesca Rossi (IBM) and Michael Schönstein (Federal Chancellery of Germany). The complete list of members and relevant outputs on AI futures can be found at https://oecd.ai/site/ai-futures.

Because of the prospective nature of part of this report and the lack of rigorous study on some topics, many of the future-oriented aspects of its contents are necessarily speculative.

This report was discussed and reviewed by members of the Expert Group from September 2023 to July 2024. It was also discussed at the OECD Working Party on Artificial Intelligence Governance (AIGO) at its November 2023 meeting. This paper was approved and declassified by written procedure by the Digital Policy Committee on 30 October 2024 and prepared for publication by the OECD Secretariat.

This report contributes to the OECD's AI in Work, Innovation, Productivity and Skills (AI-WIPS) programme, which provides policymakers with new evidence and analysis to keep abreast of the fast-evolving changes in AI capabilities and diffusion and their implications for the world of work. AI-WIPS is supported by the German Federal Ministry of Labour and Social Affairs (BMAS) and will complement the work of the German AI Observatory in the Ministry's Policy Lab Digital, Work & Society. For more information, visit https://oecd.ai/wips and https://denkfabrik-bmas.de. It also contributes to the OECD Horizontal Foresight Initiative on Anticipating and Managing Emerging Global Transformations, which seeks to develop policy frameworks and risk management approaches to increase preparedness for AI, synthetic biology and other potential transformative developments. For more information, visit https://oe.cd/global-transformations.

This report was drafted by Jamie Berryhill (AIEDT), Hamish Hobbs and Dexter Docherty (SFU) in close collaboration with Expert Group co-chairs and members. Strategic direction and editing were provided by Karine Perset, Head of AIEDT, and Rafał Kierzenkowski, OECD Senior Counsellor for Strategic Foresight. Riccardo Rapparini, Robin Staes-Polet, Michaela Sullivan-Paul, Pablo Gomez Ayerbe and Moritz von Knebel made analysis and drafting contributions. The team gratefully acknowledges the input from Expert Group members, as well as from OECD colleagues Jerry Sheehan, Audrey Plonk, Hanna-Mari Kilpelainen, Gallia Daor, Molly Lesher, Jeremy West, Luis Aranda, Alistair Nolan, Sarah Bérubé and Rashad Abelson of the Directorate for Science, Technology and Innovation (STI); Uma Kalkar of SFU; Richard May of the Directorate for Financial and Enterprise Affairs (DAF); Stijn Broecke of the Directorate for Employment, Labour and Social Affairs (ELS) and Charles Baubion, Giulia Cibrario, James Drummond, Andras Hlacs, Becky King, Craig Matasick, Mauricio Mejia and Arturo Rivera Perez of the Directorate for Public Governance (GOV). The team also thanks John Tarver and Andreia Furtado for editorial support.

Note to Delegations:

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

DSTI/CDEP/AIGO(2023)13/FINAL

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Cover image: © Kjpargeter/Shutterstock.com.

Table of contents

FIGURES

Executive summary

The swift evolution of AI technologies calls for policymakers to consider and proactively manage AI-driven change. The OECD's Expert Group on AI Futures was established to help meet this need and anticipate AI developments and their potential impacts. This initiative aims to equip governments with insights to craft forward-looking AI policies. This report discusses research and expert insights on prospective AI benefits, risks, and policy imperatives. While offering guidance for policymakers, decision-makers are encouraged to remain aware of uncertainties, actively seek diverse perspectives and vigilantly monitor the societal implications of AI innovations.

Governments can shape AI policies to steer developments toward desirable futures

The Expert Group identified characteristics of desirable AI futures through a survey, discussions and scenario exploration. These include widely distributed AI benefits; respect for human rights, privacy and intellectual property rights; more and better jobs; resilient physical, digital and societal systems; mechanisms to maximise AI security and prevent misuse; steps to prevent excessive power concentration; strong risk management practices for training, deployment and use of AI systems that may carry high risks and international and multi-stakeholder co-operation for trustworthy AI. These characteristics embody the realisation of AI's benefits and mitigating its risks. Governments can take action to help realise positive AI futures. The OECD worked with Expert Group members through the survey and discussions to identify policy and governance priorities. Annex A provides details on the methodology for doing so.

Future benefits from AI include scientific breakthroughs and better lives…

The Expert Group identified 21 potential future AI benefits. Through ranking and synthesis of these, as discussed in Annex A, it put forth **ten priority benefits** that warrant policy focus:

1. accelerated scientific progress, such as through devising new medical treatments;
2. better economic growth, productivity gains and living standards;
3. reduced inequality and poverty, aided through poverty reduction efforts and improved agriculture;
4. better approaches to address urgent and complex issues, including mitigating climate change and advancing other Sustainable Development Goals (SDGs);
5. better decision-making, sense-making and forecasting through improved analysis of present events and future predictions;
6. improved information production and distribution, including new forms of data access and sharing;
7. better healthcare and education services, such as tailored health interventions and tutoring;
8. improved job quality, including by assigning dangerous or unfulfilling tasks to AI;
9. empowered citizens, civil society and social partners, including through strengthened participation;
10. improved institutional transparency and governance, instigating monitoring and evaluation.

… but future risks from AI include harms to individuals and societies

The Expert Group identified 38 potential future AI risks. Through ranking and synthesis of these, it put forth **ten priority risks** warranting enhanced policy focus:

1. facilitation of increasingly sophisticated malicious cyber activity, including on critical systems;
2. manipulation, disinformation, fraud and resulting harms to democracy and social cohesion;
3. races to develop and deploy AI systems cause harms due to a lack of sufficient investment in AI safety and trustworthiness;
4. unexpected harms result from inadequate methods to align AI system objectives with human stakeholders' preferences and values;
5. power is concentrated in a small number of companies or countries;
6. minor to serious AI incidents and disasters occur in critical systems;
7. invasive surveillance and privacy infringement that undermine human rights and freedoms;
8. governance mechanisms and institutions unable to keep up with rapid AI evolutions;
9. AI systems lacking sufficient explainability and interpretability erode accountability;
10. exacerbated inequality or poverty within or between countries, including through risks to jobs.

Some risks were not prioritised because they were rated less important overall, though individual Expert Group member rankings varied significantly. Opinions diverged particularly about the potential risk of humans losing control of artificial general intelligence (AGI). This is a hypothetical concept whereby machines could have human-level or greater "intelligence" across a broad spectrum of contexts.

Proactive policies and governance can help to capture AI's benefits and manage risks

The Expert Group identified 66 potential policy approaches to obtain AI benefits and mitigate risks. Through ranking and synthesis of these, it put forth **ten policy priorities** to help achieve desirable AI futures:

1. establish clearer rules, including on liability, for AI harms to remove uncertainties and promote adoption;
2. consider approaches to restrict or prevent certain "red line" AI uses;
3. require or promote the disclosure of key information about some types of AI systems;
4. ensure risk management procedures are followed throughout the lifecycle of AI systems that may pose a high risk;
5. mitigate competitive race dynamics in AI development and deployment that could limit fair competition and result in harms, including through international co-operation;
6. invest in research on AI safety and trustworthiness approaches, including AI alignment, capability evaluations, interpretability, explainability and transparency;
7. facilitate educational, retraining and reskilling opportunities to help address labour market disruptions and the growing need for AI skills;
8. empower stakeholders and society to help build trust and reinforce democracy;
9. mitigate excessive power concentration;
10. take targeted actions to advance specific future AI benefits.

Governments recognise the importance of these issues, but more needs to be done

Policy initiatives recognise the importance of these issues. Recent developments include the revision of the OECD AI Principles; finalisation of the European Union AI Act and Council of Europe Framework Convention on AI and Human Rights, Democracy and the Rule of Law; executive actions in countries such as the United States, the launch of national AI safety and research institutes; commitments endorsed by AI companies; efforts to increase relevant talent in government and apply existing regulation to the context of AI; public investments in AI research and development and initiatives of the United Nations and its agencies. Efforts on the horizon, such as the EU Liability Directive, may also advance beneficial AI. Yet, opportunities exist to take more concrete action. Governments should consider how best to implement priority policy actions and strengthen their capacities to help anticipate and shape AI futures.

1 Identifying desirable AI futures

Governments should consider the medium- and long-term implications of AI

The medium to long-term implications of rapidly advancing AI systems remain largely unknown and fiercely debated. Experts raise a range of potential future risks from AI, some of which are already becoming visible. At the same time, experts and others expect AI to deliver significant or even revolutionary benefits. Future-focused activities can help better understand AI's possible longer-term impacts and begin shaping them in the present to seize AI's benefits while managing its risks.

To this end, the OECD Expert Group on AI Futures ("Expert Group") is a multi-disciplinary group of 70 leading AI experts that helps address future AI challenges and opportunities by providing insights into the possible AI trajectories and impacts and by equipping governments with the knowledge and tools necessary to develop forward-looking AI policies.[1]

Policy actions today can help achieve desirable future scenarios

The Expert Group, through a survey, discussions, and scenario exploration exercises, presented its views on the characteristics of desirable AI futures in society and governance (see methodology in Annex A). These desirable futures embody the realisation of potential future AI benefits and the mitigation of key future risks. Positive futures will not occur automatically; they demand concrete action by policymakers, companies, and other AI actors.

Benefits from AI would be widely distributed

AI can accelerate scientific research and generate solutions that contribute to breakthroughs in areas such as healthcare and climate change. Certain policies could enable innovation in trustworthy AI, of which benefits would be shared widely within and between countries and equitably distributed across stakeholder groups, sectors and the public, while preventing system deployments or uses with substantial potential for harm. All countries, including emerging and developing economies, would benefit from AI's socio-economic potentials.

AI would empower people, civil society organisations (CSOs) and social partners

People would be empowered through AI, such as through new data-driven tools to make more informed decisions, including a focus on women and marginalised communities. Governments would facilitate this by leveraging AI to engage with citizens and incorporate their views into policymaking, thus reinforcing democracy and participation in public life. The capabilities of CSOs and social partners such as trade unions would be strengthened by AI, allowing them to better connect with and gather insights from citizens and workers. Through new means to analyse open government data and outputs, AI would enable CSOs worldwide to provide stronger independent oversight of government. This oversight role would be further facilitated by disclosure requirements or norms for certain AI systems that help understand their functioning and foster an ecosystem of independent evaluators. In the workplace, the use of AI would be trustworthy, and its benefits would be distributed fairly, with workers and social partners also able to leverage AI to bolster organising and inform collective bargaining. The public would have access to reliable, authentic information, enhanced and personalised education and reskilling opportunities.

Human rights, including privacy, would be respected

Developers and deployers of AI systems and third parties such as auditors would widely use benchmarks, evaluations, and technical tools to detect, mitigate, and correct harmful bias and discrimination. Frameworks and practices to ensure that AI systems are designed, developed, deployed, and used in accordance with human rights would be available and widely adopted. Policies and solutions to protect personal data would be in place, especially for use cases that may carry high risk and systems that may impact vulnerable populations.

Intellectual property rights would be respected and clarified if needed

Model developers would have clear guidance on which data can be used to train models and which data are protected by copyright. Rightsholders and other content generators would be empowered to make educated decisions about how their data and content are used.

Robust technical, procedural and educational tools would help keep AI systems transparent, explainable and aligned with human stakeholders' values

AI actors—those actively participating in the AI system lifecycle, including organisations and individuals that deploy or operate AI—could leverage robust procedures, technical approaches, and other methods to provide strong assurance that AI systems are safe and trustworthy. This would include ensuring appropriate transparency and explainability and aligning system behaviours with the values of human stakeholders.

Physical, digital and societal systems and ecosystems would be resilient

Technical tools and other protective measures against AI-facilitated malicious cyber activity would be developed and available to AI developers and deployers. Critical infrastructure, physical and information security requirements would be adapted to reflect risks posed by the use of AI. To help ensure the resilience of societal systems, government initiatives would help the transition of labour markets, including reskilling efforts and considering new social safety nets. In addition, a portfolio of efforts at international and domestic levels would reinforce democracy and information integrity, including via effective processes enabling free and fair elections and mitigating mass distribution of disinformation.

Effective mechanisms maximise AI security and prevent misuse by bad actors

AI systems would be designed, deployed and overseen in a way that minimises risks of misuse by malicious actors. AI security risks, more broadly, would be identified, mitigated and monitored through standardised processes.

Appropriate policies and measures would prevent excessive power concentration

Decisions pertaining to the development, deployment and use of AI systems with significant impacts on societies and economies would be decentralised where possible, with appropriate transparency measures, accountability processes, liability rules and effective oversight. With regard to AI and key inputs and enablers, mechanisms would be in place to mitigate the undue concentration of market, economic or political power in the hands of one or very few providers.

Governance measures would include strong risk management practices, including for training, deployment and use of AI systems that may carry high risks

There would be appropriately defined and enforced risk management approaches for developing, deploying, and using AI systems, particularly those that may pose elevated risks. These approaches would include transparent risk assessments, risk mitigation procedures, and red lines prohibiting uses representing unacceptable risks. Organisations involved in overseeing risk management, including those in the public sector, would have sufficient mandates, authorities and in-house inter-disciplinary skills and capacities to understand and oversee such approaches.

International and multi-stakeholder co-operation would facilitate safe and trustworthy AI

Strong co-operation would result in collective global, cross-sectoral rules, commitments and information sharing to promote AI safety and trustworthiness.

Governments' AI foresight efforts are expanding

Governments are working to build strategic foresight capacities (Box 1.1) through programmes such as the OECD Government Foresight Community and public sector foresight efforts (OECD, 2024[1]; 2023[2]; OECD/CAF, 2022[3]). This is critical given the rapid development of AI, its unknowns and the potential costs of falling behind. Expert Group members and OECD work[2] encourage governments to develop strong AI foresight capacities to continuously anticipate futures that may emerge. This can help them understand where, when and how to intervene, including to prepare for plausible changes.

Policymakers can build their strategic foresight capacities by defining a concrete value proposition for foresight efforts in policy and decision-making systems and processes, investing in strategic foresight and identifying and addressing challenges that hinder comprehensive approaches to strategic foresight, such as bureaucratic silos and barriers to dialogue with non-governmental actors to understand AI impacts.

Box 1.1. Strategic foresight can help anticipate potential AI futures

Due to the wide-reaching future impacts of AI, strategic foresight researchers are often drawn from a range of disciplines, use various methods and put forward diverse arguments. Strategic foresight combines these approaches to build a nuanced understanding of assumptions, including through:

- **Expert surveys and consultations** are used to generate forecasts in the form of timelines for potential AI developments and milestones. They can inform analysis and recommendations on how to address relevant ethical, social and technical issues. Such approaches are also useful for identifying areas of consensus and disagreement among experts.
- **Scenario planning and road mapping** create and explore hypothetical future narratives based on different assumptions and variables. By capturing uncertainties, interrelationships and possible time frames for specific AI trajectories, it is possible to identify and build an understanding of potential risks and opportunities.
- **Trend and data analysis** involves collecting, processing and interpreting large amounts of data related to the current state of AI in order to identify patterns, correlations, insights and opportunities. These can then be extrapolated to decipher the prospective AI future landscape.
- **Horizon scanning** means detecting signs of potential changes, such as through considering early signals, trends in adjacent domains, wild cards (low-probability, large-effect events) and matters at the margins of current thinking that challenge past assumptions.
- **Others**. Literature reviews uncover existing knowledge, gaps or controversies. Consultation and polling can help measure public awareness, trust and expectations of AI and its impacts. Empirical evidence can help validate and improve theoretical models and scenarios and philosophical analysis can help better make sense of current developments in the field.

Source: https://www.oecd.org/en/about/programmes/strategic-foresight, (OECD, 2017[4]; Honorof, 2023[5]).

2 AI's potential future benefits

Key messages

- The Expert Group on AI Futures put forth **ten priority benefits** for enhanced policy focus:
  1. accelerated scientific progress;
  2. better economic growth, productivity gains and living standards;
  3. reduced inequality and poverty;
  4. better approaches to urgent and complex issues, including mitigating climate change and advancing other Sustainable Development Goals (SDGs);
  5. better decision-making, sense-making and forecasting;
  6. improved information production and distribution;
  7. better healthcare and education services;
  8. improved job quality;
  9. empowered citizens, civil society and social partners;
  10. improved institutional transparency and governance, instigating monitoring and evaluation.
- Key considerations to better promote AI's potential benefits in AI policies and initiatives include:
  - National and international policy initiatives often recognise the importance of the benefits above, but could take more decisive action to seize them.
  - Considerations of job quality are important, but they are presently often overshadowed by concerns regarding job quantity (OECD, 2023[6]).
  - AI's potential to improve institutional transparency and empower civil society and social partners should be more widely recognised, as should research on concrete use cases.

The Expert Group identified ten priority AI benefits for enhanced policy focus

The Expert Group on AI Futures ("Expert Group") identified 21 potential future AI benefits. Through ranking and synthesis of these, it put forth **ten priority benefits** for enhanced policy focus, many of which are already starting to become visible (see Annex A for methodology). These benefits help to implement the five value-based OECD AI Principles (2024[7]): 1.1) Inclusive growth, sustainable development and well-being; 1.2) Respect for the rule of law, human rights and democratic values, including fairness and privacy; 1.3) Transparency and explainability; 1.4) Robustness, security and safety and 1.5) Accountability.

BENEFIT 1: Accelerated scientific progress

So far, rapid AI advances have led to groundbreaking applications in science

Key areas of progress in AI include robotics, nuclear fusion, drug discovery, digital simulations, antibody generation and protein folding, with AlphaFold as a key example (Azizzadenesheli et al., 2024[8]; Stanford, 2023[9]).[3] However, AI's contribution to science is just beginning. In some areas, the technology may have achieved less than anticipated. For example, some found that AI contributed little to research during the COVID-19 pandemic (OECD, 2023[10]). Furthermore, AI has so far mostly contributed to breakthroughs in a narrow set of natural and physical sciences, while similar transformations in other disciplines, such as social sciences, have progressed less despite high expectations (Manning, Zhu and Horton, 2024[11]).

Accelerated productivity in science could be one of AI's most valuable applications

AI can help scientists become more productive. Through large language model (LLM)-based research assistants, laboratory robots and facilitation of scientific and technological breakthroughs, many expect AI to continue to reshape science and innovation (Franca, 2023[12]; OECD, 2023[10]). AI can be viewed as the "invention of a method of invention" (Griliches, 1957[13]), resulting in AI tools and assistants that can perform an increasing number of research tasks (Bianchini, Müller and Pelletier, 2022[14]).

Expert Group members highlighted that generative AI systems may increasingly be used as components in agentic AI systems that perform a range of functions with growing autonomy and less human involvement, as well as with increasing adaptiveness to evolving conditions (OECD.AI, 2023[15]). Many experts expect that AI will transform scientific progress in the coming decades and help address widespread medical challenges, such as treating cancer and neurodegenerative diseases like Alzheimer's (Gruetzemacher, Paradice and Lee, 2019[16]; OECD, 2023[10]).

BENEFIT 2: Better economic growth, productivity gains and living standards

AI may already be helping to make firms more productive and competitive

Evidence suggests that firms adopting AI tend to be more productive than those that do not, but a causal link is unclear. Part of the reason behind this is that AI adopters already tend to have solid digital adoption and capacities (Calvino and Fontanelli, 2023[17]). AI contributes to economic growth and productivity through the discovery of new ideas and through new efficient and effective means of conducting work (Jones, 2022[18]). Existing LLMs can increase worker productivity but cannot perform many tasks (Dell'Acqua et al., 2023[19]). As with other technologies, there is a time lag between adopting AI and seeing productivity gains materialise as organisations adapt. OECD.AI Trends & Data show global venture capital investments in AI start-ups growing 400% between 2015 and 2022. Demand for professionals with AI skills has more than doubled between 2018 and 2023, and the latest developments in generative AI are only increasing this momentum. Such investments and skill demand today can be expected to impact economic growth tomorrow.

In the future, economic gains from AI-enabled productivity growth could be significant

Predictions for future economic gains from using AI vary, with some estimates ranging from a 1-7% rise in global GDP by 2033 to a speculative ten-fold increase over decades if hypothetical forms of artificial general intelligence (AGI) are created (Acemoglu, 2024[20]; Goldman Sachs, 2023[21]; Russell, 2022[22]).[4]

Accelerated innovation and integration of AI tools into business processes could create new types of jobs, yield productivity gains in almost all sectors and enhance workers' productivity in many tasks currently handled by elite subject matter experts (Autor, 2024[23]; OECD, 2023[24]). Expert Group members highlighted that this may take the form of a "J curve", with an initial drop in productivity due to the costs of integrating and adopting AI, followed by significant growth in the value of these investments (Brynjolfsson, Rock and Syverson, 2020[25]; OECD.AI, 2023[15]). Numerous countries and institutions have highlighted that benefitting from AI-driven economic growth requires investment in complementary assets, such as skills, but also in systemic changes, such as reducing economic and gender inequities and building societal systems that are more resilient to disruption (OECD, 2023[24]; [6]; Anderson and Sutherland, 2024[26]; UNESCO/OECD/IDB, 2022[27]).

BENEFIT 3: Reduced inequality and poverty

So far, the impacts of AI on inequality within countries appear mixed, but evidence points to increased inequality between countries

Among OECD countries, there is no indication to date that AI has affected wage inequality between occupations. Still, evidence suggests that AI may be associated with lower wage inequality *within* occupations (Georgieff, 2024[28]). One explanation is that low performers within an occupation benefit more from using AI, producing an equalising effect. Beyond wage inequalities, some experts and researchers suggest that AI may exacerbate existing digital divides, such as urban-rural divides and the gender digital divide, for example, because of a lack of access to AI education, infrastructure and resources. At the same time, AI may also help address digital inclusion issues, such as by enhancing economic opportunities for remote individuals (Božić, 2023[29]; Gottschalk and Weise, 2023[30]; Bentley et al., 2024[31]).

Internationally, AI developments are highly concentrated in a few countries, raising concerns that AI's benefits are unevenly distributed (OECD, 2024[32]). Advanced economies exhibit higher exposure to AI, given their higher share of employment in high-skilled jobs that are more exposed to AI. Still, their workers are more likely to be complemented rather than replaced by AI, and their economies are more likely to benefit from productivity enhancements as a result. In contrast, lower wages in developing economies could reduce AI adoption incentives and related productivity benefits. AI systems developed in a limited set of countries may not fit the social and institutional context in others, limiting the potential for their deployment and use. Access to important inputs like AI talent and compute are also distributed unequally worldwide, as discussed under Risk 5 on power concentration in Chapter 3.

AI could drive reductions in inequality and poverty through targeted, collective action

Provided equitable distribution of benefits, some experts put forward that everyone could have living standards currently seen as comfortable (UC Berkeley, 2021[33]). Others note that AI can contribute to robust and tailored poverty reduction efforts based on timely and relevant data, help enhance the resilience and efficiency of agricultural activities and assist in improving other conditions that contribute to inequality, such as access to education, health, financial programmes and well-being services (Goralski and Tan, 2022[34]; Javaid et al., 2023[35]; Mhlanga, 2021[36]; OECD, 2019[37]; WEF, 2024[38]). Broadening the reach of AI's benefits depends on successful, targeted collective action. Without this, Expert Group members and others were concerned that AI may increase inequality (Chapter 3, Risk 10) (OECD.AI, 2023[15]; Gates, 2023[39]).

BENEFIT 4: Better approaches to urgent and complex issues, including mitigating climate change and advancing other SDGs

AI can help to achieve most SDGs and is already assisting in combatting climate change

Addressing complex global challenges, such as those in the SDGs, demands advanced capacities to synthesise, understand and act upon large amounts of information and coordinate responses nationally and internationally. Expert consultations have found that various forms of machine learning (ML) and AI could help achieve most of the 169 SDG targets but may hinder the achievement of 69 (Vinuesa et al., 2020[40]). For example, AI is assisting in mitigating climate change impacts through systems that track changes in weather patterns, sea level rise and disaster risk, with policymakers also taking into account the environmental burden that AI itself may pose (Earthna, 2023[41]; OECD.AI, 2023[42]; OECD, 2022[43]).

