A guide for policy-makers:

Evaluating rapidly developing technologies including AI, large language models and beyond

International Science Council

This paper is an update of “A framework for evaluating rapidly developing digital and related technologies: AI, large language models and beyond”. This paper is published by The International Science Council, 5 rue Auguste Vacquerie, 75116 Paris, France

To cite this report:
International Science Council 2024. A guide for policy-makers: Evaluating rapidly developing technologies including AI, large language models and beyond, Paris, France, International Science Council.
DOI: 10.24948/2024.07
https://council.science/publications/policy-makers-guide-framework-digital-technologies/

About the International Science Council

The International Science Council (ISC) works at the global level to catalyze and convene scientific expertise, advice and influence on issues of major concern to both science and society. The ISC has a growing global membership that brings together over 240 organizations, including international scientific unions and associations from natural and social sciences, and the humanities, and national and regional scientific organizations such as academies and research councils.

The International Science Council is exploring AI and new technologies as part of its initiatives in its Action Plan, such as *Converging Science and Technology in a Digital Era*, and through its think tank, the *Centre for Science Futures*.

Cover Photograph: tadamichi on iStock

Contents

Introduction

Rapidly emerging technologies present challenging issues when it comes to their use, governance and potential regulation. The ongoing policy and public debates on artificial intelligence (AI) and its use have brought these issues into acute focus. Broad principles for AI have been announced by UNESCO, the OECD, the UN and others, including the United Kingdom's Bletchley Declaration, and there are emerging jurisdictional attempts at regulation of aspects of the technology through, for example, the European Union (EU) AI Act or the recent United States AI Executive Order.

While the use of AI is discussed at length in these and other fora, across geopolitical divides and in countries at all income levels, there remains an ontological gap between the development of high-level principles and their incorporation into practice through either regulatory, policy, governance or stewardship approaches. The path from principle to practice is poorly defined, but given the nature and cadence of AI development and application, the variety of interest involved and the range of possible applications, any approach cannot be overly generic or prescriptive.

For these reasons, the non-governmental scientific community continues to play a particular role. The International Science Council (ISC) – with its pluralistic membership from the social and natural sciences – released a discussion paper in October 2023 presenting a preliminary analytical framework that considered the risks, benefits, threats and opportunities associated with rapidly moving digital technology. While it was developed to consider AI, it is inherently technology agnostic and can be applied to a range of emerging and disruptive technologies, such as synthetic biology and quantum. That discussion paper invited feedback from academics and policy-makers. The overwhelming feedback made conducting such an analysis necessary and stood as a valuable approach to address emerging technologies like AI.

The purpose of the framework is to provide a tool to inform all stakeholders – including governments, trade negotiators, regulators, civil society and industry – of the evolution of these technologies to help them frame how they might consider the implications, positive or negative, of the technology itself, and more specifically its particular application. This analytical framework has been developed independent of government and industry interests. It is maximally pluralistic in its perspectives, encompassing all aspects of the technology and its implications based on extensive consultation and feedback.

This discussion paper for policy-makers is not intended to proscribe a regulatory regime, but rather to suggest an adaptive and evolving analytical framework which could underpin any assessment and regulatory processes that might be developed by stakeholders, including governments and the multilateral system.

As decision-makers globally and nationally consider appropriate policy settings and levers to balance the risks and rewards of a new technology such as AI, the analytical framework is intended as a complementary tool to ensure the full suite of potential implications are adequately reflected.

Background: why an analytical framework?

The rapid emergence of technologies with the complexity and implications of AI is driving many claims of great benefit. However, it also provokes fears of significant risks, from individual to geostrategic level.[1] Much of the discussion to date has been considered in a binary sense as publicly expressed views tend to take place at the extreme ends of the spectrum. The claims made for or against AI are often hyperbolic and – given the nature of the technology – difficult to assess.