AI could enable humanity to solve complex challenges and further advance the SDGs

By augmenting humans' capacity to comprehend and execute complex tasks and providing new forms of collaborative tools, AI could play a significant role in anticipating, mitigating and managing the complex challenges and impacts of "megatrends" as they evolve, potentially at a global scale (Haluza and Jungwirth, 2023[44]; US Department of State, 2023[45]). Expert Group members and governments have highlighted that advances in AI could lead to better societal outcomes through innovations in energy systems, climate modelling and many other relevant areas (OECD.AI, 2023[15]; White House, 2024[46]). AI could also support better international co-operation, which is necessary to address global challenges, by enabling the testing of new mechanisms to incentivise and facilitate information sharing and enabling better monitoring and evaluation of signatories' adherence to agreements (Clarke and Whittlestone, 2022[47]).

BENEFIT 5: Better decision-making, sense-making and forecasting

AI tools have already started to assist decision-making in several areas

LLMs can support individual reasoning, and evidence shows real-world benefits from AI-assisted decision-making (Brynjolfsson, Danielle and Raymond, 2023[48]). AI systems can overcome reasoning mistakes and biases by helping humans filter out "noise" and irrelevant influences that can lead to inconsistent and inaccurate decisions (Du, 2023[49]). The potential for AI systems to make data-driven decisions is leading to its adoption across a range of sectors, including within the public sector.[5] Generative AI may be increasingly able to contribute to such decision-making, given its increasing ability to score well on relevant tests (OECD, 2023[50]). However, uncertainties remain regarding the validity of tests used to assess AI performance. As of 2023, machines could not match the reasoning and creative decision-making of top human performers in a variety of contexts, suggesting a need for further improvements and enhancements to realise AI's benefits fully (Koivisto and Grassini, 2023[51]). Uncertainties also remain about the accuracy of machine outputs, with issues such as generative AI "hallucinations" remaining unresolved.

Future AI systems could help to formulate complex decisions and improve predictions

AI systems could generate novel insights, including by extrapolating from past data. Particularly if such AI-enabled extrapolations become longer-range or more abstract, they could provide firms and individuals with automated decision assistance and rapidly generated probability estimates in different domains. This could look like weather forecasting applied to political, socioeconomic or environmental developments over longer periods. Concerning generative AI systems in particular, Expert Group members noted that such systems are starting to be embedded in autonomous AI agents in increasingly complex tasks. Through human-machine collaboration, such systems could assist in guiding optimal decisions and acting as research assistants and advisors (Horvitz, 2014[52]; Russell, 2019[53]). Firms could use generative AI systems for targeted strategy consulting to make sense of numerous simultaneous and complex changes in the global economy. These systems could provide a range of services, in tasks ranging from counselling to personal assistance, making everyday life easier by supporting household decisions, such as financial planning. The availability of such services could have positive spillover effects across markets, such as reducing consumer search costs and increasing competition among providers. Such increased competition may then spur further productivity and economic growth. Policymakers could use AI to support political decisions via policy enactment simulations to assess probabilities of accomplishing desired outcomes and inform adaptations and improvement in the review of policies.

BENEFIT 6: Improved information production and distribution

AI systems already enable new forms of data collection

This includes automatically detecting and identifying items in images, audio recordings or video. Rapid progress has occurred in the capabilities and prevalence of AI-enabled sensing devices, allowing automatic speech transcription, motion detection, live image recognition and a wide range of tasks that previously required human labour (Zhang, Wang and Lee, 2023[54]; OECD, 2023[10]).

AI-enabled sensors could facilitate new and expanded forms of data access and sharing

Simultaneous advances in several forms of novel data gathering, such as satellite imagery, may result in new forms of geospatial intelligence and global modelling (Sathyaraj et al., 2024[55]). Autonomous systems embedded in all types of devices —such as drones, self-driving vehicles and other robotic and remote sensing instrumentation integrated with AI— also produce new data-gathering capabilities. AI can also help to maximise the quality and utility of data, as well as humans' and machines' ability to process and analyse it (Jarrahi et al., 2023[56]). Expert Group members emphasised, though, that data collection should be targeted and deliberate to mitigate risks of invasive surveillance (Chapter 3, Risk 7) and harms related to the environmental footprint that the storage and processing of such data may leave (OECD, 2022[43]).

BENEFIT 7: Better healthcare and education services

AI systems are increasingly capable of offering personalised services

Smart assistants and AI customer service agents demonstrate the potential for AI-driven services. In healthcare, AI systems provide real-time data and insights about patients. They can save lives by detecting anomalies and facilitating preventative action, though currently to a limited degree. AI systems have shown promise in trials for medical decision-making, though implementation is early (Vasey, 2022[57]). AI for education also has great potential, but adoption in education systems has been slow compared to other sectors, and past educational technologies, including those based on ML, have sometimes failed to deliver (Barnum, 2023[58]). As of 2023, most AI efforts in education involve guidance related to using LLMs in the classroom, and most AI uses in this field did not take advantage of the latest generative approaches and were limited to providing recommendations, diagnostics or feedback (Fadel et al., 2024[59]).

Future AI systems could enable enhanced and personalised services at scale

Leveraging AI, people could have access to "Everything as a Service" (EaaS) or suites of agents that can help them achieve tasks through human-machine collaboration (Russell, 2019[53]). Healthcare and education are particularly promising areas of focus, with experts in a recent survey expressing positive views on AI's potential impact in these areas (Rainie and Anderson, 2024[60]). In healthcare, for example, AI could result in tailored and preventative interventions and informed behavioural "nudges" leading to better outcomes, help health professionals provide more time for care and yield new techniques to unlock value from vast health data assets – 97% of which remain untapped in OECD countries (Sumner et al., 2023[61]; Bennett Institute, 2024[62]; OECD, 2024[63]). AI advances could also help ease projected healthcare workforce shortages of 3.5 million by 2030 in OECD countries (OECD, 2024[63]).

AI could lower barriers to entry in education, especially in low- and middle-income countries. By increasing accessibility and lowering the price of knowledge acquisition, AI could open up a new supply of skilled labour (Fan and Qiang, 2023[64]; WEF, 2023[65]; Demaidi, 2023[66]). AI could help democratise autonomous learning by providing support tailored to the needs of individuals through personalised tutoring, including for those with special needs (Bond, 2023[67]). Student outcomes could be enhanced by redefining how, where and what students learn (OECD, 2023[68]; Fadel et al., 2024[59]; Fariani, Junus and Santoso, 2023[69]). Developing tailored tutoring AI systems currently represents an engineering challenge, but simpler tools like chatbots, AI-generated teaching content and teacher and student support tools can be expected at an increased scale in the near-term (Huang and Rust, 2021[70]; Flavián and Casaló, 2021[71]).

BENEFIT 8: Improved job quality

AI can already improve employees' income and performance

AI already has positive impacts on job quality, often by automating some dangerous or tedious tasks, thereby improving workers' well-being (OECD, 2023[6]). Nearly two-thirds of workers surveyed by the OECD reported that AI improved their enjoyment of work. As an example of AI in action, wearable AI devices and AI sensors already provide real-time assessments of high-risk movements or situations that threaten workers' safety in the automotive industry (Hart, 2023[72]). Employers have used the resulting insights to protect the physical safety of workers, improve factory conditions and prevent accidents. However, caution must be exercised to ensure such uses of AI do not impose adverse effects on workers.

Monotonous or dangerous tasks could be further performed by AI systems

In a recent survey of hundreds of experts across fields, 77% said that AI will positively impact people's day-to-day work activities by 2040 (Rainie and Anderson, 2024[60]). However, opinion polls among workers indicate that 32% of those in ICT and 14% of those in hospitality, services and arts industries expect AI to benefit them more than it will hurt them (Kechhar, 2023[73]). This demonstrates that the expected impact of AI on work is not equally distributed.

AI could contribute to positive psychological effects in the workplace if less-stimulating tasks are allocated to AI systems and workers can devote their time to more fulfilling pursuits (Jia et al., 2024[74]). In the public sector, for instance, AI can reduce the time public officials invest in monotonous tasks (OECD, 2024[75]). Expert Group members noted that historically, general-purpose technologies have resulted in new, high-quality jobs in novel fields, which may contribute to both job quality and quantity (OECD.AI, 2023[15]). In industries such as construction and manufacturing, humans could complete dangerous tasks remotely with AI-enabled robotics and high-quality simulations could reduce accidents.

BENEFIT 9: Empowered citizens, civil society and social partners

AI is opening up unprecedented mechanisms for data analysis and public engagement

In one example, a civil society organisation (CSOs) used an AI system to audit public expenses and "in a week revealed more suspicious claims than what the responsible governmental agency did in a year" (Savaget, Chiarini and Evans, 2019[76]). CSOs such as charities are increasingly aware of the potential of AI: in 2024, 61% of surveyed charities in the United Kingdom (UK) reported using AI daily, nearly doubling the result from 2023 (Legraien, 2024[77]). However, fewer than 25% of respondents felt prepared to respond to AI opportunities and challenges. In the workplace, social partners such as trade unions and business associations bargain and build social dialogue on the use of AI, but they also use AI to inform workers of their rights and better understand their experiences (OECD, 2023[24]). Despite the recent uptake of AI and increased interest by other AI actors in supporting opportunities for such organisations, their widespread use of AI has often been limited by funding, skills and operational constraints (Savaget, Chiarini and Evans, 2019[76]; Government of Denmark, 2024[78]; OECD, 2023[24]).

AI could help expand the scale and scope of oversight and representation activities

AI can bolster CSOs and social partners by facilitating new forms of digital social services and constituent engagement capacities (Sanchez, 2021[79]). Expert Group members highlighted that tools such as AI assistants could help a broad range of organisations, including grassroots and community organisations, to undertake more complex tasks or scale up their operations (OECD.AI, 2023[80]). Synergies enabled through transparent and accessible institutions may further this benefit (Benefit 10).

BENEFIT 10: Improved institutional transparency and governance, instigating monitoring and evaluation

So far, AI has been used to oversee public programmes and engage with the public

AI has already been applied to public sector datasets to identify and manage corruption risks and promote integrity and efficiency (Ugale and Hall, 2024[81]). AI supports institutional communications and helps facilitate participatory exercises.[6] It also builds institutional capacity within governments to more effectively monitor, enforce and evaluate policies, reducing burdens and improving policy effectiveness for government and businesses. Many governments have taken steps to ensure their algorithms and AI use are transparent and accountable (OECD, 2023[82]). AI-enabled technologies can offer alternative channels for governments to communicate with citizens and provide efficient tools to gather and analyse societal perspectives (OECD, 2023[83]; 2022[84]). Some governments have successfully trialled the use of AI for purposes of preference aggregation, mass deliberation and consensus brokering (Tsai et al., 2024[85]).

AI could change societal norms and expectations regarding institutional transparency

AI's ability to sort, filter, and summarise vast amounts of information could lower barriers to disclosure, making it easier for governments to be transparent to the public. It could also help citizens understand complex governmental processes and open up opportunities for broader public engagement and scrutiny by civil society organisations. This could accelerate institutional reforms, with AI tools assisting civil society organisations in tasks such as monitoring and evaluation and better targeting of services, information dissemination, and advocacy materials (Efthymiou, Alevizos and Sidiropoulos, 2023[86]). Such advancements can strengthen trust in governments and reinforce democracy (OECD, 2022[87]).

Policy efforts recognise potential future benefits, but gaps may exist

An OECD review of AI policy efforts[7] found that they often highlight the importance of AI's potential to accelerate scientific progress, advance economic growth and productivity gains and raise living standards. To a lesser extent, they acknowledge AI's potential in addressing complex and accelerating issues; assisting decision-making, sense-making and forecasting and powering beneficial services. Yet, recognition of positive impacts on job *quality* tends to be limited and overshadowed by issues of job *quantity* and associated topics of training and capacity building (OECD, 2023[24]). Recognition of benefits for organisational transparency and empowering civil society seems light, except regarding opening government data to support training AI systems and providing transparency in automated decisions (e.g., explaining public benefits determinations). Non-exhaustive examples of actions include:

- **Scientific progress.** The European Union (EU) AI Act (2024[88]) provides exemptions from certain rules and limits copyright protections for AI systems used for scientific research. France has developed an AI for Science, Science for AI Centre (AISSAI).[8] The UK invested EUR 117 million (equivalent) to create AI research hubs in relevant areas (UKRI, 2024[89]). The US AI Executive Order on the Safe, Secure and Trustworthy Development and Use of AI (AI EO) (2023[90]) includes requirements to build foundation models for applied science and support healthcare AI research.[9]
- **Growth, productivity and increased well-being.** Many national AI initiatives provide funding for growth and productivity investments, including promoting job transition and creation. For example, Germany budgeted EUR 5 billion to implement its national strategy by 2025 (OECD, 2024[1]).

- **Reduced inequality and poverty.** The EU AI Act features measures to promote diversity, non-discrimination, fairness and accessibility. Outcomes of the first AI Safety Summit hosted in 2023 involve the UK and global partners making investments to boost AI efforts to accelerate development in emerging and developing economies (UK FCDO, 2023[91]).
- **Approaches to complex and accelerating issues.** The Frontier AI Safety Commitments signed by 16 AI companies pledge to develop AI systems to help address global challenges (UK DSIT, 2024[92]). UN agencies' efforts can be seen in the International Telecommunication Union (ITU)'s AI for Good webinar series, which helps AI actors to connect and identify AI solutions, and UNESCO's support of an International Research Centre on AI (IRCAI) at the Jožef Stefan Institute.[10] The EU AI Act includes allowances for personal data processing in sandboxes for certain use cases, including addressing green transition and climate change. The UK's AI research hubs include a focus on the environment and power efficiency. The US AI EO includes requirements for using AI to enable the provision of clean electric power, developing foundation models that streamline environmental reviews while improving outcomes and encouraging private companies and academia to develop AI tools to mitigate and adapt to climate change.
- **Decision-making, sense-making and forecasting.** Many national AI initiatives include actions for using AI to achieve this benefit in a variety of areas. For example, Finland's (2023[93]) legislation on automated decision-making and Israel's (2023[94]) public call to identify companies to use AI to support public decision-making processes. However, efforts focus more on decision-making than on sense-making and forecasting. As related to the previous benefit, AI is being used to inform decision-making to manage climate change, such as through digital twins to simulate real-time energy grid management to forecast and optimise energy consumption (OECD.AI, 2022[95]).
- **Improved information production and dissemination.** Many governments have initiatives to produce and enhance the value and re-usability of public information and data, including through the use of AI (OECD, 2023[96]; 2024[75]). The US National AI Resource (2024[97]) aims to connect US researchers to the computational, data and training resources needed to advance AI research**.**
- **Beneficial AI services.** Many national AI efforts focus on enhanced digital services. For example, the US (2023[98]) lists more than 700 government AI use cases, many of which are digital services. The UK (2023[99]) has made significant financial investments in AI medical services in particular. Several countries are pursuing AI in education, including via hackathons and funding investments for creating AI tools (OECD, 2023[68]; UK DfE, 2023[100]). However, Expert Group members found public investment in and use of AI in education to be low relative to its positive potential.
- **Improved job quality.** Efforts here focus more on mitigating AI workplace harms than using AI. The amended EU Directive 2002/14EC (2002[101]) and a number of national policies, such as the US AI EO, seek to mitigate job quality harms that could be caused by AI through actions including promoting employee engagement on AI and counteracting workplace surveillance (European Parliament, 2021[102]; Bell and Korinek, 2024[103]). Intergovernmental organisations are also active in conducting relevant analyses, such as the OECD's AI in Work, Innovation, Productivity and Skills (AI-WIPS) programme and efforts of the International Labour Organisation (ILO).
- **Empowered civil society and social partners.** The US National AI Resource (2024[97]) and the UK AI Research Resource (2024[89]) seek to ensure that AI research resources are broadly accessible. The OECD Employment Outlook (2023[24]) looks at efforts related to social partners.
- **Improved institutional transparency.** Several initiatives seek to use AI for public engagement, such as governments' use of the open-source crowd engagement AI application Polis (Computational Democracy, 2023[104]). Some initiatives also aim to make the public sector's use of AI more transparent, such as the UK (2024[105]) Algorithmic Transparency Recording Standard.

3 Potential future AI risks

Key messages

- The OECD Expert Group on AI Futures put forth **ten priority risks** for enhanced policy focus:
  1. facilitation of increasingly sophisticated malicious cyber activity;
  2. manipulation, disinformation, fraud and resulting harms to democracy and social cohesion;
  3. races to develop and deploy AI systems cause harms due to a lack of sufficient investment in AI safety and trustworthiness;
  4. unexpected harms result from inadequate methods to align AI system objectives with human stakeholders' preferences and values;
  5. power is concentrated in a small number of companies or countries;
  6. minor to serious AI incidents and disasters occur in critical systems;
  7. invasive surveillance and privacy infringement;
  8. governance mechanisms and institutions unable to keep up with rapid AI evolutions;
  9. AI systems lacking sufficient explainability and interpretability erode accountability;
  10. exacerbated inequality or poverty within or between countries.
- Key considerations to better mitigate AI's potential risks in AI policies and initiatives include:
  - The creation of national, multilateral and regional bodies for AI safety science, R&D and testing reflects the attention devoted to many of these potential risks.
  - Even more agile and flexible approaches and networks may be needed to keep up with the pace of AI advancements.
  - Effective methods are still needed to ensure AI systems' objectives are aligned with human stakeholders' preferences and values.
  - Competitive "race" dynamics between companies and countries and significant market concentration in key AI sectors merit stronger policy focus.

The Expert Group identified ten priority AI risks for enhanced policy focus

The Expert Group on AI Futures ("Expert Group") identified 38 potential future AI risks. Through ranking and synthesis of these, it put forth **ten priority risks** for enhanced policy focus, many of which are starting to become visible (see Annex A for methodology). While the benefits discussed in Chapter 2 facilitate trustworthy AI, as embodied in the OECD AI Principles, the priority risks represent potential hindrances.

Although ten priority risks are highlighted, others also deserve attention. Notably, the risks that generated the most disagreement among experts (red in Annex B, Figure B.2) may warrant further discussion. There were particularly diverging views on the likelihood and consequences of humans losing control of artificial general intelligence (AGI) systems—an issue central to debated AI catastrophic risk scenarios. Such variance suggests the need for deeper investigation, with Expert Group members agreeing that its outputs should discuss diverging views rather than require consensus (OECD.AI, 2023[106]). AGI refers to hypothetical future AI systems with human-level or greater intelligence across a broad spectrum of contexts

(OECD, 2024[107]). There is substantial debate and uncertainty amongst experts about if or when such systems might be developed and even whether the milestone is well-defined. However, the development of AGI is the goal of several AI companies (Altman, 2023[108]; Deepmind, 2024[109]).

RISK 1: Facilitation of increasingly sophisticated malicious cyber activity

AI systems are already increasing the incidence and severity of malicious cyber activity

Although many efforts involve using AI to *mitigate* cybersecurity risks, AI systems have reduced the level of effort needed for malicious cyber activity that would have previously required significant time investment by human experts (UK DSIT, 2023[110]). Large language models' (LLMs) ability to generate software "exploits" is the focus of significant attention (Klimek, 2023[111]; Maraju, Rashu and Sagi, 2024[112]). Generative AI is estimated to have contributed to an 8% increase in cyberattacks over the first half of 2023, notably in the form of phishing emails, keystroke monitoring malware and ransomware (Mascellino, 2023[113]). Evidence suggests that several state-based actors are using LLMs to pursue new cyberattack approaches, including identifying vulnerabilities and assisting with generating content for phishing campaigns (OpenAI, 2024[114]). While current advanced AI systems can help execute basic cyberattacks, they do not appear capable of sophisticated, multi-step autonomous attacks (UK DSIT, 2024[115]).

Future AI-facilitated malicious cyber activity could disrupt critical

For years, researchers have cautioned that AI-enabled infrastructure hacking and ransomware attacks are likely to increase and evolve. AI tools can increase the capacity of malicious actors to inflict damage, particularly by lowering the skill and cost required to execute malicious cyber activity that can destabilise societies and cause virtual or physical harm (Fassihi, 2023[116]; Brundage et al., 2018[117]). For example, AI systems could facilitate malicious cyber activity to infiltrate nuclear or healthcare facilities, energy infrastructure or other critical digital systems (Gerstein and Leidy, 2024[118]; Puwal, 2024[119]). AI could lead to novel forms of malicious cyber activity by subverting AI-integrated systems, such as hacking autonomous vehicles to cause them to crash or tampering with medical images to generate false cancer detection positives (Brundage et al., 2018[117]; Yamin et al., 2021[120]). Like other digital systems, AI systems could be the target of malicious cyber activity (NIST, 2024[121]). Experts raised concerns that the value of the online ecosystem, digital technologies and digitally-enabled infrastructure could erode if security does not keep pace with expanding, evolving threats (OECD.AI, 2023[15]; Pupillo et al., 2021[122]).

RISK 2: Manipulation, disinformation, fraud and resulting harms to democracy and social cohesion

AI is already beginning to amplify disinformation and online manipulation of people through the production and dissemination of convincing synthetic content

While AI has significant potential to help fight false and misleading content online, AI-enabled mis- and disinformation is a top concern for governments (OECD.AI, 2022[123]; OECD, 2023[124]).[11] Existing AI systems can produce convincing disinformation, with numerous cases of AI-generated disinformation on sensitive political issues reaching wide viewership (Bontcheva et al., 2024[125]). Evidence is mixed with regard to how well humans can identify AI-generated text (OECD, 2024[126]; Casal and Kessler, 2023[127]). Regarding images, humans have been shown to perceive artificially generated faces as more "real" than actual faces (Nightingale and Farid, 2022[128]). Many experts have raised the potential for AI-enabled mis- and dis-information as a significant threat to electoral systems, but evidence to date suggests related impacts have been limited (Simon, McBride and Altay, 2024[129]).

Future AI-enabled manipulation and disinformation could affect information ecosystems

Future AI systems could amplify the scale and severity of mis- and disinformation and help scale up fraud and scams like spear phishing (UK DSIT, 2024[115]). Sophisticated models could help to precisely tailor and broadly deploy messages to individuals based on psychological profiles via "personalised persuasion" (Matz et al., 2024[130]). "Compositional deepfakes" could help craft credible yet unreal narratives and make it increasingly difficult to distinguish fact from fiction (Horvitz, 2022[131]). Some believe AI systems could facilitate mass manipulation of people, criminal coercion, such as automated blackmail, and fraud, such as through digital impersonation (Horvitz, 2022[132]; Khan, 2023[133]; Fletcher, Tzani and Ioannou, 2024[134]). These risks could be exacerbated with anthropomorphised AI systems—made to seem human (Deshpande et al., 2023[135]). A related challenge is information pollution, where the increased prevalence of AI-generated outputs leads to decreased quality of information online, contributing to misinformation (Vincent, 2023[136]; Lorenz, Perset and Berryhill, 2023[137]). At societal scales, AI-enabled mis- and disinformation is a fundamental threat to the information ecosystem and the fact-based exchange of information that underpins science and democracy (Ognyanova et al., 2020[138]; OECD, 2022[139]). These issues could cause lasting damage to social cohesion, democratic principles and human rights (OECD, 2024[140]). However, there is uncertainty about this future risk, with some experts arguing that fears about AI-enabled mis- and disinformation may be overblown (Simon, Altay and Mercier, 2024[141]).

RISK 3: Races to develop and deploy AI systems cause harms due to a lack of sufficient investment in AI safety and trustworthiness

Rapid product releases and subsequent issues suggest race dynamics are underway

Effective competition in providing AI services is likely to be important in ensuring consumers and economies fully benefit from the technology (OECD, 2024[142]). However, some experts suggest that this pressure—along with unclear liability and accountability allocations and regulatory requirements—may contribute to an underemphasis on AI ethics and safety (Chow and Perrigo, 2023[143]; Li, 2023[144]). Since late 2022, AI companies have rapidly released new or enhanced products, which were sometimes found to exhibit significant shortcomings afterwards. Relatedly, pressure among some firms and individuals to adopt and use such systems as soon as they are available could exacerbate risks (Clarke and Whittlestone, 2022[47]). In addition to companies, competitive pressures may cause race dynamics between countries (NSCIA, 2021[145]; de Neufville and Baum, 2021[146]; Bremmer and Suleyman, 2023[147]).

Further competitive dynamics could promote the rapid development and deployment of AI systems without sufficient efforts to ensure they are trustworthy and safe

Race dynamics may increase the risks of AI incidents, and some experts believe that, in the absence of effective governance and regulation, short-term gains could come at the expense of long-term societal goals (Hendrycks, Mazeika and Woodside, 2023[148]). Regarding races between *companies*, damage could be caused by premature deployment of products without sufficient safety measures to beat competitors to the market. This is especially the case where clear first-mover advantages and "winner-takes-all" dynamics (Askell et al., 2019[149]; Vipra and Myers West, 2023[150]) exist. Due to possible trade-offs, some experts suggest companies may be compelled to prioritise performance and speed over safety or to dedicate significant resources, such as compute and talent, to developing products without commensurate resources devoted to safety and trustworthiness (Askell et al., 2019[149]; Leike, 2022[151]; Kahn, 2024[152]). Some experts also caution that AI race dynamics between *countries* could escalate international conflict, perhaps inadvertently, and cause tensions in international co-operation on intellectual property, regulation or international governance approaches (Johnson, 2020[153]; Garfinkel, 2019[154]; Roberts et al., 2024[155]).

RISK 4: Unexpected harms result from inadequate methods to align AI system objectives with human stakeholders' preferences and values

Today, some experts view AI misalignment as a foundational, unresolved issue

AI system objectives can be explicit or implicit. It can be challenging for the developer or user of an AI system to specify explicit objectives in a manner that ensures the system implements them in a way that aligns with the human's intent. For instance, spelling out the user's true aims can often be too difficult or result in an inefficient system (Gabriel, 2020[156]). Thus, the objectives programmed into the AI system are often high-level, conceptual or proxy objectives. This can result in unanticipated consequences. The degree of difference between an AI system's actions in seeking to achieve the explicit objective and the intents or values of humans is sometimes referred to as misalignment.

Some experts suggest that some kinds of AI may develop goals of self-preservation, self-improvement and resource acquisition (Bales, D'Alessandro and Kirk-Giannini, 2024[157]). Another issue is generalisation, or "ensuring that the outputs translate from their training contexts to the real world as intended" (UK DSIT, 2024[115]). Ensuring that AI systems act in accordance with user intent or in line with shared human values (insofar as these are identifiable) has been called an "unsolved problem" by some experts (Hendrycks et al., 2022[158]). Evidence of AI misalignment can be observed today in reward hacking, where an AI system finds unforeseen and potentially harmful ways of achieving a goal (Skalse et al., 2022[159]). While a few methods to increase alignment exist, such as using specific types of human feedback when training AI models, they generally have limited scalability and can introduce new biases (Casper et al., 2023[160]). Some fields of research, including those sometimes labelled as AI alignment, investigate ways to verifiably ensure the behaviour of AI systems aligns with human preferences (Russell, 2019[53]; Dung, 2023[161]; Bekenova et al., 2022[162]). However, other researchers suggest that alignment should focus on norms and constraints related to the system's function rather than on preferences (Zhi-Xuan et al., 2024[163]).

AI misalignment harms may grow as AI systems are more widely deployed

Some experts believe that this issue could result in AI systems that act in ways that undermine underlying, implicit human interests, which could escalate as AI becomes more deeply integrated into societies and economies. Misaligned AI systems pursuing human-defined objectives in unanticipated and undesirable ways could have impacts that range from harmful bias and harmful recommendations all the way to potentially catastrophic consequences (Skalse et al., 2022[159]; Russell, 2019[53]; OECD.AI, 2023[164]).

Expert Group members expected generative AI systems to be increasingly used as components in AI agents that perform a wide range of functions with increasing levels of autonomy (OECD.AI, 2023[15]). These increasing levels of autonomy and integration into complex sequences of tasks with reduced human oversight could increase the importance of alignment.

RISK 5: Power is concentrated in a small number of companies or countries

Foundation models and key AI inputs are already concentrated among dominant players

Concentration of market power with AI is often related to control over access to resources, including data and compute needed to train advanced AI models (Buchanan, 2020[165]). Demand for compute in particular has grown dramatically (OECD, 2023[166]). The ability to obtain compute, influenced by availability and costs, influences which organisations can build AI systems, the types of systems that get built and who benefits from the systems and profits that accrue from provisioning them (Vipra and Myers West, 2023[150]).[12] This is further complicated by vulnerabilities in the semiconductor supply chain (Haramboure et al., 2023[167]). Industry practices like "bundling" together products and services drive self-reinforcing network effects whereby having more users and data, in turn, helps to further improve AI models (UC

Berkeley, 2021[33]; OECD, 2024[142]). Relatedly, mergers, acquisitions, strategic investments and partnerships, as well as vertical integration in which a firm operates at multiple levels of the value chain, can hinder competition and make it difficult to survive for companies without such integration or ability to partner with major players (UK CMA, 2023[168]; 2024[169]). Open-source AI models can help democratise access, though there is debate among experts with regard whether open sourcing certain highly capable AI models could pose risks that outweigh the benefits (Mozilla, 2023[170]; Seger et al., 2023[171]).

In the future, some firms and countries could wield even more power in AI markets based on their market dominance, financial resources and technological capabilities

A lack of access to AI resources and structural factors such as economies of scale, first-mover advantages, acquisitions and partnerships pose risks to competition in AI markets (OECD, 2024[142]). If market power were concentrated in the hands of one or a few dominant firms, they could secure major economic gains, potentially at the expense of smaller firms, governments and academic institutions that lack the resources to catch up (Vipra and Myers West, 2023[150]; Cockburn, Henderson and Stern, 2018[172]). This concentration in market or economic power could shift dynamics regarding political power, with the potential for incumbents to play a disproportionate role in influencing policy (Bettelle, 2023[173]; AI Now Institute, 2023[174]). This could have particularly negative impacts for developing and emerging economies.

Expert Group members cautioned that if few providers dominate AI markets, society may become dependent on them (OECD.AI, 2023[15]). This could place decisions about key infrastructure and services under the control of a limited number of providers, risking a lack of meaningful choice or democratic oversight over essentially corporate decisions. If governments are unable to build internal capacity and expertise and promote effective competition, some suggested that inequality could worsen (see Risk 10). They also suggested that policymakers should create approaches to help ensure that developers' decisions align with the public interest and that there is a healthy, competitive market for AI products and services. If this risk were taken further, some experts argue that those with market control over key AI systems or ecosystems could use this advantage to strengthen political power, potentially facilitating wide-scale subjugation and/or authoritarianism, especially with regard to use by state-based actors (Funk, Shahbaz and Vesteinsson, 2023[175]; Clarke and Whittlestone, 2022[47]; Dizikes, 2023[176]).

RISK 6: Minor to serious AI incidents and disasters occur in critical systems

AI is increasingly deployed in a variety of critical systems

These include air traffic control, financial, nuclear and military systems (Laplante et al., 2020[177]; Zwetsloot and Dafoe, 2019[178]). A failure in a critical sector can cause cascading effects, as observed in the 2010 stock market "flash crash" where issues related to high-frequency trading bots caused a drop in the Dow Jones Industrial Average by 9% in a few minutes (Makhija, Chacko and Kukreja, 2024[179]). The chair of the United States Securities and Exchange Commission (SEC) cautioned that a financial crash is likely as AI is further adopted to manage financial markets unless regulation is put in place (Carter, 2023[180]).

In the future, failures of AI in critical systems could cause major harms

AI is expected to be increasingly integrated into critical systems, with potentially severe risks if they prove unreliable in unanticipated ways or if improperly assured systems are used (Laplante et al., 2020[177]; OECD.AI, 2022[181]). AI systems in consumer products, such as airplanes or autonomous vehicles, may also be hazardous if poorly designed or implemented (CSET, 2021[182]). The magnitude of such risks is likely to increase as AI systems become more complex and prevalent (Bianchi, Cercas Curry and Hovy, 2023[183]). AI supply chains involve multiple layers of dependency, with downstream providers often reliant upon foundation model providers for the functioning of their AI systems. This dependency could introduce correlated failures across different systems if an issue arises in the underlying foundation model or with the model provider, potentially jeopardising critical infrastructure and services (OECD.AI, 2023[15]).

RISK 7: Invasive surveillance and privacy infringement

AI-enabled surveillance is already being leveraged with negative consequences

The Carnegie Endowment for International Peace (2022[184]) found that 97 of 179 (54%) countries analysed are using AI technologies for public surveillance.[13] AI systems have been used to make sensitive inferences, such as sexual orientation, political preferences, income and potential future criminality (Wang and Kosinski, 2018[185]; Kosinski, 2021[186]; Staab et al., 2023[187]; Stark and Hutson, 2021[188]). Such uses can erode privacy, fuel automated discrimination and suppress political opposition. The misuse of biometrics amplifies this risk. Facial recognition biases are often mentioned, yet identification through characteristics like demeanour, walking style or heartbeat is also possible (Privacy Ticker, 2019[189]; Hambling, 2019[190]). Civil society organisations and other groups have called for a ban on AI-enabled mass surveillance.[14] The EU AI Act (2024[88]) includes the prohibition of most forms of real-time biometric identification in public spaces. Still, actions are needed to bridge AI, data and privacy communities and find ways to promote innovation and seize AI's benefits while protecting privacy rights (OECD, 2024[191]).

In the future, AI could enable invasive surveillance at scales not previously feasible

In a recent survey of hundreds of experts across fields, 79% said that AI will have a negative impact on people's privacy by 2040, a concern shared by the general public (Rainie and Anderson, 2024[60]; Fazlioglu, 2024[192]). Increasing data collection and AI capabilities can allow countries to upgrade surveillance capabilities. In the future, some suggest that machine-listening abilities could reach a level where they could simultaneously understand all conversations or monitor a vast number of CCTV cameras (Russell, 2019[53]; Brumfiel, 2023[193]). AI-enabled surveillance of employees could also undermine labour rights and job quality, including by suppressing collective bargaining (OECD, 2023[24]; Scott, 2024[194]). The misuse of biometrics, such as facial recognition systems, could limit freedom of expression and assembly, amplify discrimination and result in targeted manipulation of individuals or groups (Access Now, 2021[195]; AI.gov, 2022[196]; Latif et al., 2022[197]; OECD, 2022[198]). More generally, Expert Group members noted that AI could facilitate the scaling up of enforcement of laws, rules and policies, with associated benefits and risks.

RISK 8: Governance mechanisms and institutions unable to keep up with rapid AI evolutions

The pace of and uncertainties related to AI development present novel challenges

Some experts suggest contributing factors include novel challenges and dilemmas brought about by the use of AI, financial and information asymmetries between technology developers and governments that hinder effective regulatory responses, lobbying efforts to maintain the status-quo and market incentives that favour rapid technological advancement without appropriate governance considerations (Clarke and Whittlestone, 2022[47]; Metz, 2023[199]). Additionally, the inherent difficulty of predicting and addressing the societal impacts of new technologies until they are well-developed and widely adopted, often referred to as the Collingridge Dilemma, adds further complexity (OECD, 2020[200]). Governmental complexities can also contribute to this challenge. For instance, general-purpose AI models may challenge siloed governance structures, which can feature organisations with competition or uncertainty in mandates or necessitate complex governance arrangements to align actors for efficient and effective regulatory policy making (OECD, forthcoming[201]). The availability of in-house public sector capacities and tools may also be insufficient for effective policymaking and enforcement that maximises benefits and minimises costs.

The inability to keep up with AI advancements could lead to inadequate governance

Advancements in AI systems are occurring at a rapid pace, making proactive policy action increasingly difficult. This is further complicated by the increasing technical expertise and capabilities needed to govern AI systems. Societies might take a long time to understand transformative changes and implement desired courses of action, making it difficult to influence present and future developments and further compounding challenges over time (Grallet and Pons, 2023[202]). Policymakers will need to consider, though, that an overly precautionary approach can stifle innovation (Draghi, 2024[203]), with Expert Group members finding that improper regulatory models and delays in obtaining real-world value from AI to also represent potential future risks, albeit at a lower level of importance (see Annex B). Proactive approaches to address this risk will need to be flexible and well-informed.

RISK 9: AI systems lacking sufficient explainability and interpretability erode accountability

Leading AI systems based on deep learning are still very difficult to understand

Systems based on deep learning are “black boxes”, meaning that it is difficult to describe how they produce a given output. This makes it hard to detect and mitigate harmful biases and produces challenges in determining accountability when issues arise. The field of interpretable and explainable AI has gained attention and support in recent years (see Chapter 4), including by many standard-setting bodies, but technical challenges remain (Gao and Guan, 2023[204]).

In the future, black box AI systems could undertake important societal functions, eroding understanding and accountability

As AI systems become increasingly integrated into economic and societal functions, black box systems could exacerbate other AI risks. For instance, it is difficult to determine whether black box AI systems are aligned to human stakeholders’ preferences and which preferences are embedded in a system (Christian, 2020[205]). Organisations and individuals could overly rely on and have a false sense of trust in seemingly efficient yet potentially flawed AI systems (Russell, 2019[53]). Because flaws may be unobservable, the risk of AI incidents and the perpetuation of harmful bias may increase (OECD.AI, 2022[206]). The issue can erode the accountability of AI actors and disempower the public by limiting their ability to make informed decisions or potentially making them subject to opaque, flawed AI-driven decisions (Lima et al., 2022[207]).

RISK 10: Exacerbated inequality or poverty within or between countries

Little evidence shows negative labour demand impacts, but bias is well documented

As of 2023, there is little evidence of AI-induced negative impacts on labour demand, but this may be because AI adoption remains low (OECD, 2023[24]). Other evidence suggests, though, that national investments in AI may be associated with higher levels of domestic income inequality (Cornelli and Frost, 2023[208]). Further, AI systems perpetuating harmful biases, introducing inequities and producing unequal gains have been observed across a range of sectors (Bender et al., 2021[209]; NIST, 2022[210]; Larsson, White and Bogusz, 2024[211]). Bias can take the form of often-discussed issues of data and algorithmic bias, but it can also be seen in ways that AI usage intersects with institutional and social biases, both human (errors in human thinking) and systemic (practices that advantage some groups over others), which are often overlooked (NIST, 2022[210]). As related to data bias, Expert Group members and others have cautioned that digital data divides—where some groups are more represented in data than others—limit the potential for AI benefits, such as personalised AI services, leaving them only useful and accurate for data-rich populations (UNESCO, 2019[212]; Perry and Lee, 2019[213]; Dieterle, Dede and Walker, 2024[214]).

AI could increase social, economic and digital divides and block development pathways

In a recent survey of hundreds of experts across fields, 70% said that AI will have a negative impact on wealth inequalities by 2040 (Rainie and Anderson, 2024[60]), a concern shared by the general public (Modhvadia, 2023[215]). If future gains from AI accrue inequitably, some suggest it could drive up inequality and displace workers without generating equivalent new opportunities (Acemoglu and Restrepo, 2020[216]). Around 27% of jobs today are in occupations at high risk of automation (OECD, 2023[24]). This could become more severe as AI becomes more ubiquitous, with some experts believing that inequitably accrued benefits and harms may worsen inequality in the short-term and fundamentally alter wage and employment levels by displacing labour in the long-term (Bell and Korinek, 2024[103]). The OECD (2023[24]) found that while AI is capable of automating non-routine tasks, its future impacts on labour demand are ambiguous, depending on the balance between the displacement of human labour by AI, the increase in labour demand because of the greater productivity AI brings and the creation of new jobs caused by AI adoption.

Beyond labour markets, automated discrimination, if unmitigated, could unfairly prevent access to goods and services, such as housing and jobs, for some individuals and groups. Emerging and developing economies may be particularly disadvantaged, as advanced economies leading the AI transformation may be better able to absorb change, automation could result in "reshoring" investments for outsourced work and inexpensive labour could be exploited to support AI advancements, such as screening harmful and extreme content to help add safeguards to AI systems (Grallet and Pons, 2023[202]; IMF, 2024[217]; Muggah, 2023[218]). Expert Group members and others note that efforts to mitigate AI harms centre on advanced economies and on issues that already receive attention from policymakers and media, potentially leaving their resulting solutions poorly matched for other contexts and underrepresented populations and languages further behind (Muggah, 2023[218]; OECD.AI, 2023[15]).

Policy efforts could help manage future risks, but some gaps may exist

An OECD review of AI policy efforts[7] found that they often highlight the importance of AI risks related to facilitating increasingly sophisticated malicious cyber activity; AI evolving too fast for governance to keep up; a lack of sufficient explainability and interpretability; manipulation, disinformation and fraud; invasive surveillance and privacy infringement and exacerbating inequality or poverty. There is less recognition of harms resulting from inadequate AI alignment methods, at least under some conceptions of alignment, but this appears to be increasing. While some policy efforts recognise the risk of competitive race dynamics among countries, there appears to be less recognition of such dynamics among companies. Similarly, recognition of power concentration tends to focus more on power held by countries than by companies, beyond related but less direct calls to enhance competition and support for AI startups. However, the recent US Executive Order on the Safe, Secure, and Trustworthy Development and Use of AI (AI EO) (2023[90]), includes concrete language on this.

Specific actions to mitigate these risks tend to be less prevalent than the aforementioned recognition that they are important. However, governments are increasingly transitioning from principles to actionable policy and governance. The G7 (2023[219]) Hiroshima AI Process Code of Conduct for Organisations Developing Advanced AI Systems includes cross-cutting commitments that may help address priority risks. The growing number of national, multilateral and regional bodies for AI safety, science, R&D and testing can also provide cross-cutting actions to mitigate many of these risks. Examples include AI safety institutes and units in several countries (see Chapter 4, Policy Action 6). Additional examples include:

- **Facilitation of increasingly sophisticated malicious cyber activity.** The Frontier AI Safety Commitments put forth by the governments of the United Kingdom (UK) and Korea have been signed by 16 global AI companies, pledging investments in cybersecurity and incentives for third-party discovery and reporting of vulnerabilities (UK DSIT, 2024[92]). The EU AI Act imposes robustness and cybersecurity requirements for high-risk AI systems, while taking into account underlying digital infrastructure. It also requires more stringent obligations for general-purpose AI systems that could pose a "systemic risk", such as those that could contribute to disruptions in critical sectors or democratic processes or lower barriers to entry to offensive cyber capabilities. A global coalition of 18 countries issued joint AI security guidelines (UK NCSC, 2023[220]). The US AI EO calls for best practices for managing cyber risks in specific areas, such as financial institutions, and a pilot to use AI to discover and remediate vulnerabilities in government digital systems.

- **Manipulation, disinformation, fraud and harms to democracy and social cohesion.** The Frontier AI Safety Commitments include pledges to deploy mechanisms that enable users to understand if content is AI-generated and to prioritise research on societal risks posed by advanced AI systems. The EU AI Act imposes detection and disclosure obligations on providers of very large online platforms and search engines to assess systemic risks related to AI-generated content, including disinformation and threats to democratic processes. It also mandates that the outputs of AI systems are recognisable as such. The US AI EO requires the establishment of guidance for the federal government on content authentication and labelling. The OECD Hub on Information Integrity is working with governments to promote access to information that helps enable individuals to be exposed to a variety of ideas, make informed choices and exercise their rights.[15]
- **Races to develop and deploy advanced AI systems.** The EU AI Act and US AI EO place controls on key AI inputs and reporting requirements for AI systems trained above certain compute thresholds. The EU AI Act may also mitigate races deploying products by instituting a range of requirements to protect EU citizens from potential harms. The US (2023[221]) has also put forth voluntary commitments for technology companies to promote safe, secure and transparent development and use of AI. Several items touched on in Chapter 4 are also relevant.
- **Inadequate AI alignment methods.** The EU AI Act imposes stringent obligations for general-purpose AI systems that could pose a systemic risk. The language indicates that international approaches have identified the need to pay attention to unintended issues of control relating to alignment with human intent, among others, when considering this.
- **Power concentration**. Most efforts focus on market power. The US AI EO describes "addressing risks from dominant firms' use of key assets" to disadvantage competitors and providing opportunities for small and medium-sized enterprises (SMEs) among its policies and principles. The EU (2024[222]) has a variety of support tools for SMEs. Some governments are facilitating access to AI resources, such as compute and hardware. Examples include public sector-provided supercomputer access for SMEs in Serbia and similar plans in the EU (Lomas, 2024[223]), a variety of national semiconductor projects and funding programmes and the provision of regulatory sandboxes (Vipra and Myers West, 2023[150]). Some governments, such as France (2024[224]), are investing in open-source AI. Emerging regulatory approaches, such as investigating cloud computing concentrations in the US (2023[225]), could be useful tools. Competition authorities have a critical role in promoting fair competition in AI markets, as evidenced by a joint G7 communiqué and a joint statement by the EU, UK and US on AI competition issues.[16] The OECD has work focused on market concentration and power,[17] finding that ensuring the right balance on enforcement and reforms that provide governments with a holistic view of the market is preferable to dramatic reforms in competition law or crude regulatory solutions (2018[226]).
- **AI incidents and disasters in critical systems.** The EU AI Act classifies the use of AI in critical infrastructure as "high-risk", which imposes regulatory obligations and incident reporting. It also requires stringent obligations for general-purpose AI systems that could pose a systemic risk by making it easier to interfere with critical infrastructure. The UK has classified data centres as critical national infrastructure to boost protections from cyber criminals (2024[227]). The OECD has developed an AI Incidents Monitor (AIM) and is working through its Expert Group on AI Incidents to develop a common definition for AI incidents and related terminology (OECD, 2023[228];

forthcoming[229]; 2024[230]). The IEEE (2024[231]) Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems represents a step toward mitigating incidents.

- **Invasive surveillance and privacy infringement.** Many national or multilateral entities seek to address privacy risks from AI-enabled data collection, such as through the EU General Data Protection Regulation (GDPR) (EP, 2020[232]) and the US AI EO. The Council of Europe's (CoE) Framework Convention on AI and Human Rights, Democracy and the Rule of Law ("Framework Convention on AI") (2024[233]) includes a variety of requirements related to privacy and personal data protection. The EU AI Act classifies all remote biometric identification systems to be high-risk and subject to strict requirements, with its use in public spaces prohibited beyond some law enforcement exemptions. The high-risk designation also applies to the use of AI to monitor workers' performance and behaviours. It also bans social scoring, emotion recognition systems in workplaces and educational institutions and biometric categorisation systems that deduce or infer things like political opinions, race or sexual orientation. The OECD recently created an Expert Group on AI, Data and Privacy to further explore this area (OECD.AI, 2024[234]).
- **Governance mechanisms and institutions unable to keep up with rapid AI evolutions.** The UK hosting the first AI Safety Summit was noted as being prompted by this challenge (Stacey and Milmo, 2023[235]). Instruments that provide for experimentation and flexibility are increasing, such as sandboxes in Colombia, Estonia, France, Norway, Spain and others; as well as efforts to put in place agile governance processes in Colombia and Japan.[18] The OECD (2021[236]) Recommendation for Agile Regulatory Governance to Harness Innovation includes relevant commitments agreed to by OECD countries, with relevant concepts and tools also embedded in the OECD (2024[237]) Framework for Anticipatory Governance of Emerging Technologies. The CoE Framework Convention on AI calls upon each party to enable, as appropriate, the establishment of controlled environments for developing, experimenting and testing AI systems. Some countries also have initiatives to boost in-house public sector AI capacities. This includes, for instance, dedicated training programmes in Colombia (2019[238]) or courses available globally through Elements of AI (2024[239]); specialised AI recruitment, such as US National AI Talent Surge (2024[240]) and the creation and staffing of AI and data labs throughout the German federal government (OECD, 2024[1]). Finally, efforts to strengthen governmental strategic foresight efforts, as discussed in Chapter 1, can help to address this risk.
- **AI systems lack explainability and interpretability.** Several policies include a right to an explanation for AI outputs. The EU AI Act includes requirements for transparency and explainability of certain types of AI systems and calls for the development of sandboxes to facilitate explainable AI. The US Defense Advanced Research Projects Agency (DARPA) conducted research and published an Explainable AI (XAI) toolkit.[19] Many national AI initiatives include requirements for transparency, traceability and explainability (Nannini, Balayn and Smith, 2023[241]).[20]
- **Exacerbated inequality or poverty.** The CoE Framework Convention on AI includes requirements related to equality and non-discrimination. The US secured voluntary commitments from leading AI companies to ensure AI does not promote harmful bias and discrimination. During the first AI Safety Summit, the UK and several partners pledged GBP 80 million as part of a development programme focused on Africa to combat inequality and boost prosperity (UK FCDO, 2023[91]).

4 Priority policy actions

Key messages

- The OECD Expert Group on AI Futures put forth **ten priority policy actions** to encourage the obtention of future AI benefits and mitigation of risks:
  1. establish clearer rules, including on liability, for AI harms;
  2. consider approaches to restrict or prevent certain “red line” AI uses;
  3. require or promote the disclosure of key information about some types of AI systems;
  4. ensure risk management procedures are followed throughout the lifecycle of AI systems that may pose a high risk;
  5. mitigate competitive race dynamics in AI development and deployment that could limit fair competition and result in harms;
  6. invest in research on AI safety and trustworthiness approaches, including AI alignment, capability evaluations, interpretability, explainability and transparency;
  7. facilitate educational, retraining and reskilling opportunities to help address labour market disruptions and the growing need for AI skills;
  8. empower stakeholders and society to help build trust and reinforce democracy;
  9. mitigate excessive power concentration;
  10. take targeted actions to advance specific future AI benefits.
- Policymakers could consider how best to pursue priority policy actions when developing or reviewing national AI strategies, policies and initiatives.
- There has been a significant increase in policy initiatives, especially in recent months, and several of these have been implemented successfully in some governments.

The Expert Group identified ten priority policy actions

Policymakers can consider multiple options to nurture future AI benefits while mitigating potential risks. Taken together, a portfolio of approaches could provide “defence-in-depth,” whereby a comprehensive set of approaches rather than a single policy helps secure AI benefits and reduce AI risks (NIST, 2024[242]).

The Expert Group on AI Futures (“Expert Group”) identified 66 potential policy approaches to obtain AI benefits and mitigate risks. Through ranking and synthesis of these, it put forth **ten priority policy actions** for enhanced focus (see Annex A for methodology). The actions complement and build upon the five recommendations for policymakers contained in the OECD AI Principles (2024[7]): 2.1) Investing in AI R&D, 2.2) Fostering an inclusive AI-enabling ecosystem, 2.3) Shaping an enabling interoperable governance and policy environment for AI, 2.4) Building human capacity and preparing for labour market transformation and 2.5) International co-operation for trustworthy AI.

Each priority policy action and examples of relevant policy efforts are discussed below. Overall, AI policy efforts[7] often reflect the importance of carrying out all ten priority items, though specific action plans to implement them are less frequent.

POLICY ACTION 1: Establish clearer rules, including on liability, for AI harms

Clear rules promote AI accountability and adoption by removing uncertainty

The concept of harm is central to AI standards and safety and regulations. The OECD (2024[230]) definition of AI incidents and hazards includes the following harms:

- injury or harm to the health of a person or groups of people;
- disruption of the management and operation of critical infrastructure;
- violations of human rights or a breach of obligations under the applicable law intended to protect labour and intellectual property rights;
- harm to property, communities or the environment.

The definition is not intended to address in what way an AI system or its sub-components may be related to or responsible for any harm that may occur in an AI incident.

AI incidents are being reported at a rapid pace.[21] To address harms from AI incidents, updating or clarifying safety and liability rules and frameworks is viewed as a promising avenue (Narayanan and Potkewitz, 2023[243]; Forum on Information & Democracy, 2024[244]; European Union, 2024[88]).[22] Concern about preventing harms and addressing liability for AI-caused damage is a key barrier to AI adoption by European Union (EU) businesses (OECD, 2023[245]). Clear safety and liability rules could improve predictability and consistency. However, they may require complex technical analysis to determine the root of problems. To determine liability, AI actors must consider several factors and different parties' roles in the AI system value chain. Parties include AI developers; operators of AI systems, such as companies using AI-enabled products; suppliers and sellers of AI products; end-users; rightsholders and other actors along the value chain (EC, 2022[246]). Establishing clear liability could also involve new legal agreements among actors or new interpretations of liability provisions in existing contracts (Kumar and Nagle, 2020[247]; Villasenor, 2019[248]). Once clear liability rules are established, Expert Group members emphasised that effective enforcement mechanisms will be critical (OECD.AI, 2023[249]).

Recent and emerging public policy efforts

The currently adopted revision of the EU Product Liability Directive[23] and planned EU AI Liability Directive (2022[250]) may make bringing claims for AI-caused harm easier and establish a clear cause-and-effect relationship between actions and damage attribution (Bollans, 2023[251]). In the US, companies may be subject to enforcement action by the Federal Trade Commission (FTC) under the FTC Act (FTC, 2024[252]; DiResta and Sherman, 2023[253]),[24] with the US Executive Order on the Safe, Secure, and Trustworthy Development and Use of AI (AI EO) encouraging the FTC to consider using its authority to ensure consumers and workers are protected from AI harms (2023[90]).

POLICY ACTION 2: Consider approaches to restrict or prevent certain “red line” AI uses

Red lines can help demarcate and enforce limits regarding unacceptable uses of AI

To promote trustworthy AI, some experts and civil society organisations have called for “red lines” against uses of AI that may fail to respect human rights or privacy rights. Some of the red lines asserted by these experts or organisations include mass surveillance, monitoring public spaces, exacerbating discrimination and manipulating human behaviour (EDRi, 2021[254]; Janjeva et al., 2023[255]). Others include using AI systems that conduct malicious cyber activity, self-replicate autonomously, provide advice on biological or chemical weapons, defame real people, facilitate large-scale influence operations or select and engage attack targets autonomously (Russell, 2024[256]; Bengio, 2024[257]; UN, 2024[258]). The use of lethal autonomous weapons systems (LAWS) represents a contentious issue among countries and a potential red line. LAWS could change warfare, with some already being used today (Trager and Luca, 2022[259]). In identifying red lines, policymakers and other AI actors can reflect on and articulate the potential positive uses and impacts of AI to help identify those not aligned with pro-societal outcomes.

Recent and emerging public policy efforts

The EU AI Act (2024[88]) prohibits certain uses of AI systems, including biometric identification in public spaces, with some law enforcement exemptions; unfair commercial practices; social scoring; emotion recognition systems in workplaces and educational institutions and biometric categorisation systems that infer characteristics like political opinions, race or sexual orientation. The Council of Europe's (CoE) Framework Convention on AI Human Rights, Democracy and the Rule of Law ("Framework Convention on AI") (2024[233]) requires that each party assess the need for a moratorium or ban on AI use cases that it considers incompatible with respect for human rights, the functioning of democracy or the rule of law. Canada's proposed AI and Data Act (AIDA) (2023[260]) would create new criminal legal provisions for AI-specific offences. The US FTC has taken enforcement action related to some uses of AI for facial recognition and robocalls (FTC, 2023[261]; Swenson, 2024[262]). The Frontier AI Safety Commitments (2024[92]) pledge to set thresholds and risk mitigation measures for AI risks that, unless adequately mitigated, would be deemed intolerable. While not red lines per se, efforts to identify unreasonable or intolerable risks may assist in identifying potential red lines. These efforts are complemented by international dialogues between industry and academia on AI risks and red lines, such as the International Dialogues on AI Safety (IDAIS).[25]

Although clear red lines may be useful, building international consensus on where to draw them has proven very challenging. For example, the United Nations (UN) Convention on Certain Conventional Weapons (CCW) has been discussing the concept of "meaningful human control" of autonomous weapons systems since 2013 (UNODA, 2023[263]). Recent progress includes a declaration signed by more than 50 countries aiming to build international consensus around responsible behaviour and a UN General Assembly resolution to undertake rigorous study and seek broad views from governments and other stakeholders on ways to address challenges raised by LAWS (US Department of State, 2023[264]; UN, 2023[265]).

POLICY ACTION 3: Require or promote the disclosure of key information about some types of AI systems

As AI impacts an increasingly wide range of activities, transparency about the nature and use of AI systems becomes more important. Disclosure requirements or commitments can reduce information asymmetries between providers and users and help users make better decisions, as occurs in other sectors. Disclosures may include model cards containing standardised information on AI models and datasheets documenting training data's characteristics, such as their purpose, composition, intended uses, maintenance and potential harmful biases. Disclosures can detail the AI system developer's safety and responsibility practices and ensure that humans know when and how they interact with an AI system. Tensions may exist between some types of disclosure obligations and commitments and the protection of intellectual property and trade secrets, which will need to be navigated by policymakers and AI actors (Mylly, 2023[266]).

Recent and emerging public policy efforts

The EU AI Act, CoE Framework Convention on AI and national rules in countries like Israel[26] require adopting measures to ensure adequate transparency and oversight tailored to specific AI contexts and risks and that persons interacting with AI systems are, as appropriate for the context, notified as such. The EU AI Act and US AI EO include requirements for disclosing when content is generated by AI, with the latter requiring a process for companies developing certain types of AI systems to report to the government details on model training, testing and data ownership and the labelling of synthetic content. The proposed EU AI Liability Directive has been interpreted to mandate the disclosure of evidence related to systems in which an individual claims damages, such as logs and datasets (Nannini, Balayn and Smith, 2023[241]). Mandatory requirements may not always be necessary, with relevant recommendations, voluntary commitments and standards also emerging. The US National Institute of Standards and Technology (NIST) AI Risk Management Framework (2023[267]) recommends that AI developers and deployers publish information about their AI systems and underlying data, how they are used and identified adverse incidents and outputs. The Frontier AI Safety Commitments pledge external transparency on a variety of AI safety and risk management practices, with the UK also securing commitments from firms developing advanced AI systems to publish their safety policies and grant government access to their models for safety evaluations (UK DSIT, 2023[268]). Voluntary commitments have been established for the responsible development and sharing of AI-generated content, such as the Partnership on AI's (PAI) (2023[269]) Responsible Practices for Synthetic Media, supported by leading technology and media companies.

POLICY ACTION 4: Ensure risk management procedures are followed throughout the lifecycle of AI systems that may pose a high risk

For some AI systems, the context of their development or use may pose a higher risk. This can relate to their scale (seriousness and probability of adverse impact), scope (breadth of application, such as number of individuals affected) or optionality (degree of choice as to whether to be subject to the effects of an AI system) (OECD, 2022[270]). Risk management procedures can help to identify which systems or contexts pose higher risks to mitigate them. Risk management for AI systems that may carry high risks needs to be informed by guidance on which levels of risk are acceptable for different uses and contexts. Risk management is needed both before and after the deployment of AI systems.

Risks should be managed prior to deployment and monitored afterwards

There are a variety of proposed mechanisms for proactively managing risks from AI prior to deployment, many of which are still in the early stages of development. These include risk management frameworks, impact assessment methodologies, go-no-go policies, protective actions to mitigate major risks, evaluation and response procedures and accountability processes. Such approaches should take into account risks related to AI systems' limitations and capabilities, as well as contexts of use. For advanced AI systems in particular, another example is "responsible scaling policies" (RSPs), which commit to actions based on risk assessment of AI system capabilities. When identifying potentially dangerous capabilities, RSPs often set thresholds that trigger actions to slow or cease development. Companies developing advanced AI systems such as Anthropic (2023[271]) and OpenAI (2023[272]) have developed and are using RSPs. Other potential schemes include licensing regimes for developers of models that may carry high risks, sandboxes to test AI systems in controlled environments and stronger protections against theft and misuse of models (Malgieri and Pasquale, 2024[273]; OECD, 2023[274]; Navo et al., 2023[275]). After deployment, it is important for AI actors to continue monitoring system behaviours to determine what expected or unexpected risks may be materialising, and many frameworks provide for such monitoring.

Market deployment is a common target for regulatory action

The EU's product safety regulation is a case in point for regulatory action focused primarily on development and market deployment. Some believe that developers of systems that may carry high risks should have to prove their systems do not pose societal risks—a "safety case" regime where such systems could only be put on the market after verification (Anderljung and Korinek, 2024[276]). However, this could create barriers to market entry that require further consideration. Developers' internal policies could help achieve this within companies, but some experts underscore the need for a third-party ecosystem of evaluators and auditors that can assess risk independently (Groves, 2024[277]; Raji et al., 2022[278]). There is some guidance to assist the safe deployment of foundation models,[9] including by PAI (2023[279]). Some others suggest tiered or structured access around deployment based on risk levels, capabilities or end-user competency (Seger et al., 2023[171]; Shevlane, 2022[280]). Others propose approaches to correct deployed systems retroactively, such as product recall-type provisions if safety concerns arise or "deployment correction" incident response practices (Tartaro, 2023[281]; O'Brien, 2023[282]).

Recent and emerging public policy efforts

The OECD (2023[283]) Guidelines for Multinational Enterprises on Responsible Business Conduct, which sets out expectations for businesses in identifying and addressing AI-related harms, among other things, and the G7's Hiroshima AI Process International Guiding Principles for Advanced AI Systems and Code of Conduct, can set baseline standards to manage risks (2023[284]; [219]). The OECD.AI Catalogue of Tools and Metrics for Trustworthy AI aggregates hundreds of resources that can help AI actors build and deploy trustworthy AI systems. The OECD (2023[285]) has also mapped common guideposts to promote interoperability in risk management, including CoE's draft Human Rights, Democracy and Rule of Law Impact Assessment (HUDERIA),[27] relevant standards by ISO and IEEE, and NIST's Risk Management Framework (2023[267]). The ISO/IEC 42001:2023 standard provides a framework for managing risk and opportunities.[28] The EU (through the EU AI Act), the UK (2024[286]) and the US (through its AI EO) have introduced reporting and oversight requirements for developers of certain systems regarding model training and safety. The EU AI Act also mandates a number of requirements for high-risk AI systems, including on risk management, and requests technical standards to be developed by European Standardisation Organisations. Companies signing the Frontier AI Safety Commitments pledge to assess and manage risks of AI systems, not develop or deploy AI models if the risks cannot be sufficiently mitigated, and set thresholds and monitoring mechanisms to identify and mitigate intolerable risks. The OECD conducted a public consultation on AI risk thresholds to inform future work on this topic.[29] The UK and US secured commitments from leading AI developers to share advanced AI models with the government for testing before market deployment.[30] Testing and assessment bodies and national, multilateral or regional bodies, such as safety institutes (see Policy Action 6), are increasingly playing a role in facilitating risk management, including through building testing and assessment ecosystems.

POLICY ACTION 5: Mitigate competitive race dynamics in AI development and deployment that could limit fair competition and result in harms

Effective competition in providing AI services will be important in ensuring consumers and economies fully benefit from AI (OECD, 2024[142]). However, some experts have raised concerns that unmitigated races between companies and countries to lead in AI development could result in an insufficient focus on ensuring governance approaches are in place to promote trustworthiness in AI systems. To address these dynamics, Expert Group members favoured increased international collaboration, with many members endorsing further development and implementation of international good practice principles, norms and standards; international AI oversight authorities, as akin to an International Atomic Energy Agency (IAEA) for AI; international expert panels on AI, as akin to an Intergovernmental Panel on Climate Change (IPCC) for AI and joint scientific projects, as akin to a European Organisation for Nuclear Research (CERN) for AI (OECD.AI, 2023[15]). They also suggested fostering a global ecosystem of AI, data and governance experts to work with regulators, and to consider the development of additional AI treaties. Such international AI governance efforts should include representatives and experts from emerging and developing economies. Corporate governance principles, extensive red-teaming and public investments in trustworthy AI were also seen as ways to help mitigate these dynamics. Efforts to promote fair competition in AI markets are also key, as discussed below under Policy Action 9.

Recent and emerging public policy efforts

The CoE Framework Convention on AI demonstrates the potential for binding international treaties. Inspired by the IPCC, the UK spearheaded an International Scientific Report on the Safety of Advanced AI, with an interim report published at the Seoul AI Summit in May 2024 ([115]) and a final report to be presented at France's AI Action Summit in February 2025. The OECD and UN announced an enhanced collaboration on global AI governance,[31] with the UN High-level Advisory Body on AI (2024[258]) recommending an IPCC-type independent scientific panel for AI with support from UN agencies, the OECD and its Global Partnership on AI (GPAI) and other international institutions. Risk management efforts under Policy Action 4 and the international network of AI Safety Institutes under Policy Action 6 may also help.

POLICY ACTION 6: Invest in research on AI safety and trustworthiness approaches, including AI alignment, capability evaluations, interpretability, explainability and transparency

"AI safety" is a broad term that encompasses different legal, technical, procedural and educational approaches to prevent AI-related harms. Of highest priority to Expert Group members are approaches focusing on:

- Alignment of AI systems with human stakeholders' values and preferences. AI alignment is a growing field of research with the goal to ensure that AI systems' behaviour reliably aligns with the intents and values of designers, users, and other stakeholders. Existing alignment methods, such as reinforcement learning from human feedback (RLHF), are limited in their ability to scale and can introduce new harmful biases, which calls for further research (Ji et al., 2024[287]; Casper et al., 2023[160]). The risks from inadequate AI alignment methods are discussed in more detail in Chapter 3, Risk 4.
- Assessments, evaluations, and assurance processes for capabilities that can lead to dangerous uses and seek to develop methods to assess those capabilities of AI systems (OECD.AI, 2023[288]). These processes include impact and risk assessments and algorithmic audits covering the AI system lifecycle. They can incorporate benchmarking to allow systems to be compared, red teaming, incident reporting, and other transparency measures (Brennan, 2023[289]; US NSTC, 2023[290]; Ji, 2023[291]; OECD, 2023[228]).
- Robustness research seeks to improve the ability of AI systems to withstand or overcome adverse conditions, as related to errors and the intentional exploitation of model vulnerabilities (OECD, 2024[7]; Tocchetti et al., 2022[292]).

Policy approaches, such as funding for research and development and incentives, can encourage progress on AI safety challenges. Commitments or requirements for AI developers and other AI actors can also encourage investment in relevant research.

Policy instruments and investments in AI safety could help mitigate AI risks that were not rated as high priority by the Expert Group as a whole but were pressing for some. In particular, some experts argued that they do not believe existing risk management activities and policy efforts adequately consider the potential risk of humans losing control over hypothetical future misaligned artificial general intelligence (AGI) systems (Bengio, 2024[257]; Cass-Beggs, 2024[293]; Faggella, 2024[294]; Taylor, 2023[295]),[4] which they see as an extreme potential outcome of insufficient AI safety measures, especially those for AI alignment. They argue that, regardless of testing, developers and deployers of some types of advanced AI systems have such a limited understanding of their system's capabilities or how they may react to novel scenarios that they cannot guarantee their safety. This "loss of control" is a subject of disagreement among Expert Group members and the broader AI community. Expert Group discussions suggested that this may be due to diverging views about the premise of AGI in general. Notably, though, the level of priority accorded to the risks posed by the absence of methods to ensure alignment (Risk 4), alongside the prioritisation of investments in AI safety, indicate pathways for collaboration on the issue, despite underlying disagreements about AGI.

In addition, further investments in **explainable, interpretable and transparent AI** may also be helpful.[32] Progress here could help reduce harmful bias and other harms by making AI systems' decision processes more visible, thereby allowing AI actors and users to make or request corrections. Greater interpretability could also help make AI systems more truthful and reduce the occurrence of so-called hallucinations (Evans et al., 2021[296]; Sahoo et al., 2024[297]). Advancements in these areas could also help detect machine behaviours and understand rationales for systems generating outputs that are not aligned with user intent or broader human values.

Recent and emerging public policy efforts

The EU AI Act, by setting rules for high-risk AI systems and general-purpose AI models, the G7 Hiroshima Process and the Safety Summits started by the UK solidified international focus on **AI safety**. The first AI Summit, held by the UK in November 2023, resulted in the Bletchley Declaration signed by 28 countries and the EU. It recognised that "risks may arise from potential intentional misuse or unintended issues of control relating to alignment with human intent" and committed to the safe development of AI (UK DSIT, 2023[298]). At the follow-on 2024 Summit, Korea and the UK secured Frontier AI Safety Commitments from leading AI companies, including pledges to implement AI safety and transparency measures. The next summit will take place in France in early 2025.[33]

The EU, Japan, Singapore, the UK, and the US have each launched an AI Safety Institute or unit,[34] with these and six additional countries deciding to form an international network of institutes (UK DSIT, 2024[299]). NIST, which houses the US institute, also runs an AI Safety Consortium with more than 200 organisations working toward AI safety. The EU AI Act imposes obligations for high-risk AI systems, including on robustness and cybersecurity and general-purpose AI models, including for general-purpose AI models that could pose a systemic risk. Several governments have developed frameworks or requirements for AI impact assessments, audits and transparency, including Canada, Denmark, France, Mexico, the UK, Uruguay, the US and a joint initiative among five European Supreme Audit Institutions.[35] In addition, the US AI EO launched an initiative to create guidance and benchmarks for evaluating and auditing AI capabilities, focusing on those that could cause harm.

Many government initiatives have focused on **interpretability and explainability,** such as those in the OECD.AI Database of National AI Policies & Strategies.[36]

POLICY ACTION 7: Facilitate educational, retraining and reskilling opportunities to help address labour market disruptions and the growing need for AI skills

The OECD (2023[24]) anticipates that the skills needed to develop, adopt, and use AI will become more important and that existing public policy efforts in many countries will be insufficient. Acquiring skills in AI development requires a combination of formal higher education and on-the-job learning. In contrast, AI adoption and use can require a range of AI literacy skills, with further efforts needed to define and measure these skills and identify education and training requirements. Training, among other activities, should be provided to higher-skilled workers, managers, and vulnerable groups to enable AI adoption and promote equity. OECD efforts have also found that AI is expanding the set of jobs at risk of automation, with some experts noting that further focus is needed on educational and retraining efforts to address potential risks of long-term structural unemployment and help workers adapt and move into new roles (UC Berkeley, 2021[33]). Collective bargaining and social dialogue are important to facilitate the AI transition for both workers and employers (OECD, 2023[24]).

Recent and emerging public policy efforts

Many national initiatives call for AI reskilling for public servants, such as the US AI EO's requirements for AI hiring and training across the federal government and the UK's (2024[300]) online courses on generative AI. Other initiatives focus on upskilling society, such as the globally available Elements of AI course (2024[239]). AI is increasingly being introduced into school curricula for children, youth and teachers, such as through Australia's (2023[301]) Framework for Generative AI in Schools and Croatia's "AI – From Concept To Implementation" initiative (2024[302]). The EU AI Act specifically includes a legal provision on AI literacy for providers and developers of AI systems. The *OECD Employment Outlook* (2023[68]) discusses actions related to collective bargaining and social dialogue regarding AI.

POLICY ACTION 8: Empower stakeholders and society to help build trust and reinforce democracy

Engaging diverse stakeholders early in the technology development cycle enriches the understanding of issues and fosters trust. It helps align technological innovation with societal needs (OECD, 2024[303]), and public engagement on AI influences how well harms are mitigated (UK Government Office for Science, 2023[304]). Engagement can help empower stakeholders, including the public, and build trust in government and its ability to lead AI policy and governance (OECD, 2023[83]). Declining trust in public institutions impedes the government's ability to address pressing challenges, and in 2023, 44% of people had no or low trust in their national government (OECD, 2024[305]). Engagement activities should include stakeholders from emerging and developing economies, as AI solutions are deployed on a global scale. Collective bargaining and social dialogue also have an important role to play in supporting and empowering workers in the AI transition (OECD, 2023[68]). Efforts to strengthen representation, participation and openness in public life are key to reinforcing democracy (OECD, 2024[306]).

Elections have been a strong area of focus for empowering society and protecting democratic principles, both to counter AI-enabled disinformation and deepfakes and building resilience, as well as the potential to use AI to strengthen electoral processes by pre-empting fraud and disenfranchisement and lowering barriers to entry for underfunded candidates (Eisen et al., 2023[307]). AI could also assist in improving deliberative processes and helping citizens to better connect with political candidates and better understand their policy positions (Schneier, 2023[308]; Panditharatne, Weiner and Kriner, 2023[309]).

Recent and emerging public policy efforts

A flagship OECD *Reinforcing Democracy* initiative,[37] with focus areas including combating disinformation and enhancing representation and participation in public life, calls for new actions in the face of challenges exacerbated by AI. Related transparency measures under Policy Action 3 can help achieve these aims. The *OECD Employment Outlook* (2023[68]) discusses actions related to collective bargaining and social dialogue regarding AI. The EU AI Act includes provisions to strengthen current protections for IP rights and EU Data Act includes provisions that give users more control over how some data they generate are used (EC, 2024[310]). Officials in some countries have conducted extensive public engagement on AI, such as public consultations by the White House Office of Science and Technology Policy (OSTP) in the US when developing the Blueprint for an AI Bill of Rights (2022[311]) and to obtain views on AI threats and opportunities (2023[312]).

POLICY ACTION 9: Mitigate excessive power concentration

The development and use of AI can centralise market, economic, political or military power in new ways which may not be appropriately managed by existing governance approaches. Expert Group members highlighted the importance of considering, as appropriate, a range of governance tools to address excess power concentrations. These include the potential for new, revised or clarified regulations; strong regulatory enforcement; ensuring competition authorities have sufficient resources and capacities; oversight and tracking of highly capable AI systems and computational capacity; distribution of benefits (e.g., promoting distributed ownership and international technical standards); promotion of AI ecosystems to support market-adapted AI solutions; and provision of AI enablers and resources as digital public goods, such as government-funded compute and data repositories and internationally pooled resources to fund research by universities, nonprofits and researchers in emerging and developing economies. Recent OECD (2024[142]) work emphasises that access to quality data and computing power may be key to developing competitive AI markets and that effective competition in AI markets is key to mitigating entrenched market power.

Recent and emerging public policy efforts

Many countries have initiatives to open and promote the re-use of government data, which can be used to train AI systems (OECD, 2023[96]). More recent efforts also seek to facilitate access to compute power for SMEs and researchers (OECD, 2023[82]; [166]). Efforts also exist to promote market competition,[38] with several competition authorities focusing on AI markets and investing in monitoring and knowledge building (OECD, 2024[142]). This includes efforts by the UK Competition & Markets Authority (CMA) to monitor and evaluate competition issues related to foundation models.[39] In the US, the AI EO requires the government to prioritise certain funding for technical assistance and resources for small businesses to help commercialise AI and includes actions for “addressing risks from dominant firms’ use of key assets” and other recent (2024[313]) policy calls on government agencies to use public procurement as a lever for promoting a competitive AI market. Some governments are undertaking antitrust investigations, such as the US inquiries into AI partnerships among tech firms (FTC, 2024[314]).

POLICY ACTION 10: Targeted actions to advance specific future AI benefits

Expert Group members found that while policy discussions and efforts often acknowledge the potential benefits of AI, actions proposed or underway often do not explicitly aim to achieve these benefits. Instead, they address them more indirectly. Expert Group members recommend that governments take more direct action when focusing on and investing in ways to achieve priority benefits (Chapter 2).

Recent and emerging public policy efforts

Chapter 2 on potential future AI benefits includes examples of recent and emerging policy efforts to capture each priority AI benefit.

Annex A. Item identification and prioritisation methodology note

The identification and prioritisation of potential future artificial intelligence (AI) benefits, risks, and policy actions discussed in this report reflect the views of the OECD Expert Group AI Futures ("Expert Group"), as supported by the OECD Secretariat. The Expert Group was launched in July 2023.[40]

To build an initial knowledge base to inform early discussions of the Expert Group, the OECD conducted an extensive literature review, which surfaced approximately 250 relevant sources of policy, research, expert opinion and philosophical inquiry on potential trajectories and impacts of AI. These sources comprised web pages, blogs, media articles, academic and research reports, books, videos and films, policy documents, briefs, and event proceedings.

The OECD analysed these sources and identified an initial set of 17 potential future benefits, 36 potential risks and 68 potential policy actions. These items were touched on at the first meeting of the Expert Group on 13 July 2023, as well as asynchronously afterwards via email. Expert Group members provided feedback on items through an interactive feedback board.[41] In addition, an open discussion was held on the OECD.AI Policy Observatory website asking, "What do you see as the most significant potential benefits and risks of AI 10+ years from now?".[42]

Based on experts' feedback and the responses to the open discussion, the OECD refined the list of potential future AI benefits, risks and policy actions, including adding some items and consolidating others, to arrive at a final set. The OECD developed a prioritisation survey based on the final set to gauge Expert Group members' subjective opinions on the importance and actionability of each item on a 0-10 scale.[43] The survey also allowed the experts to provide open-ended feedback on the survey questions and design, allowing them to suggest revised or additional areas for analysis. The survey was conducted August-September 2023. The box below provides the instructions for the expert group members. Of the 61 members of the Expert Group at the time, the OECD received responses from 53, for an overall response rate of 87%. The results of the survey are reported in Annex B.

The survey was not designed to be a scientific instrument but rather a gauge of subjective opinions of expert group members regarding potential AI benefits, risks and policy actions based on their own lived experience and expertise in relevant fields.

Informed by the survey results, the OECD and Expert Group worked together to identify characteristics of positive AI futures in September 2023, further elaborated and refined through discussions and communications, with the current characteristics identified in Chapter 1 of this report. Also informed by the survey results, subsequent discussions and communications through February 2024 further prioritised which potential future AI benefits, risks and policy actions that Expert Group members generally agreed to be of the highest priority for policy action. This took into account the views of new members who had joined the Expert Group since the survey was conducted. This resulted in some items being consolidated, streamlined or rephrased when compared to the survey results.

The OECD leveraged the sources and findings uncovered in the literature review, plus additional material identified afterwards, along with the survey results, to draft this report. While the prioritised items indicate Expert Group members' views on the high-level priorities and concepts, they do not necessarily represent a consensus of opinions on the more in-depth discussion under each item or for the discussion on policy actions underway and potential gaps. However, all Expert Group members had an opportunity to review the draft chapters in May 2024 and discuss it at an Expert Group meeting in July 2024, and the OECD incorporated their feedback into the draft. In addition, the draft was reviewed by members of the OECD Working Party on Artificial Intelligence Governance (AIGO) in September-October 2024, and the OECD Secretariat incorporated their feedback.

Box A A.1. Instructions provided to survey participants

The risks, benefits and solutions presented on this survey have been identified through extensive OECD research, as well discussions held in previous events and roundtables.

You will be asked to provide a ranking of each item along two axes:

1. **Importance**. In your opinion, how important is it that governments focus on this item? Your thinking should weigh both the magnitude of potential impacts from a given risk, benefit or policy solution and the probability of these impacts. In assessing impacts, you could, for instance, take into account the level of harm that you perceive for potential risks, the societal or economic good that could be yielded by potential benefits or the magnitude of positive change that could be brought about by implementing potential solutions. Please use the following examples to give a sense of the rating scale:
   - **0**. This is in no way important for governments and the international community.
   - **2**.This is an issue of marginal importance for governments and the international community.
   - **4**. This is a somewhat important issue.
   - **6**. This is an important issue, but not among the most important.
   - **8**. This is among the most important issues for governments and the international community.
   - **10**. This is the most important issue for governments and the international community.
2. **Actionability**. Assuming that political will exists, based on all of the factors that you can think of (e.g., feasibility, level of complexity, ease of implementation, current and perceived future technical ability and financial resources, etc.), what is your opinion on how actionable the item is in terms of the ability of a group of like-minded countries to make a significant impact with regard to mitigating potential risks, yielding potential benefits, and putting in place potential solutions? For this item, it may be useful to ground your thinking in the medium-term (over the next 10-20 years).
   - A rating of **zero (0)** implies that you think there is no meaningful way for governments to mitigate this risk, contribute to seizing this opportunity or effectively implement this solution, even through collective action.
   - Ratings in the middle – **five (5)** – imply that a group of like-minded governments could have some agency in mitigating a risk, seizing an opportunity or effectively implementing solutions, but that their success may be partial, uncertain or require an unusually large commitment of resources or high level of global collaboration.
   - A rating of **ten (10)** implies near certainty of almost entirely mitigating the risk, realizing the benefits or effectively implementing solutions based solely on the actions of like-minded governments.

Your results may be charted in two ways. First, the OECD may plot the responses on a matrix to identify different categories for the items.

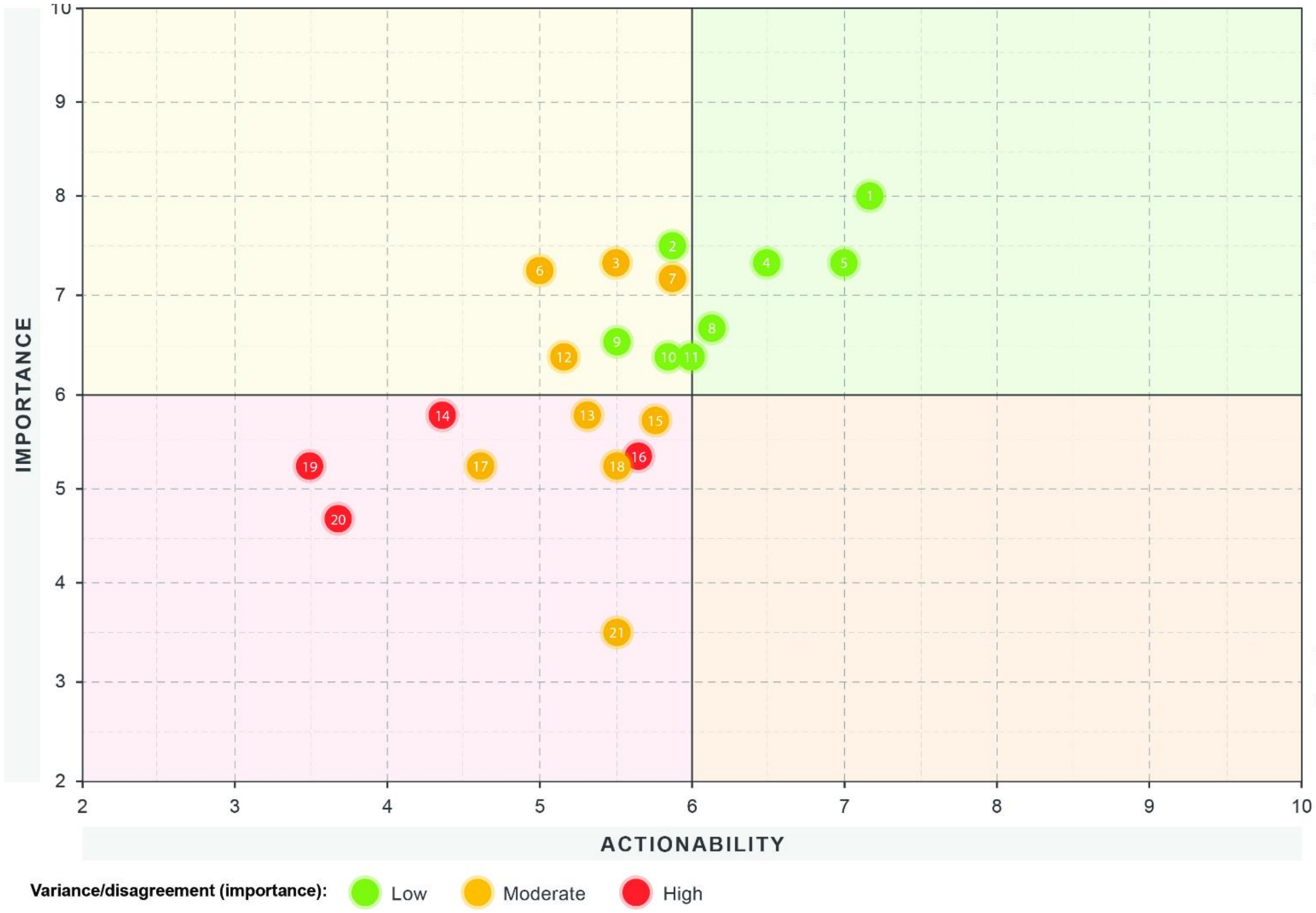
Secondly, the comparison of responses from participants may help surface areas of constructive disagreement or controversy, which could help identify areas ripe for discussion and provide additional context for the matrix.

Your answers will naturally be **subjective** based on your own thoughts and opinions. This is fine, and the future report will be clear about this.

Source: Survey of the OECD Expert Group on AI Futures.

Annex B. Ranking of potential AI benefits, risks and policy actions

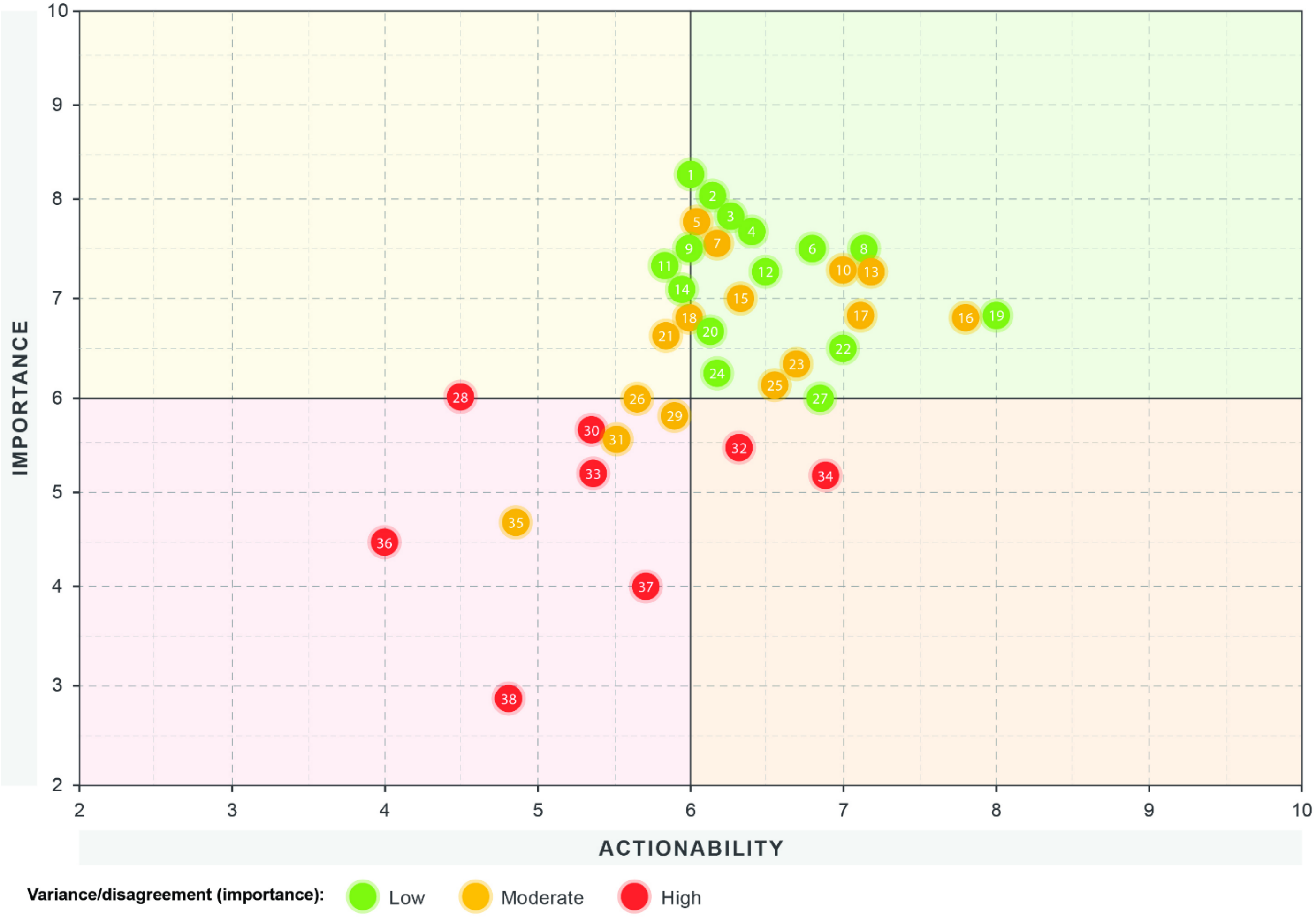
Figure B.1. Experts identified and ranked 21 potential future AI benefits



1. Beneficial scientific progress
2. Economic growth & raised living standards
3. Address complex & accelerating issues
4. Assist decision making, forecasting
5. Beneficial AI services (e.g. healthcare, education)
6. Reduce inequality and/or poverty
7. Transform innovation processes
8. Positive impacts on job quality
9. Improve information production and/or distribution
10. Enhanced civil society capabilities
11. Make institutions more transparent
12. Reduce conflict (detection/deterrence)
13. Job creation - AI creates new tasks
14. Improved international co-operation
15. AI oversight to ensure compliance with agreements
16. Using AI to understand/train other AI
17. Improve multi-agent cooperation (Multiple AI systems coordinate rapidly at scale)
18. Increased creative capacity
19. Reduce conflict (through shared prosperity)
20. Support humans to be more moral/ethical
21. Building/populating VR environments.

Note: See Annex A for details on the source and rating scale.

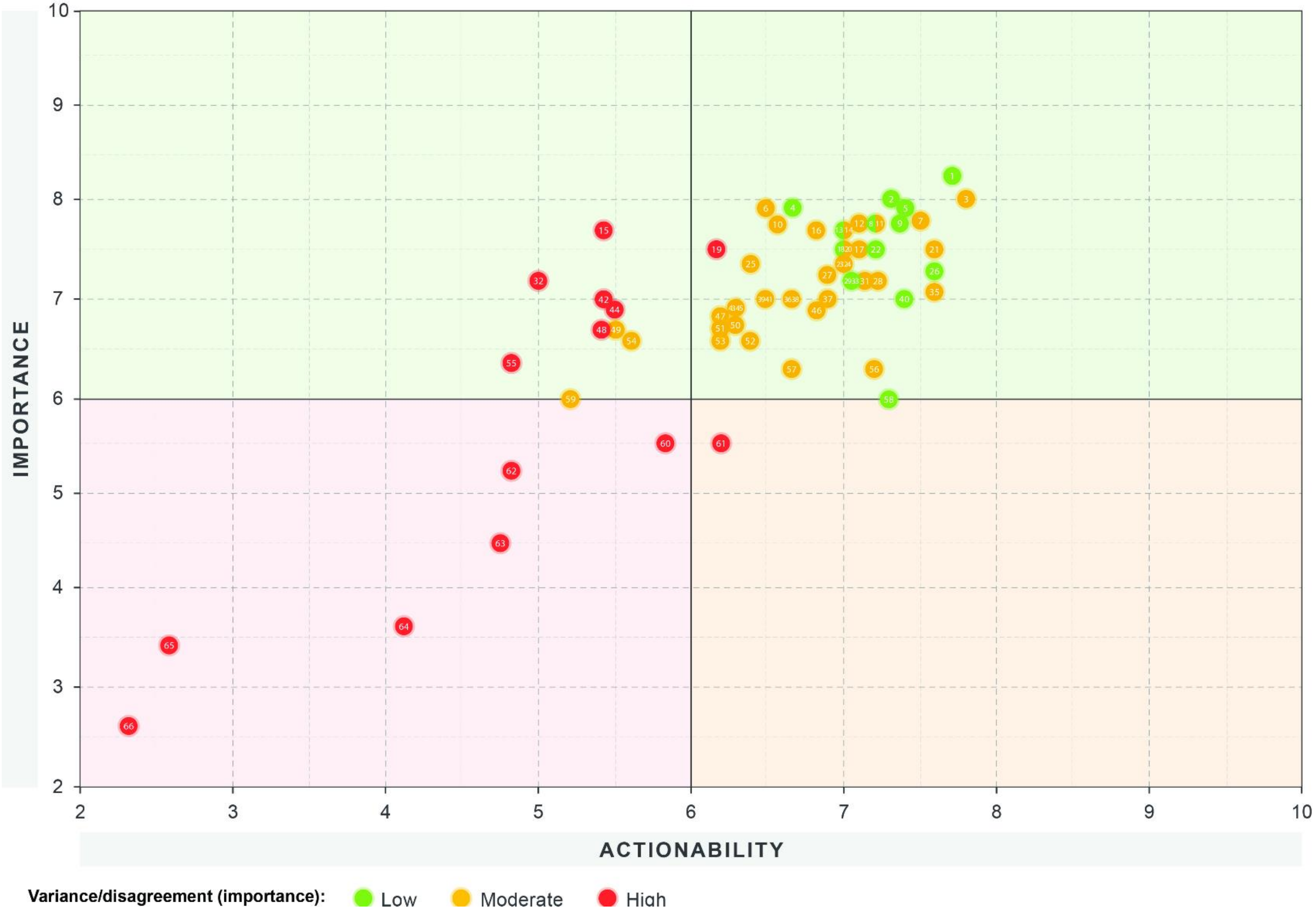
Figure B.2. Experts identified and ranked 38 potential future AI risks



1. Cyberattacks
2. Manipulation/disinformation/fraud
3. Harms to democracy
4. Racing dynamics
5. Lack of alignment methods
6. Concentrations of power
7. Information pollution
8. AI incidents and disasters
9. Extreme power concentration
10. Surveillance
11. Pacing problem
12. Lack of explain/interpret-ability
13. Reduced privacy
14. Exacerbate inequality
15. Kinetic attacks (e.g. LAWS)
16. Inadequate auditing/assurance
17. Biometrics misuse
18. Harmful scientific discovoery
19. Unclear legal accountability
20. Employment level shocks
21. Worsened conflict
22. Improper regulatory models
23. Lack of digital identity limits trust
24. Difficulty evaluating AI harms
25. AI governance blind spots
26. Over-reliance on AI systems
27. Disempowered public
28. AI changes human behaviour
29. Divergent AI development pathways
30. Misaligned AGI cannot be controlled
31. Multi-agent conflicts
32. Negative job quality impacts
33. Gaps in neuroscience impact AI oversight
34. AI environmental harms
35. Discord in the AI community
36. Loss of human heterogeneity
37. Delays in real-world AI value
38. Machines deserve rights but are denied

Note: See Annex A for details on the source and rating scale.

Figure B.3. Experts identified and ranked 66 potential future AI policy actions



Note: See Annex A for details on the source and rating scale. See table B.1 for a numbered list of the individual actions.

Table B.1. Experts identified and ranked 66 potential policy actions in Figure B.3.

1: Liability rules for AI-caused harm	**23**: Oversight of AI systems and compute	**45**: Research collaboration framework
2: R&D - AI safety	**24**: Better multi-disciplinary integration	**46**: Regional research hubs
3: AI systems disclosure when interacting with humans	**25**: Mitigate power concentrations - distribute benefits	**47**: Research on human preferences
4: R&D - AI alignment	**26**: Moonshots/mission-oriented approaches	**48**: Make human preferences the goal of AI systems
5: R&D - dangerous capabilities assessments, evals, assurance	**27**: Stronger regulatory enforcement	**49**: Standards and qualifications for fact-checking
6: AI red-lines (prohibit some use cases)	**28**: R&D - AI quality assurance	**50**: Individual ownership of private data
7: Require disclosure of key info (e.g., safety practices, model cards)	**29**: R&D - benchmarking AI for societal impact	**51**: Empower civil society
8: R&D – Interpret/explainability, transparency	**30**: International regulatory oversight (e.g., IAEA for AI)	**52**: Provenance/watermarking systems
9: Education & retraining	**31**: Secure/interoperable digital identity	**53**: Penalties for purveying improper information
10: Reinforce democratic processes	**32**: Right to mental security	**54**: Promote human/AI collaboration
11: Foster an AI-literate public	**33**: Privacy preserving technologies	**55**: Share superintelligent AI technology
12: Controlled release of AI models	**34**: Mitigate power concentrations - provide enablers/resources	**56**: Bias bounty programmes
13: R&D - robustness	**35**: Responsible corporate governance models	**57**: Strengthen labour unions/protections
14: Strengthen social safety nets	**36**: National AI regulatory authorities	**58**: R&D - model editing/finetuning
15: Ban LAWS	**37**: Media platforms showing content from reputable sources	**59**: Consider a future when employment isn't required
16: Controlled development of AI models	**38**: Traceability	**60**: Log AI interactions with humans
17: AI certification/auditing ecosystem	**39**: International panel (e.g., IPCC for AI)	**61**: International declaration on AGI risks
18: Mitigate power concentrations - regulation	**40**: R&D - fairness	**62**: Universal Basic Income (UBI)
19: Ban machines impersonating humans	**41**: Global ecosystem - experts working with regulators	**63**: Insurance policies in case of automation
20: Public engagement	**42**: International treaties	**64**: Tax automation
21: Dynamic regulatory processes, experimentation	**43**: International AI research body (e.g., CERN for AI)	**65**: Advanced AI R&D moratorium
22: Good practice principles/standards/norms	**44**: R&D - Truthful AI	**66**: Ban advanced AI

References

Access Now (2021), *Ban Biometric Surveillance*, https://www.accessnow.org/campaign/ban-biometric-surveillance. [195]

Acemoglu, D. (2024), *The Simple Macroeconomics of AI*, Economic Policy, https://www.economic-policy.org/wp-content/uploads/2024/04/EcPol-2024-016_Proof_hi_Acemoglu.pdf. [20]

Acemoglu, D. and P. Restrepo (2020), "The wrong kind of AI? Artificial intelligence and the future of labour demand", *Cambridge Journal of Regions, Economy and Society*, Vol. 13/1, pp. 25-35, https://doi.org/10.1093/cjres/rsz022. [216]

AI Now Institute (2023), *2023 Landscape: Confronting Tech Power*, https://ainowinstitute.org/2023-landscape. [174]

AI.gov (2023), *The Government is Using AI to Better Serve the Public*, https://ai.gov/ai-use-cases. [98]

AI.gov (2022), *Request for Information (RFI) on Public and Private Sector Uses of Biometric Technologies: Responses*, https://www.ai.gov/rfi/2022/86-FR-56300/Barrett-Biometric-RFI-2022.pdf. [196]

Altman, S. (2023), *Planning for AGI and Beyond*, https://openai.com/index/planning-for-agi-and-beyond/ (accessed on 23 May 2024). [108]

Anderljung, M. and A. Korinek (2024), *Frontier AI Regulation: Safeguards Amid Rapid Progress*, https://www.lawfaremedia.org/article/frontier-ai-regulation-safeguards-amid-rapid-progress. [276]

Anderson, B. and E. Sutherland (2024), "Collective action for responsible AI in health", *OECD Artificial Intelligence Papers*, No. 10, OECD Publishing, Paris, https://doi.org/10.1787/f2050177-en. [26]

Anthropic (2023), *Anthropic's Responsible Scaling Policy*, https://www.anthropic.com/news/anthropics-responsible-scaling-policy. [271]

Askell, A. et al. (2019), *The Role of Cooperation in Responsible AI Development*, https://arxiv.org/pdf/1907.04534. [149]

Australian Government Department of Education (2023), *Australian Framework for Generative Artificial Intelligence (AI) in Schools*, https://www.education.gov.au/schooling/resources/australian-framework-generative-artificial-intelligence-ai-schools. [301]

Autor, D. (2024), *Applying AI to Rebuild Middle Class Jobs*, National Bureau of Economic Research, Cambridge, MA, https://doi.org/10.3386/w32140. [23]

Azizzadenesheli, K. et al. (2024), *Neural Operators for Accelerating Scientific Simulations and Design*, https://arxiv.org/pdf/2309.15325. [8]

Bales, A., W. D'Alessandro and C. Kirk-Giannini (2024), *Artificial Intelligence: Arguments for Catastrophic Risk*, https://arxiv.org/pdf/2401.15487. [157]

Barnum, M. (2023), *Mark Zuckerberg tried to revolutionize American education with technology. It didn't go as planned.*, https://www.chalkbeat.org/2023/10/4/23903768/mark-zuckerberg-czi-schools-personalized-learning-technology-summit/. [58]

Bekenova, Z. et al. (2022), "Artificial Intelligence, Value Alignment and Rationality", *TalTech Journal of European Studies*, Vol. 12/1, pp. 79-98, https://doi.org/10.2478/bjes-2022-0004. [162]

Bell, S. and A. Korinek (2024), "AI's Economic Peril", *Journal of Democracy*, Vol. 34/4, https://www.journalofdemocracy.org/articles/ais-economic-peril/. [103]

Bender, E. et al. (2021), "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–23*, Vol. Association for Computing Machinery, Inc., https://doi.org/10.1145/3442188.3445922. [209]

Bengio, Y. (2024), *Government Interventions to Avert Future Catastrophic AI Risks*, https://hdsr.mitpress.mit.edu/pub/w974bwb0/release/2. [257]

Bennett Institute (2024), *Public Health gets personal: The case for an AI-driven personalised prevention platform*, https://www.bennettinstitute.cam.ac.uk/wp-content/uploads/2024/03/The-case-for-an-AI-driven-personalised-prevention-platform.pdf. [62]

Bentley, S. et al. (2024), "The digital divide in action: how experiences of digital technology shape future relationships with artificial intelligence", *AI and Ethics*, https://doi.org/10.1007/s43681-024-00452-3. [31]

Bettelle, J. (2023), *AI's "Oppenheimer Moment" Is Bullshit*, https://battellemedia.com/archives/2023/05/ais-oppenheimer-moment-is-bullshit. [173]

Bianchi, F., A. Cercas Curry and D. Hovy (2023), "Viewpoint: Artificial Intelligence Accidents Waiting to Happen?", *Journal of Artificial Intelligence Research*, Vol. 76, pp. 193-199, https://doi.org/10.1613/jair.1.14263. [183]

Bianchini, S., M. Müller and P. Pelletier (2022), "Artificial intelligence in science: An emerging general method of invention", *Research Policy*, Vol. 51/10, p. 104604, https://doi.org/10.1016/j.respol.2022.104604. [14]

Bollans, S. (2023), *EU Artificial Intelligence Liability Directive*, https://www.shlegal.com/insights/eu-artificial-intelligence-liability-directive. [251]

Bond, M. (2023), *Levelling the educational playing field with artificial intelligence*, https://www.open.ac.uk/blogs/digitalaccessadvisor/index.php/2023/08/22/levelling-the-educational-playing-field-with-artificial-intelligence/. [67]

Bontcheva, K. et al. (2024), *Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities*, European Digital Media Observatory, https://edmo.eu/wp-content/uploads/2023/12/Generative-AI-and-Disinformation_-White-Paper-v8.pdf. [125]

Božić, V. (2023), *Artifical Intelligence as the Reason and the Solution of Digital Divide*, https://www.langedutech.com/letjournal/index.php/let/article/view/53. [29]

Bremmer, I. and M. Suleyman (2023), *The AI Power Paradox: Can States Learn to Govern Artificial Intelligence—Before It's Too Late?*, https://www.foreignaffairs.com/world/artificial-intelligence-power-paradox. [147]

Brennan, J. (2023), *AI assurance? Assessing and mitigating risks across the AI lifecycle*, https://www.adalovelaceinstitute.org/report/risks-ai-systems. [289]

Brumfiel, G. (2023), *How AI is revolutionizing how governments conduct surveillance*, https://www.npr.org/2023/06/13/1181868277/how-ai-is-revolutionizing-how-governments-conduct-surveillance. [193]

Brundage, M. et al. (2018), *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf. [117]

Brynjolfsson, E., L. Danielle and L. Raymond (2023), *Generative AI at Work*, National Bureau of Economic Research, https://doi.org/10.3386/w31161. [48]

Brynjolfsson, E., D. Rock and C. Syverson (2020), *The Productivity J-Curve: How intangibles complement general purpose technologies*, https://www.aeaweb.org/articles?id=10.1257/mac.20180386. [25]

Buchanan, B. (2020), *The AI Triad and What It Means for National Security Strategy*, Center for Security and Emerging Technology, https://doi.org/10.51593/20200021. [165]

Calvino, F. and L. Fontanelli (2023), "A portrait of AI adopters across countries: Firm characteristics, assets' complementarities and productivity"*, OECD Science, Technology and Industry Working Papers*, No. 2023/02, OECD Publishing, Paris, https://doi.org/10.1787/0fb79bb9-en. [17]

Carter, T. (2023), *It is 'nearly unavoidable' that AI will cause a financial crash within a decade, SEC head says*, https://www.businessinsider.in/stock-market/news/it-is-nearly-unavoidable-that-ai-will-cause-a-financial-crash-within-a-decade-sec-head-says/articleshow/104473885.cms. [180]

Casal, J. and M. Kessler (2023), "Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing", *Research Methods in Applied Linguistics*, Vol. 2/3, p. 100068, https://doi.org/10.1016/j.rmal.2023.100068. [127]

Casper, S. et al. (2023), *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*, https://arxiv.org/abs/2307.15217. [160]

Cass-Beggs, D. (2024), *A Welcome Voice for Canada on the Future of AI*, https://www.techpolicy.press/a-welcome-voice-for-canada-on-the-future-of-ai/. [293]

Central Digital and Data Office (2024), *Artificial Intelligence: introducing our series of online courses on generative AI*, https://cddo.blog.gov.uk/2024/01/19/artificial-intelligence-introducing-our-series-of-online-courses-on-generative-ai/. [300]

Chow, A. and B. Perrigo (2023), *The AI Arms Race Is Changing Everything*, https://time.com/6255952/ai-impact-chatgpt-microsoft-google. [143]

Christian, B. (2020), “The Alignment Problem”, in *The Alignment Problem: Machine Learning and Human Values*, W.W. Norton & Company. [205]

Clarke, S. and J. Whittlestone (2022), “A Survey of the Potential Long-term Impacts of AI”, *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, https://doi.org/10.1145/3514094.3534131. [47]

Cockburn, M., R. Henderson and S. Stern (2018), *The Impact of Artificial Intelligence on Innovation*, National Bureau of Economic Research, https://www.nber.org/papers/w24449. [172]

CoE (2024), *The Framework Convention on Artificial Intelligence*, https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence. [233]

Computational Democracy (2023), *Featured Case Studies*, https://compdemocracy.org/Case-studies/. [104]

Cornelli, G. and J. Frost (2023), *Artificial intelligence, services globalisation and income inequality*, https://www.bis.org/publ/work1135.htm. [208]

CSET (2021), *AI Accidents: An Emerging Threat*, Center for Security and Emerging Technology, https://doi.org/10.51593/20200072. [182]

de Neufville, R. and S. Baum (2021), “Collective action on artificial intelligence: A primer and review”, *Technology in Society*, Vol. 66, p. 101649, https://doi.org/10.1016/j.techsoc.2021.101649. [146]

Deepmind, G. (2024), *Our Mission*, https://deepmind.google/about/ (accessed on 21 June 2024). [109]

Dell’Acqua, F. et al. (2023), “Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality”, *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.4573321. [19]

Demaidi, M. (2023), “Artificial intelligence national strategy in a developing country”, *AI & Society*, https://doi.org/10.1007/s00146-023-01779-x. [66]

Deshpande, A. et al. (2023), *Anthropomorphization of AI: Opportunities and Risks*, https://aclanthology.org/2023.nllp-1.1.pdf. [135]

Dieterle, E., C. Dede and M. Walker (2024), “The cyclical ethical effects of using artificial intelligence in education”, *AI &amp; SOCIETY*, Vol. 39/2, pp. 633-643, https://doi.org/10.1007/s00146-022-01497-w. [214]

DiResta, A. and Z. Sherman (2023), *The FTC Is Regulating AI: A Comprehensive Analysis*, https://www.hklaw.com/en/insights/publications/2023/07/the-ftc-is-regulating-ai-a-comprehensive-analysis. [253]

Dizikes, P. (2023), *How an “AI-tocracy” emerges*, https://news.mit.edu/2023/how-ai-tocracy-emerges-0713. [176]

Draghi, M. (2024), *The future of European competitiveness*, European Commission, https://commission.europa.eu/topics/strengthening-european-competitiveness/eu-competitiveness-looking-ahead_en. [203]

Du, M. (2023), “Machine vs. human, who makes a better judgment on innovation? Take GPT-4 [49] for example”, *Frontiers in Artificial Intelligence*, Vol. 6, https://doi.org/10.3389/frai.2023.1206516.

Dung, L. (2023), “Current cases of AI misalignment and their implications for future risks”, *Synthese*, Vol. 202/138, https://doi.org/10.1007/s11229-023-04367-0. [161]

Earthna (2023), *AI Solutions to Combat Climate Change*, https://www.earthna.qa/sites/default/files/inline-files/AI%20Solutions%20to%20combat%20climate%20change-English.pdf. [41]

EC (2024), *European Data Act enters into force, putting in place new rules for a fair and innovative data economy*, https://digital-strategy.ec.europa.eu/en/news/european-data-act-enters-force-putting-place-new-rules-fair-and-innovative-data-economy. [310]

EC (2022), *Impact assessment - Proposal for a Directive on adapting non contractual civil liability rules to artificial intelligence*, European Commission, https://commission.europa.eu/document/download/a25ea208-9a1d-483b-ab71-bcd1905e9000_en?filename=1_4_197608_impact_asse_dir_ai_en.pdf. [246]

EDRi (2021), *Civil society calls for AI red lines in the European Union’s Artificial Intelligence proposal*, https://edri.org/our-work/civil-society-call-for-ai-red-lines-in-the-european-unions-artificial-intelligence-proposal. [254]

Efthymiou, I., A. Alevizos and S. Sidiropoulos (2023), *The Role of Artificial Intelligence in Revolutionizing NGOs’ Work*, https://www.researchgate.net/publication/372824365_The_Role_of_Artificial_Intelligence_in_Revolutionizing_NGOs'_Work. [86]

Eisen, N. et al. (2023), *AI can strengthen U.S. democracy—and weaken it*, https://www.brookings.edu/articles/ai-can-strengthen-u-s-democracy-and-weaken-it/. [307]

Elements of AI (2024), *Elements of AI*, https://www.elementsofai.com. [239]

EP (2020), *The Impact of the General Data Protection (GDPR) on Artificial Intelligence*, https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf. [232]

EU (2024), *EU support tools for SMEs*, https://europa.eu/youreurope/business/running-business/eu-support-tools-sme/index_en.htm. [222]

EU (2002), *Directive 2002/14/EC - informing and consulting employees*, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02002L0014-20151009. [101]

European Parliament (2022), *Artificial intelligence liability directive*, https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739342/EPRS_BRI(2023)739342_EN.pdf. [250]

European Parliament (2021), *Improving Working Conditions Using Artificial Intelligence*, https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662911/IPOL_STU(2021)662911_EN.pdf. [102]

European Union (2024), *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*, https://eur-lex.europa.eu/eli/reg/2024/1689/oj. [88]

Evans, O. et al. (2021), *Truthful AI: Developing and governing AI that does not lie*, https://arxiv.org/abs/2110.06674. [296]

Fadel, C. et al. (2024), *Education for the Age of AI*, https://curriculumredesign.org/our-work/education-for-the-age-of-ai. [59]

Faggella, D. (2024), *Introducing 'The Trajectory' – A Specific Editorial Focus on Power and Artificial General Intelligence*, https://emerj.com/emerj-team-updates/introducing-the-trajectory-editorial-focus-on-power-and-artificial-general-intelligence. [294]

Fan, Q. and C. Qiang (2023), *Tipping the scales: AI's dual impact on developing nations*, https://blogs.worldbank.org/en/digital-development/tipping-the-scales--ai-s-dual-impact-on-developing-nations. [64]

Fariani, R., K. Junus and H. Santoso (2023), "A Systematic Literature Review on Personalised Learning in the Higher Education Context", *Technology, Knowledge and Learning*, Vol. 28/2, pp. 449–476, https://doi.org/10.1007/s10758-022-09628-4. [69]

Fassihi, F. (2023), *U.N. Officials Urge Regulation of Artificial Intelligence*, https://www.nytimes.com/2023/07/18/world/un-security-council-ai.html. [116]

Fazlioglu, M. (2024), *Consumer Perspectives of Privacy and Artificial Intelligence*, https://iapp.org/resources/article/consumer-perspectives-of-privacy-and-ai/. [192]

Feldstein, S. (2022), *AI & Big Data Global Surveillance Index (2022 updated)*, https://data.mendeley.com/datasets/gjhf5y4xjp/4. [184]

Flavián, C. and L. Casaló (2021), "Artificial intelligence in services: current trends, benefits and challenges", *The Service Industries Journal*, Vol. 41/13-14, pp. 853–859, https://doi.org/10.1080/02642069.2021.1989177. [71]

Fletcher, R., C. Tzani and M. Ioannou (2024), "The dark side of Artificial Intelligence – Risks arising in dating applications", *Assessment and Development Matters*, Vol. 16/1, pp. 17-23, https://doi.org/10.53841/bpsadm.2024.16.1.17. [134]

Forum on Information & Democracy (2024), *AI as a Public Good: Ensuring Democratic Control of AI in the Information Space*, https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf. [244]

Franca, C. (2023), *AI empowering research: 10 ways how science can benefit from AI*, https://arxiv.org/abs/2307.10265. [12]

France Commission on AI (2024), *Our AI: Our Ambition for France*, https://www.info.gouv.fr/upload/media/content/0001/09/02cbcb40c3541390be391feb3d963a4126b12598.pdf. [224]

FTC (2024), *FTC Launches Inquiry into Generative AI Investments and Partnerships*, https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-launches-inquiry-generative-ai-investments-partnerships. [314]

FTC (2024), *FTC Proposes New Protections to Combat AI Impersonation of Individuals*, https://www.ftc.gov/news-events/news/press-releases/2024/02/ftc-proposes-new-protections-combat-ai-impersonation-individuals. [252]

FTC (2023), *An Inquiry into Cloud Computing Business Practices: The Federal Trade Commission is seeking public comments*, https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/03/inquiry-cloud-computing-business-practices-federal-trade-commission-seeking-public-comments. [225]

FTC (2023), *Rite Aid Banned from Using AI Facial Recognition After FTC Says Retailer Deployed Technology without Reasonable Safeguards*, https://www.ftc.gov/news-events/news/press-releases/2023/12/rite-aid-banned-using-ai-facial-recognition-after-ftc-says-retailer-deployed-technology-without. [261]

Funk, A., A. Shahbaz and K. Vesteinsson (2023), *The Repressive Power of Artificial Intelligence*, https://www.newamerica.org/planetary-politics/briefs/power-governance-ai-public-good/. [175]

G7 (2023), *Hiroshima Process International Code of Conduct for Advanced AI Systems*, https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems. [219]

G7 (2023), *Hiroshima Process International Guiding Principles for Advanced AI system*, https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-guiding-principles-advanced-ai-system. [284]

Gabriel, I. (2020), “Artificial Intelligence, Values, and Alignment”, *Minds and Machines*, Vol. 30/3, pp. 411-437, https://doi.org/10.1007/s11023-020-09539-2. [156]

Gao, L. and L. Guan (2023), *Interpretability of Machine Learning: Recent Advances and Future Prospects*, https://arxiv.org/pdf/2305.00537. [204]

Garfinkel, B. (2019), “How Does the Offense-Defense Balance Scale?”, *Journal of Strategic Studies*, Vol. 42/6, pp. 736-763, https://doi.org/10.1080/01402390.2019.1631810. [154]

Gates, B. (2023), *The Age of AI has begun*, https://www.gatesnotes.com/The-Age-of-AI-Has-Begun. [39]

Georgieff, A. (2024), *Artificial intelligence and wage inequality*, OECD Publishing, https://doi.org/10.1787/bf98a45c-en. [28]

Gerstein, D. and E. Leidy (2024), *Emerging Technology and Risk Analysis: Artificial intelligence and critical infrastructure*, https://www.rand.org/content/dam/rand/pubs/research_reports/RRA2800/RRA2873-1/RAND_RRA2873-1.pdf. [118]

Goldman Sachs (2023), *Generative AI could raise global GDP by 7%*, https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html. [21]

Goralski, M. and T. Tan (2022), “Artificial intelligence and poverty alleviation: Emerging innovations and their implications for management education and sustainable development”, *The International Journal of Management Education*, Vol. 20/3, p. 100662, https://doi.org/10.1016/j.ijme.2022.100662. [34]

Gottschalk, F. and C. Weise (2023), “Digital equity and inclusion in education: An overview of practice and policy in OECD countries”*, OECD Education Working Papers*, No. 299, OECD Publishing, Paris, https://doi.org/10.1787/7cb15030-en. [30]

GOV.UK (2024), *Algorithmic Transparency Recording Standard*, https://www.gov.uk/government/publications/algorithmic-transparency-template. [105]

Government of Canada (2023), *Artificial Intelligence and Data Act*, https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act. [260]

Government of Croatia (2024), *AI – From Concept To Implementation*, https://www.carnet.hr/wp-content/uploads/2024/04/Kurikulum-izvannastavne-aktivnosti-za-osnovne-skole_Umjetna-inteligencija.pdf. [302]

Government of Denmark (2024), *Digital Democracy Initiative*, https://digitaldemocracyinitiative.net/-/media/6193cf77b746491db6a041cfb43d1d91.ashx. [78]

Grallet, G. and H. Pons (2023), *Yuval Noah Harari (Sapiens) versus Yann Le Cun (Meta) on artificial intelligence*, https://www.lepoint.fr/sciences-nature/yuval-harari-sapiens-versus-yann-le-cun-meta-on-artificial-intelligence-11-05-2023-2519782_1924.php. [202]

Griliches, Z. (1957), “Hybrid Corn: An Exploration in the Economics of Technological Change”, *Econometrica*, Vol. 25/4, p. 501, https://doi.org/10.2307/1905380. [13]

Groves, L. (2024), *Code & conduct: How to create third-party auditing regimes for AI systems*, https://www.adalovelaceinstitute.org/report/code-conduct-ai/. [277]

Gruetzemacher, R., D. Paradice and K. Lee (2019), *Forecasting Transformative AI: An Expert Survey*, https://arxiv.org/abs/1901.08579. [16]

Haluza, D. and D. Jungwirth (2023), “Artificial Intelligence and Ten Societal Megatrends: An Exploratory Study Using GPT-3”, *Systems*, Vol. 11/3, p. 120, https://doi.org/10.3390/systems11030120. [44]

Hambling, D. (2019), *The Pentagon has a laser that can identify people from a distance—by their heartbeat*, https://www.technologyreview.com/2019/06/27/238884/the-pentagon-has-a-laser-that-can-identify-people-from-a-distanceby-their-heartbeat. [190]

Haramboure, A. et al. (2023), “Vulnerabilities in the semiconductor supply chain”*, OECD Science, Technology and Industry Working Papers*, No. 2023/05, OECD Publishing, Paris, https://doi.org/10.1787/6bed616f-en. [167]

Hart, M. (2023), *AI in action: Reshaping ergonomic safety strategies in the automotive industry*, https://www.shponline.co.uk/ergonomics/ai-in-action-reshaping-ergonomic-safety-strategies-in-the-automotive-industry/. [72]

Hendrycks, D., M. Mazeika and T. Woodside (2023), *An Overview of Catastrophic AI Risks*, https://arxiv.org/abs/2306.12001. [148]

Hendrycks, D. et al. (2022), *Unsolved Problems in ML Safety*, https://arxiv.org/abs/2109.13916. [158]

Honorof, M. (2023), *The future of AI could hinge on two philosophical concepts*, https://www.tomsguide.com/features/ai-philosophy-solipsism-blockhead. [5]

Horvitz, E. (2022), *Artificial Intelligence and Cybersecurity: Rising challenges and promising directions*, https://www.armed-services.senate.gov/imo/media/doc/5.3.22%20Eric%20Horvitz%20Testimony.pdf. [132]

Horvitz, E. (2022), *On the Horizon: Interactive and Compositional Deepfakes*, [131] https://arxiv.org/abs/2209.01714.

Horvitz, E. (2014), *Reflections and Framing: One-Hundred Year Study on Artificial Intelligence: Reflections and Framing*, https://ai100.stanford.edu/reflections-and-framing. [52]

Huang, M. and R. Rust (2021), “Engaged to a Robot? The Role of AI in Service”, *Journal of Service Research*, Vol. 24/1, pp. 30–41, https://doi.org/10.1177/1094670520902266. [70]

IEEE (2024), *IEEE Approved Draft Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems*, https://standards.ieee.org/ieee/7009/7096/. [231]

IMF (2024), *Gen-AI: Artificial Intelligence and the Future of Work*, https://www.imf.org/-/media/Files/Publications/SDN/2024/English/SDNEA2024001.ashx. [217]

Israel Ministry of Innovation, Science and Technology (2023), *A call for applications to receive support for government projects based on artificial intelligence*, https://www.gov.il/he/pages/rfp16082023. [94]

Janjeva, A. et al. (2023), *The Rapid Rise of Generative AI: Assessing risks to safety and security*, https://cetas.turing.ac.uk/publications/rapid-rise-generative-ai. [255]

Jarrahi, M. et al. (2023), “Artificial intelligence and knowledge management: A partnership between human and AI”, *Business Horizons*, Vol. 66/1, pp. 87-99, https://doi.org/10.1016/j.bushor.2022.03.002. [56]

Javaid, M. et al. (2023), “Understanding the potential applications of Artificial Intelligence in Agriculture Sector”, *Advanced Agrochem*, Vol. 2/1, pp. 15-30, https://doi.org/10.1016/j.aac.2022.10.001. [35]

Jia, N. et al. (2024), “When and How Artificial Intelligence Augments Employee Creativity”, *Academy of Management Journal*, Vol. 67/1, pp. 5-32, https://doi.org/10.5465/amj.2022.0426. [74]

Ji, J. (2023), *What Does AI Red-Teaming Actually Mean?*, https://cset.georgetown.edu/article/what-does-ai-red-teaming-actually-mean/. [291]

Ji, J. et al. (2024), *AI Alignment: A Comprehensive Survey*, https://arxiv.org/abs/2310.19852. [287]

Johnson, J. (2020), *Artificial Intelligence: A Threat to Strategic Stability*, https://www.airuniversity.af.edu/Portals/10/SSQ/documents/Volume-14_Issue-1/Johnson.pdf. [153]

Jones, C. (2022), “The Past and Future of Economic Growth: A Semi-Endogenous Perspective”, *Annual Review of Economics*, Vol. 14/1, pp. 125-152, https://doi.org/10.1146/annurev-economics-080521-012458. [18]

Kahn, J. (2024), *Exclusive: OpenAI promised 20% of its computing power to combat the most dangerous kind of AI—but never delivered, sources say*, https://fortune.com/2024/05/21/openai-superalignment-20-compute-commitment-never-fulfilled-sutskever-leike-altman-brockman-murati/. [152]

Karger, E. et al. (2023), *Forecasting Existential Risks: Evidence from a Long-Run Forecasting Tournament*, https://static1.squarespace.com/static/635693acf15a3e2a14a56a4a/t/64f0a7838ccbf43b6b5ee40c/1693493128111/XPT.pdf. [317]

Kechhar, R. (2023), *Workers’ views on the risk of AI to their jobs*, [73] https://www.pewresearch.org/social-trends/2023/07/26/workers-views-on-the-risk-of-ai-to-their-jobs/.

Khan, L. (2023), *We Must Regulate A.I. Here's How*, https://www.nytimes.com/2023/05/03/opinion/ai-lina-khan-ftc-technology.html. [133]

Klimek, P. (2023), *Why generative AI is a double-edged sword for the cybersecurity sector*, https://venturebeat.com/security/why-generative-ai-is-a-double-edged-sword-for-the-cybersecurity-sector/. [111]

Koivisto, M. and S. Grassini (2023), "Best humans still outperform artificial intelligence in a creative divergent thinking task", *Scientific Reports*, Vol. 13/1, https://doi.org/10.1038/s41598-023-40858-3. [51]

Kosinski, M. (2021), "Author Correction: Facial recognition technology can expose political orientation from naturalistic facial images", *Scientific Reports*, Vol. 11/1, https://doi.org/10.1038/s41598-021-02785-z. [186]

Kumar, R. and F. Nagle (2020), *The Case for AI Insurance*, https://hbr.org/2020/04/the-case-for-ai-insurance. [247]

Laplante, P. et al. (2020), "Artificial Intelligence and Critical Systems: From Hype to Reality", *Computer*, Vol. 53/11, pp. 45-52, https://doi.org/10.1109/mc.2020.3006177. [177]

Larsson, S., J. White and C. Bogusz (2024), *The Artificial Recruiter: Risks of Discrimination in Employers' Use of AI and Automated Decision-Making*, https://doi.org/doi.org/10.17645/si.7471. [211]

Latif, S. et al. (2022), *AI-Based Emotion Recognition: Promise, Peril, and Prescriptions for Prosocial Path*, https://arxiv.org/abs/2211.07290. [197]

Legraien, L. (2024), *Six in 10 charities use AI in day-to-day operations, report finds*, https://www.civilsociety.co.uk/news/six-in-10-charities-use-ai-in-day-to-day-operations-report-finds.html. [77]

Leike, J. (2022), *Three alignment taxes*, https://aligned.substack.com/p/three-alignment-taxes. [151]

Li, B. (2023), *Large Language Models, Innovation, and Capitalism*, https://belindal.github.io/blog/llm-capitalism/. [144]

Lima, G. et al. (2022), *The Conflict Between Explainable and Accountable Decision-Making Algorithms*, https://arxiv.org/abs/2205.05306. [207]

Lomas, N. (2024), *EU wants to upgrade its supercomputers to support generative AI startups*, https://techcrunch.com/2024/01/24/eu-supercomputers-for-ai-2/. [223]

Lorenz, P., K. Perset and J. Berryhill (2023), "Initial policy considerations for generative artificial intelligence", *OECD Artificial Intelligence Papers*, No. 1, OECD Publishing, Paris, https://doi.org/10.1787/fae2d1e6-en. [137]

Makhija, P., E. Chacko and M. Kukreja (2024), "A Global Taxonomy of Flash Crashes: Cases Demonstrating the Operation and Impact of High-Frequency Traders", in *Transformations in Banking, Finance and Regulation, Banking Resilience*, WORLD SCIENTIFIC (EUROPE), https://doi.org/10.1142/9781800614291_0014. [179]

Malgieri, G. and F. Pasquale (2024), “Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology”, *Computer Law &amp; Security Review*, Vol. 52, p. 105899, https://doi.org/10.1016/j.clsr.2023.105899. [273]

Manning, B., K. Zhu and J. Horton (2024), *Automated Social Science: Language Models as Scientist and Subjects*, https://arxiv.org/abs/2404.11794. [11]

Maraju, K., Rashu and T. Sagi (2024), “Hackers Weaponry: Leveraging AI Chatbots for Cyber Attacks”, in *Proceedings of the International Conference on Cybersecurity, Situational Awareness and Social Media, Springer Proceedings in Complexity*, Springer Nature Singapore, Singapore, https://doi.org/10.1007/978-981-99-6974-6_21. [112]

Mascellino, A. (2023), *Artificial Intelligence and USBs Drive 8% Rise in Cyber-Attacks*, https://www.infosecurity-magazine.com/news/ai-usbs-drive-rise-cyber-attacks. [113]

Matz, S. et al. (2024), “The potential of generative AI for personalized persuasion at scale”, *Scientific Reports*, Vol. 14/1, https://doi.org/10.1038/s41598-024-53755-0. [130]

Metaculus (2024), *When will the first general AI system be devised, tested, and publicly announced?*, https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/ (accessed on 2024). [319]

Metz, C. (2023), *What Exactly Are the Dangers Posed by A.I.?*, https://www.nytimes.com/2023/05/01/technology/ai-problems-danger-chatgpt.html. [199]

Mhlanga, D. (2021), “Artificial Intelligence in the Industry 4.0, and Its Impact on Poverty, Innovation, Infrastructure Development, and the Sustainable Development Goals: Lessons from Emerging Economies?”, *Sustainability*, Vol. 13/11, https://doi.org/10.3390/su13115788. [36]

MinTIC (2019), *Más de 25.000 colombianos podrán formarse gratis en Inteligencia Artificial y habilidades para la transformación digital gracias a MinTIC*, https://mintic.gov.co/portal/inicio/Sala-de-Prensa/Noticias/106989:Mas-de-25-000-colombianos-podran-formarse-gratis-enInteligencia-Artificial-y-habilidades-para-la-transformacion-digital-gracias-a-MinTIC. [238]

Modhvadia, R. (2023), *How do people feel about AI?*, https://www.adalovelaceinstitute.org/report/public-attitudes-ai/. [215]

Mozilla (2023), *Joint Statement on AI Safety and Openness*, https://open.mozilla.org/letter/. [170]

Muggah, R. (2023), *Artificial Intelligence Will Entrench Global Inequality*, https://foreignpolicy.com/2023/05/29/ai-regulation-global-south-artificial-intelligence. [218]

Mylly, U. (2023), “Transparent AI? Navigating Between Rules on Trade Secrets and Access to Information”, *IIC - International Review of Intellectual Property and Competition Law*, Vol. 54/7, pp. 1013-1043, https://doi.org/10.1007/s40319-023-01328-5. [266]

NAIRR (2024), *The National Artificial Intelligence Research Resource (NAIRR) Pilot*, https://nairrpilot.org. [97]

Nannini, L., A. Balayn and A. Smith (2023), “Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK”, *2023 ACM Conference on Fairness, Accountability, and Transparency*, https://doi.org/10.1145/3593013.3594074. [241]

Narayanan, S. and M. Potkewitz (2023), *A risk-based approach to assessing liability risk for AI-driven harms considering EU liability directive*, https://arxiv.org/abs/2401.11697. [243]

Navo, S. et al. (2023), *Securing Artificial Intelligence Model Weights: Interim Report*, RAND Corporation, https://doi.org/10.7249/wra2849-1. [275]

Nightingale, S. and H. Farid (2022), “Synthetic Faces Are More Trustworthy Than Real Faces”, *Journal of Vision*, Vol. 22/14, p. 3068, https://doi.org/10.1167/jov.22.14.3068. [128]

NIST (2024), *Computer Security Resource Center: defense-in-depth*, https://csrc.nist.gov/glossary/term/defense_in_depth. [242]

NIST (2024), *NIST identifies types of cyberattacks that manipulate behavior of AI systems*, https://www.nist.gov/news-events/news/2024/01/nist-identifies-types-cyberattacks-manipulate-behavior-ai-systems. [121]

NIST (2023), *AI Risk Management Framework*, https://www.nist.gov/itl/ai-risk-management-framework. [267]

NIST (2022), *Towards a standard for identifying and managing bias in artificial intelligence*, National Institute of Standards and Technology (U.S.), Gaithersburg, MD, https://doi.org/10.6028/nist.sp.1270. [210]

NSCIA (2021), *Final Report - National Security Commission on Artificial Intelligence*, https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf. [145]

O’Brien, J. (2023), *Deployment corrections: An incident response framework for frontier AI models*, https://www.iaps.ai/research/deployment-corrections. [282]

OECD (2024), *AI in Health: Huge potential, huge risks*, https://www.oecd.org/health/AI-in-health-huge-potential-huge-risks.pdf. [63]

OECD (2024), “AI, data governance and privacy: Synergies and areas of international co-operation”*, OECD Artificial Intelligence Papers*, No. 22, OECD Publishing, Paris, https://doi.org/10.1787/2476b1a4-en. [191]

OECD (2024), “Artificial intelligence, data and competition”*, OECD Artificial Intelligence Papers*, No. 18, OECD Publishing, Paris, https://doi.org/10.1787/e7e88884-en. [142]

OECD (2024), *Artificial intelligence, data and competition - Background Note*, https://one.oecd.org/document/DAF/COMP(2024)2/en/pdf. [320]

OECD (2024), “Defining AI incidents and related terms”*, OECD Artificial Intelligence Papers*, No. 16, OECD Publishing, Paris, https://doi.org/10.1787/d1a8d965-en. [230]

OECD (2024), *Facts not Fakes: Tackling Disinformation, Strengthening Information Integrity*, OECD Publishing, Paris, https://doi.org/10.1787/d909ff7a-en. [140]

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. [237]

OECD (2024), *Framework for the Anticipatory Governance of Emerging Technologies*, https://doi.org/10.1787/0248ead5-en. [303]

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. [75]

OECD (2024), *OECD Artificial Intelligence Review of Germany*, OECD Publishing, Paris, https://doi.org/10.1787/609808d6-en. [1]

OECD (2024), *OECD Digital Economy Outlook 2024 (Volume 1): Embracing the Technology Frontier*, OECD Publishing, https://doi.org/10.1787/a1689dc5-en. [107]

OECD (2024), *OECD Reinforcing Democracy Initiative*, https://www.oecd.org/governance/reinforcing-democracy. [315]

OECD (2024), *OECD Survey on Drivers of Trust in Public Institutions – 2024 Results: Building Trust in a Complex Policy Environment*, OECD Publishing, Paris, https://doi.org/10.1787/9a20554b-en. [305]

OECD (2024), *Recommendation of the Council on Artificial Intelligence*, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. [7]

OECD (2024), *Reinforcing democracy initiative*, https://www.oecd.org/en/about/programmes/reinforcing-democracy-initiative. [306]

OECD (2024), *The impact of Artificial Intelligence on productivity, distribution and growth: Key mechanisms, initial evidence and policy challenges*, OECD Publishing, https://doi.org/10.1787/8d900037-en. [32]

OECD (2024), “The OECD Truth Quest Survey: Methodology and findings”*, OECD Digital Economy Papers*, No. 369, OECD Publishing, Paris, https://doi.org/10.1787/92a94c0f-en. [126]

OECD (2023), *“Liability for damage caused by AI is one of the key barriers to AI adoption by EU businesses: Percentage of enterprises who responded that a specific barrier is applicable to their business”*, OECD Publishing, https://doi.org/10.1787/f60d3501-en. [245]

OECD (2023), “2023 OECD Open, Useful and Re-usable data (OURdata) Index: Results and key findings”*, OECD Public Governance Policy Papers*, No. 43, OECD Publishing, Paris, https://doi.org/10.1787/a37f51c3-en. [96]

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. [166]

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. [321]

OECD (2023), *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research*, OECD Publishing, Paris, https://doi.org/10.1787/a8d820bd-en. [10]

OECD (2023), “Common guideposts to promote interoperability in AI risk management”*, OECD Artificial Intelligence Papers*, No. 5, OECD Publishing, Paris, https://doi.org/10.1787/ba602d18-en. [285]

OECD (2023), *Engaging citizens in innovation policy: Why, when and how?*, OECD Publishing, https://doi.org/10.1787/ba068fa6-en. [83]

OECD (2023), *G7 Hiroshima Process on Generative Artificial Intelligence (AI): Towards a G7 Common Understanding on Generative AI*, OECD Publishing, Paris, https://doi.org/10.1787/bf3c0c60-en. [124]

OECD (2023), *Global Trends in Government Innovation 2023*, OECD Public Governance Reviews, OECD Publishing, Paris, https://doi.org/10.1787/0655b570-en. [82]

OECD (2023), *Is Education Losing the Race with Technology?: AI's Progress in Maths and Reading*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/73105f99-en. [50]

OECD (2023), *OECD Digital Education Outlook 2023: Towards an Effective Digital Education Ecosystem*, OECD Publishing, Paris, https://doi.org/10.1787/c74f03de-en. [68]

OECD (2023), *OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market*, OECD Publishing, Paris, https://doi.org/10.1787/08785bba-en. [24]

OECD (2023), *OECD Guidelines for Multinational Enterprises on Responsible Business Conduct*, OECD Publishing, Paris, https://doi.org/10.1787/81f92357-en. [283]

OECD (2023), *OECD Skills Outlook 2023: Skills for a Resilient Green and Digital Transition*, OECD Publishing, Paris, https://doi.org/10.1787/27452f29-en. [6]

OECD (2023), "Regulatory sandboxes in artificial intelligence"*, OECD Digital Economy Papers*, No. 356, OECD Publishing, Paris, https://doi.org/10.1787/8f80a0e6-en. [274]

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. [228]

OECD (2023), "Supporting decision making with strategic foresight: An emerging framework for proactive and prospective governments."*, OECD Working Papers on Public Governance*, No. 63, OECD Publishing, Paris, https://doi.org/10.1787/1d78c791-en. [2]

OECD (2022), *Building Trust and Reinforcing Democracy: Preparing the Ground for Government Action*, OECD Public Governance Reviews, OECD Publishing, Paris, https://doi.org/10.1787/76972a4a-en. [139]

OECD (2022), *Building Trust to Reinforce Democracy: Main Findings from the 2021 OECD Survey on Drivers of Trust in Public Institutions*, Building Trust in Public Institutions, OECD Publishing, Paris, https://doi.org/10.1787/b407f99c-en. [87]

OECD (2022), "Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint"*, OECD Digital Economy Papers*, No. 341, OECD Publishing, Paris, https://doi.org/10.1787/7babf571-en. [43]

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. [270]

OECD (2022), *OECD Guidelines for Citizen Participation Processes*, OECD Public Governance Reviews, OECD Publishing, Paris, https://doi.org/10.1787/f765caf6-en. [84]

OECD (2022), *The Protection and Promotion of Civic Space: Strengthening Alignment with International Standards and Guidance*, OECD Publishing, Paris, https://doi.org/10.1787/d234e975-en. [198]

OECD (2021), *Recommendation of the Council for Agile Regulatory Governance to Harness Innovation*, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0464. [236]

OECD (2020), *Anticipatory Innovation Governance: What it is, how it works, and why we need it more than ever before*, https://oecd-opsi.org/wp-content/uploads/2020/11/AnticipatoryInnovationGovernance-Note-Nov2020.pdf. [200]

OECD (2019), *How's Life in the Digital Age?: Opportunities and Risks of the Digital Transformation for People's Well-being*, OECD Publishing, Paris, https://doi.org/10.1787/9789264311800-en. [37]

OECD (2018), *Executive Summary of the hearing on Market Concentration*, OECD Publishing, https://one.oecd.org/document/DAF/COMP/M(2018)1/ANN7/FINAL/en/pdf. [226]

OECD (2017), *The Next Production Revolution: Implications for Governments and Business*, OECD Publishing, Paris, https://doi.org/10.1787/9789264271036-en. [4]

OECD (forthcoming), *Regulatory Policy Outlook 2025*, OECD Publishing. [201]

OECD (forthcoming), *Towards a common reporting framework for AI incidents and hazards*, OECD Publishing. [229]

OECD.AI (2024), *A new expert group at the OECD for policy synergies in AI, data, and privacy*, https://oecd.ai/en/wonk/expert-group-data-privacy. [234]

OECD.AI (2023), *AI legal cases are increasing: how can we prepare?*, https://oecd.ai/en/wonk/increasing-legal-cases. [249]

OECD.AI (2023), *Artificial General Intelligence: can we avoid the ultimate existential threat?*, https://oecd.ai/en/wonk/existential-threat. [164]

OECD.AI (2023), *Basic safety requirements for AI risk management*, https://oecd.ai/en/wonk/basic-safety-requirements-for-ai-risk-management. [288]

OECD.AI (2023), *OECD Expert Forum on Generative AI and AI Foresight*, See also video linked in source file., https://wp.oecd.ai/app/uploads/2023/06/Summary-Expert-Forum-on-Generative-AI-and-AI-Foresight.pdf. [80]

OECD.AI (2023), *OECD Expert Group on AI Futures – Meeting 1 (13 July 2023)*, https://wp.oecd.ai/app/uploads/2023/09/Expert-Group-on-AI-Futures-Meeting-1-Summary.pdf. [106]

OECD.AI (2023), *OECD Expert Group on AI Futures - Meeting 3*, See also video linked in source file., https://wp.oecd.ai/app/uploads/2024/01/Expert-Group-on-AI-Futures-Meeting-3-Summary.pdf. [15]

OECD.AI (2023), *We need to use AI to fight climate change*, https://oecd.ai/en/wonk/fight-climate-change. [42]

OECD.AI (2022), *Leveraging AI, big data analytics and people to fight untruths online*, https://oecd.ai/en/wonk/untruths-online. [123]

OECD.AI (2022), *New AI technologies can perpetuate old biases: some examples in the United States*, https://oecd.ai/en/wonk/ai-biases-usa. [206]

OECD.AI (2022), *Summary of OECD expert discussion on future risks from artificial intelligence*, [181] https://wp.oecd.ai/app/uploads/2023/03/OECD-Foresight-workshop-notes-1.pdf.

OECD.AI (2022), *The twin transitions: are digital technologies the key to a clean energy future?*, https://oecd.ai/en/wonk/twin-transitions. [95]

OECD/CAF (2022), *The Strategic and Responsible Use of Artificial Intelligence in the Public Sector of Latin America and the Caribbean*, OECD Public Governance Reviews, OECD Publishing, Paris, https://doi.org/10.1787/1f334543-en. [3]

Ognyanova, K. et al. (2020), “Misinformation in Action: Fake News Exposure Is Linked to Lower Trust in Media, Higher Trust in Government When Your Side Is in Power”, *Harvard Kennedy School Misinformation Review*, https://doi.org/10.37016/mr-2020-024. [138]

OpenAI (2024), *Disrupting malicious uses of AI by state-affiliated threat actors*, https://openai.com/blog/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors. [114]

OpenAI (2023), *Frontier risk and preparedness*, https://openai.com/blog/frontier-risk-and-preparedness. [272]

PAI (2023), *Guidance for Safe Foundation Model Deployment*, https://partnershiponai.org/modeldeployment/. [279]

PAI (2023), *Responsible Practices for Synthetic Media*, https://syntheticmedia.partnershiponai.org. [269]

Panditharatne, M., D. Weiner and D. Kriner (2023), *Artificial Intelligence, Participatory Democracy, and Responsive Government*, https://www.brennancenter.org/our-work/research-reports/artificial-intelligence-participatory-democracy-and-responsive-government. [309]

Perry, A. and N. Lee (2019), *AI is coming to schools, and if we’re not careful, so will its biases*, https://www.brookings.edu/articles/ai-is-coming-to-schools-and-if-were-not-careful-so-will-its-biases/. [213]

Privacy Ticker (2019), *Chinese police uses gait recognition for identification*, http://www.privacy-ticker.com/chinese-police-uses-gait-recognition-for-identification/. [189]

Pupillo, L. et al. (2021), *Artificial Intelligence and Cybersecurity*, CEPS, https://www.ceps.eu/ceps-publications/artificial-intelligence-and-cybersecurity-2/. [122]

Puwal, S. (2024), *Should artificial intelligence be banned from nuclear weapons systems?*, https://www.nato.int/docu/review/articles/2024/04/12/should-artificial-intelligence-be-banned-from-nuclear-weapons-systems/index.html. [119]

Rainie, L. and J. Anderson (2024), *Experts Imagine the Impact of Artificial Intelligence by 2040*, https://imaginingthedigitalfuture.org/wp-content/uploads/2024/02/AI2040-FINAL-White-Paper-2-2.29.24.pdf. [60]

Raji, I. et al. (2022), *Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance*, https://www.skillscommons.org/bitstream/handle/taaccct/18870/Raji_et_al_2022_Outsider_Oversight.pdf?sequence=3&isAllowed=y. [278]

Roberts, H. et al. (2024), “Global AI governance: barriers and pathways forward”, *International Affairs*, Vol. 100/3, pp. 1275-1286, https://doi.org/10.1093/ia/iiae073. [155]

Russell, S. (2024), *Framing the issues: Make AI safe or make safe AI?*, https://www.unesco.org/en/articles/framing-issues-make-ai-safe-or-make-safe-ai. [256]

Russell, S. (2022), *If We Succeed*, https://www.amacad.org/publication/if-we-succeed. [22]

Russell, S. (2019), *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking. [53]

Sahoo, P. et al. (2024), *Unveiling Hallucination in Text, Image, Video, and Audio Foundation Models: A Comprehensive Survey*, https://arxiv.org/abs/2405.09589. [297]

Sanchez, C. (2021), *Civil society can help ensure AI benefits us all. Here's how*, https://www.weforum.org/agenda/2021/07/civil-society-help-ai-benefits. [79]

Sathyaraj, P. et al. (2024), "Artificial Intelligence", in *Advances in Environmental Engineering and Green Technologies, Novel AI Applications for Advancing Earth Sciences*, IGI Global, https://doi.org/10.4018/979-8-3693-1850-8.ch001. [55]

Savaget, P., T. Chiarini and S. Evans (2019), "Empowering political participation through artificial intelligence", *Science and Public Policy*, Vol. 46/3, pp. 369-380, https://doi.org/10.1093/scipol/scy064. [76]

Schneier, B. (2023), *Ten Ways AI Will Change Democracy*, https://ash.harvard.edu/ten-ways-ai-will-change-democracy. [308]

Scott, G. (2024), *Labor Organizing and AI Surveillance in the Workplace*, https://www.law.georgetown.edu/poverty-journal/blog/labor-organizing-and-ai-surveillance-in-the-workplace/. [194]

Seger, E. et al. (2023), *Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives*, https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models. [171]

Shevlane, T. (2022), *Structured access: an emerging paradigm for safe AI deployment*, https://arxiv.org/abs/2201.05159. [280]

Simon, F., S. Altay and H. Mercier (2024), *Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown*, https://misinforeview.hks.harvard.edu/article/misinformation-reloaded-fears-about-the-impact-of-generative-ai-on-misinformation-are-overblown/. [141]

Simon, F., K. McBride and S. Altay (2024), *AI's impact on elections is being overblown*, https://www.technologyreview.com/2024/09/03/1103464/ai-impact-elections-overblown. [129]

Skalse, J. et al. (2022), *Defining and Characterizing Reward Hacking*, https://arxiv.org/abs/2209.13085. [159]

Staab, R. et al. (2023), *Beyond Memorization: Violating Privacy Via Inference with Large Language Models*, https://arxiv.org/abs/2310.07298. [187]

Stacey, K. and D. Milmo (2023), *AI developing too fast for regulators to keep up, says Oliver Dowden*, https://www.theguardian.com/technology/2023/sep/22/ai-developing-too-fast-for-regulators-to-keep-up-oliver-dowden. [235]

Stanford (2023), *Artificial Intelligence Index Report 2023*, Stanford, https://aiindex.stanford.edu/report/. [9]

Stark, L. and J. Hutson (2021), “Physiognomic Artificial Intelligence”, *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.3927300. [188]

Stein-Perlman, Z., B. Weinstein-Raun and K. Grace (2022), *2022 Expert Survey on Progress in AI*, https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/. [316]

Sumner, J. et al. (2023), “Developing an Artificial Intelligence-Driven Nudge Intervention to Improve Medication Adherence: A Human-Centred Design Approach”, *Journal of Medical Systems*, Vol. 48/1, https://doi.org/10.1007/s10916-023-02024-0. [61]

Swenson, A. (2024), *FCC bans AI-generated voices in robocalls that can deceive voters*, https://www.pbs.org/newshour/politics/fcc-bans-ai-generated-voices-in-robocalls-that-can-deceive-voters. [262]

Tartaro, A. (2023), “When things go wrong: the recall of AI systems as a last resort for ethical and lawful AI”, *AI and Ethics*, https://doi.org/10.1007/s43681-023-00327-z. [281]

Taylor, H. (2023), *Ministers not doing enough to control AI, says UK professor*, https://www.theguardian.com/technology/2023/may/13/ministers-not-doing-enough-to-control-ai-says-uk-professor. [295]

Tocchetti, A. et al. (2022), *A.I. Robustness: a Human-Centered Perspective on Technological Challenges and Opportunities*, https://arxiv.org/abs/2210.08906. [292]

Trager, R. and L. Luca (2022), *Killer Robots Are Here—and We Need to Regulate Them*, https://foreignpolicy.com/2022/05/11/killer-robots-lethal-autonomous-weapons-systems-ukraine-libya-regulation/. [259]

Tsai, L. et al. (2024), *Generative AI for Pro-Democracy Platforms*, https://mit-genai.pubpub.org/pub/mn45hexw/release/1. [85]

UC Berkeley (2021), *Positive AI Economic Futures*, World Economic Forum, https://www.weforum.org/reports/positive-ai-economic-futures. [33]

Ugale, G. and C. Hall (2024), “Generative AI for anti-corruption and integrity in government: Taking stock of promise, perils and practice”*, OECD Artificial Intelligence Papers*, No. 12, OECD Publishing, Paris, https://doi.org/10.1787/657a185a-en. [81]

UK CMA (2024), *AI Foundation Models Update paper*, https://assets.publishing.service.gov.uk/media/661941a6c1d297c6ad1dfeed/Update_Paper__1_.pdf. [169]

UK CMA (2023), *AI Foundation Models Initial Report*, https://assets.publishing.service.gov.uk/media/650449e86771b90014fdab4c/Full_Non-Confidential_Report_PDFA.pdf. [168]

UK DfE (2023), *New research paves way for Artificial Intelligence in education*, https://www.gov.uk/government/news/new-research-paves-way-for-artificial-intelligence-in-education. [100]

UK DHSC (2023), *£21 million to roll out artificial intelligence across the NHS*, https://www.gov.uk/government/news/21-million-to-roll-out-artificial-intelligence-across-the-nhs. [99]

UK DSIT (2024), *A pro-innovation approach to AI regulation: government response*, https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response. [286]

UK DSIT (2024), *Data centres to be given massive boost and protections from cyber criminals and IT blackouts*, https://www.gov.uk/government/news/data-centres-to-be-given-massive-boost-and-protections-from-cyber-criminals-and-it-blackouts. [227]

UK DSIT (2024), *Frontier AI Safety Commitments, AI Seoul Summit 2024*, https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024. [92]

UK DSIT (2024), *Global leaders agree to launch first international network of AI Safety Institutes to boost cooperation of AI*, https://www.gov.uk/government/news/global-leaders-agree-to-launch-first-international-network-of-ai-safety-institutes-to-boost-understanding-of-ai. [299]

UK DSIT (2024), *International scientific report on the safety of advanced AI: interim report*, https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai. [115]

UK DSIT (2023), *Capabilities and risks from frontier AI*, UK Department for Science, Innovation & Technology, https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf. [110]

UK DSIT (2023), *Leading frontier AI companies publish safety policies*, https://www.gov.uk/government/news/leading-frontier-ai-companies-publish-safety-policies. [268]

UK DSIT (2023), *The Bletchley Declaration by Countries Attending the AI Safety Summit*, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023. [298]

UK FCDO (2023), *UK unites with global partners to accelerate development using AI*, https://www.gov.uk/government/news/uk-unites-with-global-partners-to-accelerate-development-using-ai. [91]

UK Government Office for Science (2023), *Future Risks of Frontier AI*, https://assets.publishing.service.gov.uk/media/653bc393d10f3500139a6ac5/future-risks-of-frontier-ai-annex-a.pdf. [304]

UK NCSC (2023), *National Cyber Security Centre*, https://www.ncsc.gov.uk/news/uk-develops-new-global-guidelines-ai-security. [220]

UKRI (2024), *£100m boost in AI research will propel transformative innovations*, https://www.ukri.org/news/100m-boost-in-ai-research-will-propel-transformative-innovations/. [89]

UN (2024), *Governing AI for Humanity*, https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf. [258]

UN (2023), *Resolution adopted by the General Assembly: Lethal autonomous weapons systems*, https://documents.un.org/doc/undoc/gen/n23/431/11/pdf/n2343111.pdf?token=AscNJ007h9iTSBRZj3&fe=true. [265]

UNESCO (2019), *Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development*, https://repositorio.minedu.gob.pe/bitstream/handle/20.500.12799/6533/Artificial%20intelligence%20in%20education%20challenges%20and%20opportunities%20for%20sustainable%20development.pdf. [212]

UNESCO/OECD/IDB (2022), *The Effects of AI on the Working Lives of Women*, United Nations Educational, Scientific and Cultural Organization, Paris, https://doi.org/10.1787/14e9b92c-en. [27]

UNODA (ed.) (2023), *2023 Group of Governmental Experts on Lethal Autonomous Weapons Systems*, https://meetings.unoda.org/ccw-/convention-on-certain-conventional-weapons-group-of-governmental-experts-on-lethal-autonomous-weapons-systems-2023 (accessed on 2024). [263]

US Department of State (2023), *Artificial Intelligence for Accelerating Progress on the Sustainable Development Goals: Addressing Society's Greatest Challenges*, https://www.state.gov/artificial-intelligence-for-accelerating-progress-on-the-sustainable-development-goals-addressing-societys-greatest-challenges/. [45]

US Department of State (2023), *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy*, https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2. [264]

US NSTC (2023), *National Artificial Intelligence Research and Development Strategic Plan 2023 Update*, Select Committee on Artificial Intelligence and the National Science and Technology Council, https://www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf. [290]

VARMA (2023), *Automated decision-making*, https://www.varma.fi/en/automated-decision-making. [93]

Vasey, B. (2022), "Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI", *BMJ*, Vol. 377:e070904, p. 377, https://doi.org/10.1136/bmj-2022-070904. [57]

Villasenor, J. (2019), *Products liability law as a way to address AI harms*, https://www.brookings.edu/articles/products-liability-law-as-a-way-to-address-ai-harms/. [248]

Vincent, J. (2023), *Google and Microsoft's chatbots are already citing one another in a misinformation shitshow*, https://www.theverge.com/2023/3/22/23651564/google-microsoft-bard-bing-chatbots-misinformation. [136]

Vinuesa, R. et al. (2020), "The role of artificial intelligence in achieving the Sustainable Development Goals", *Nature communications*, https://www.nature.com/articles/s41467-019-14108-y. [40]

Vipra, J. and S. Myers West (2023), *Computational Power and AI*, https://ainowinstitute.org/publication/policy/compute-and-ai. [150]

Wang, Y. and M. Kosinski (2018), "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.", *Journal of Personality and Social Psychology*, Vol. 114/2, pp. 246-257, https://doi.org/10.1037/pspa0000098. [185]

WEF (2024), *Shaping the Future of Learning: The role of AI in Education 4.0*, World Economic [38]

Forum, https://www3.weforum.org/docs/WEF_Shaping_the_Future_of_Learning_2024.pdf.

WEF (2023), *Can AI transform learning for the world's most marginalized children?*, World Economic Forum, https://www.weforum.org/agenda/2023/10/ai-education-learning-marginalized-unicef/. [65]

White House (2024), *A Call to Service for AI Talent in the Federal Government*, https://www.whitehouse.gov/ostp/news-updates/2024/01/29/a-call-to-service-for-ai-talent-in-the-federal-government/. [240]

White House (2024), *Advancing the Responsible Acquisition of Artificial Intelligence in Government*, https://www.whitehouse.gov/wp-content/uploads/2024/10/M-24-18-AI-Acquisition-Memorandum.pdf. [313]

White House (2024), *Supercharging Research: Harnessing Artificial Intelligence To Meet Global Challenges*, https://www.whitehouse.gov/wp-content/uploads/2024/04/AI-Report_Upload_29APRIL2024_SEND-2.pdf. [46]

White House (2023), *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. [90]

White House (2023), *Request for Information; National Priorities for Artificial Intelligence*, https://www.federalregister.gov/documents/2023/05/26/2023-11346/request-for-information-national-priorities-for-artificial-intelligence. [312]

White House (2023), *Voluntary AI Commitments*, https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf. [221]

White House (2022), *Blueprint for an AI Bill of Rights*, https://www.whitehouse.gov/ostp/ai-bill-of-rights/. [311]

Yamin, M. et al. (2021), "Weaponized AI for cyber attacks", *Journal of Information Security and Applications*, Vol. 57, p. 102722, https://doi.org/10.1016/j.jisa.2020.102722. [120]

Zhang, B. et al. (2022), "Forecasting AI Progress: Evidence from a survey of machine learning researchers", https://arxiv.org/abs/2206.04132. [318]

Zhang, Z., L. Wang and C. Lee (2023), "Recent Advances in Artificial Intelligence Sensors", *Advanced Sensor Research*, Vol. 2/8, p. 2200072, https://doi.org/10.1002/adsr.202200072. [54]

Zhi-Xuan, T. et al. (2024), *Beyond Preferences in AI Alignment*, https://arxiv.org/abs/2408.16984. [163]

Zwetsloot, R. and A. Dafoe (2019), *Thinking About Risks From AI: Accidents, Misuse and Structure*, https://www.lawfaremedia.org/article/thinking-about-risks-ai-accidents-misuse-and-structure. [178]

Notes

[1] See https://oecd.ai/en/site/ai-futures.for more information on the work and outputs of the Expert Group.

[2] As emphasised by the work of the OECD Strategic Foresight Unit (SFU) (https://www.oecd.org/en/about/programmes/strategic-foresight) and Observatory of Public Sector Innovation (OPSI) (https://oecd-opsi.org/work-areas/anticipatory-innovation).

[3] See https://alphafold.ebi.ac.uk.

[4] AGI refers to hypothetical future AI systems that exceed human-level intelligence across a broad spectrum of domains and contexts. There is substantial debate and uncertainty amongst experts about when or if such systems might be developed. Recent expert surveys typically collect probability distributions on the emergence of AGI or "human-level" machine intelligence (HLMI) or different AI milestones. In these, respondents generally forecasts a median probability of 50% that human-level AI is likely to be achieved in the second half of the 21st century, with many surveys hovering around the year 2060 (Stein-Perlman, Weinstein-Raun and Grace, 2022[316]; Karger et al., 2023[317]; Zhang et al., 2022[318]; Gruetzemacher, Paradice and Lee, 2019[16]). Crowdsourced forecasts yield results earlier than the expert surveys, with forecasting platform Metalculus—an online aggregator of results from a large community of forecasters—showing a 50% probability of "the first general AI system" being achieved by 2033, as of August 2024 (Metaculus, 2024[319]).

[5] See, for instance, Estonia's Government Office efforts to develop a tool for data-driven decision making in government: https://reform-support.ec.europa.eu/publications-0/government-data-driven-decision-making-dddm-framework-implementation_en.

[6] See examples at https://oecd-opsi.org/case_type/opsi/?_innovation_tags=artificial-intelligence-ai.

[7] This includes international efforts, such as the European Union's AI in Science brief and AI Act; the Council of Europe's Framework Convention on AI; outputs of the Group of Seven (G7), such as its Hiroshima AI Process; the Santiago Declaration among Latin American and Caribbean states; reporting from the UN AI Advisory Board; the Windhoek Statement on AI in Southern Africa; outcomes of the UK AI Safety Summit, such as the Bletchley Declaration on AI Safety; and global guidelines on AI security by the UK and US; among others. This also includes national efforts, such as national AI strategies, policies, guidance and statements in Brazil (national AI strategy), Canada (proposed AI and Data Act – AIDA, Generative AI guide), India (national AI strategy), Israel (AI Policy – Regulation and Ethics), Namibia (Ministerial statement), the United Arab Emirates (National Programme for AI) the United Kingdom (Pro-innovation approach to AI regulation, AI for Development), the United States (Executive Order on AI, Blueprint for an AI Bill of Rights, voluntary AI commitments for companies), and joint UK-US (guidelines for AI security) among others included in the OECD.AI Database of National AI Policies & Strategies.

[8] https://www.cnrs.fr/en/cnrsinfo/aissai-centre-crossroads-science-and-ai.

[9] There are various definitions of “foundation model” in use today. For example, the US *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* defines it as "an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts” (White House, 2023[90]).

[10] See https://aiforgood.itu.int and https://ircai.org, respectively.

[11] *Misinformation* refers to false or misleading information that is shared unknowingly and is not intended to deliberately deceive, manipulate or inflict harm on a person, social group, organisation or country. Importantly, the spreader does not create or fabricate the initial misinformation content. *Disinformation* refers to verifiably false or misleading information that is knowingly and intentionally created and shared for economic gain or to deliberately deceive, manipulate or inflict harm on a person, social group, organisation or country. Fake news, synthetic media, including deepfakes, and hoaxes are forms of disinformation, among others. See (OECD, 2024[126]) for additional information.

[12] For instance, the demand for Graphics Processing Units (GPUs) has reached unprecedented levels, further increasing costs. It can be difficult to identify the cost of these chips, as there are multiple distribution outlets, but they are likely to be in the range of multiple thousands of euros (OECD, 2024[320]).

[13] The index does not distinguish between legitimate and illegitimate uses of AI surveillance techniques. Rather, the purpose of the research is to show how new surveillance capabilities are transforming governments’ ability to monitor and track individuals or groups.

[14] This includes civil society organisations (see www.accessnow.org/campaign/ban-biometric-surveillance, https://reclaimyourface.eu, www.accessnow.org/wp-content/uploads/2021/11/joint-statement-EU-AIA.pdf and https://tiremeurostodasuamira.org.br), EU enforcement bodies (see https://t.ly/OV-z8) and UN Special Rapporteurs (see https://spcommreports.ohchr.org/TMResultsBase/DownLoadPublicCommunicationFile?gId=27594)

[15] https://www.oecd.org/stories/dis-misinformation-hub.

[16] See https://www.ftc.gov/legal-library/browse/joint-statement-competition-generative-ai-foundation-models-ai-products.

[17] See https://www.oecd.org/competition/market-concentration.htm and https://www.oecd.org/daf/competition/market-power-in-the-digital-economy-and-competition-policy.htm, respectively.

[18] See https://sandbox.datos.gov.co, https://www.kratid.ee/en/kratitoe-portfell, https://www.cnil.fr/en/sandbox-cnil-launches-call-projects-artificial-intelligence-public-services, https://www.datatilsynet.no/en/regulations-and-tools/sandbox-for-artificial-intelligence, https://espanadigital.gob.es/lineas-de-actuacion/sandbox-regulatorio-de-ia, https://t.ly/DhzJo and https://www.meti.go.jp/shingikai/mono_info_service/governance_model_kento/pdf/20220808_2.pdf, respectively.

[19] See https://www.darpa.mil/program/explainable-artificial-intelligence.

[20] See examples in Switzerland (https://www.admin.ch/gov/fr/accueil/documentation/communiques.msg-id-81319.html), the UK (https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence) and the US (https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf).

[21] See the OECD AI Incidents Monitor (AIM), https://oecd.ai/incidents.

[22] There are various theoretical and regulatory policy approaches to designing optimal incentive structures for preventing harm and compensating damages when they occur. The two main policy design options are ex-ante regulation, which focuses on safety rules, and ex-post regulation, which addresses liability and compensation. Ex-ante regulation aims to prevent harm by mandating a certain level of checks before products or services are introduced to the market, while tort law (liability) compensates victims when damage has occurred. Safety and liability rules complement each other, as both aim to allocate risks between developers and users and to incentivize investments in product safety, albeit at different stages. The EU AI Act, for example, adopts a product safety approach, requiring providers of high-risk AI systems to comply with specific mandatory requirements before the AI system can be introduced to the EU market, thus reducing the probability of liability and litigation after the AI system is in use.

[23] See https://www.consilium.europa.eu/en/press/press-releases/2024/10/10/eu-brings-product-liability-rules-in-line-with-digital-age-and-circular-economy.

[24] Section 5 of the FTC Act grants the FTC power to investigate and prevent deceptive trade practices. See https://www.federalreserve.gov/boarddocs/supmanual/cch/200806/ftca.pdf.

[25] See https://idais.ai.

[26] See https://www.gov.il/BlobFolder/policy/ai_2023/en/Israels%20AI%20Policy%202023.pdf.

[27] See https://rm.coe.int/20240704-ecn-9-2024-webinar-huderia/1680b0d26c.

[28] See https://www.iso.org/standard/81230.html.

[29] See https://oecd.ai/wonk/seeking-your-views-public-consultation-on-risk-thresholds-for-advanced-ai-systems-deadline-10-september.

[30] See https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response and https://www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research, respectively.

[31] See https://oe.cd/oecd-un-ai-announcement.

[32] Though these topics are important, they are not described as in-depth as some other topics in this report because the OECD has covered them extensively in previous products. For instance, see https://oecd.ai/en/wonk/ai-principles/22-transparency-and-explainability and (OECD, 2023[321]).

[33] See https://www.gov.uk/government/topical-events/ai-safety-summit-2023, https://aiseoulsummit.kr, and https://www.elysee.fr/en/emmanuel-macron/2024/05/22/gathering-of-frances-top-ai-talents, respectively.

[34] See https://digital-strategy.ec.europa.eu/en/policies/ai-office, https://aisi.go.jp, https://t.ly/vCtd1, https://www.gov.uk/government/publications/ai-safety-institute-overview, and https://www.nist.gov/aisi, respectively.

[35] See https://t.ly/9Vag8, https://oecd-auditors-alliance.org/content/auditing-algorithms, https://guides.etalab.gouv.fr/algorithmes, https://www.gob.mx/cms/uploads/attachment/file/415644/Consolidado_Comentarios_Consulta_IA__1_.pdf, https://www.gov.uk/government/news/uk-government-publishes-pioneering-standard-for-algorithmic-transparency, https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/politicas-y-gestion/evaluacion-impacto-algoritmico, https://www.gao.gov/products/gao-21-519sp, and https://www.auditingalgorithms.net, respectively.

[36] See interactive map at https://oecd.ai/en/dashboards/ai-principles/P7.

[37] See https://oe.cd/reinforcing-democracy-initiative (OECD, 2024[315]).

[38] See https://oe.cd/comp-mktst for dedicated OECD efforts on Market Studies and Competition.

[39] See, for example, outputs from 2023 (UK CMA[168]) and 2024 (UK CMA, 2024[169]). See Chapter 4 in (OECD, 2024[142]) for details on other public initiatives considering competition in AI markets.

[40] See https://oecd.ai/wonk/futures.

[41] See https://easyretro.io/publicboard/C6taCu8yMVcTOzJWC61GaOzIJNT2/ebbbf6e1-0f1d-4da4-b6fc-4d32c7ecd6dd for a version of the feedback board that includes the potential future benefits and risks, and https://easyretro.io/publicboard/Lg97hwaJe8MJWJTe5uKGfjjtInh1/f63b59a0-0531-456e-82f2-d9bea1463fb9 for a version containing the potential future policy actions. These do not include the feedback provided by Expert Group members. The final set of items covered in this report reflects that final version of the risks, benefits and policy actions, taking into account feedback from the experts and subsequent discussions and reviews of the report.

[42] https://oecd.ai/en/network-of-experts/ai-futures/discussions/future-benefits-risks.

[43] A publicly accessible version of this survey is available at https://oe.cd/ai-futures-survey.