A more pragmatic approach is necessary where hyperbole is replaced with calibrated and more granular evaluations. AI technology will continue to evolve, and history shows that virtually every technology has both beneficial and harmful uses. The question is, therefore: how can we achieve the beneficial outcomes from this technology, while reducing the risk of harmful consequences, some of which could be existential in magnitude?

The future is always uncertain, but there are sufficient credible and expert voices regarding AI and generative AI to encourage a relatively precautionary approach. In addition, a systems approach is necessary as AI is a class of technologies with broad use and application by multiple types of users. This means that the full context must be considered when considering the implications of an any AI use for individuals, social life, civic life, societal life and in the global context.

Unlike most other technologies, for digital and related technologies, the time between development, release and application is extremely short, largely driven by the interests of the production companies or agencies. By its very nature – and given it is based on the digital backbone – AI will have applications that are rapidly pervasive, as has already been seen with the development of large language models. As a result, some properties may only become apparent after release, meaning there is the risk of unforeseen consequences, both malevolent and benevolent.

Important societal values dimensions, particularly across different regions and cultures, will influence how any use is perceived and accepted. Furthermore, geostrategic interests are already dominating discussion, with sovereign and multilateral interests continuously intersecting and thus driving competition and division.

To date, much of the regulation of a virtual technology has largely been seen through the lens of ‘principles’ and voluntary compliance, although with the EU AI Act[2] and similar we are seeing a shift to more enforceable but somewhat narrow regulations. Establishing an effective global or national technology governance and/or regulatory system remains challenging and there is no obvious solution. Multiple layers of risk-informed decision-making will be needed along the chain, from inventor to producer, to user, to government and to the multilateral system.

While high-level principles have been promulgated by UNESCO, the OECD, the European Commission and the UN, amongst others, and various high-level discussions continue regarding issues of potential governance, regulation, ethics and safety, there is a large gap between such principles and a governance or regulatory framework. This needs to be addressed.

As a starting point, the ISC considers developing a taxonomy of considerations that any developer, regulator, policy adviser, consumer or decision-maker could reference. Given the broad implications of these technologies, such a taxonomy must consider the totality of implications rather than a narrowly focused framing. Global fragmentation is increasing due to the influence of geostrategic interests on decision-making, and given the urgency of this technology, it is essential for independent and neutral voices to persistently champion a unified and inclusive approach.

---

[1] Hindustan Times. 2023. *G20 must set up an international panel on technological change*. https://www.hindustantimes.com/opinion/g20-must-set-up-an-international-panel-on-technological-change-101679237287848.html

[2] The EU Artificial Intelligence Act. 2023. https://artificialintelligenceact.eu/

The development of an analytical framework

The ISC is the primary global non-governmental organization integrating natural and social sciences. Its global and disciplinary reach means it is well placed to generate independent and globally relevant advice to inform the complex choices ahead, particularly as the current voices in this arena are largely from industry or from the policy and political communities of the major technological powers.

Following a period of extensive discussion, which included the consideration of a non-governmental assessment process, the ISC concluded that its most useful contribution would be to produce an adaptive analytic framework that can be used as the basis for discourse and decision-making by all stakeholders, including during any formal assessment processes that emerge.

The preliminary analytical framework, which was released for discussion and feedback in October 2023, took the form of an overarching checklist designed for use by both government and non-governmental institutions. The framework identified and explored the potential of a technology such as AI and its derivatives through a wide lens that encompasses human and societal wellbeing, as well as external factors such as economics, politics, the environment and security. Some aspects of the checklist may be more relevant than others, depending on the context, but better decisions seem more likely if all domains are considered, even if some can be quickly identified as irrelevant in particular cases. This is the inherent value of a checklist approach.

The preliminary framework was derived from previous work and thinking, including the International Network for Governmental Science Advice's (INGSA) report on digital wellbeing[3] and the OECD Framework for the Classification of AI Systems,[4] to present the totality of the potential opportunities, risks and impacts of AI. These previous products were more restricted in their intent given their time and context; there is a need for an overarching framework that presents the full range of issues both in the short and longer term.

Since its release, the discussion paper has received significant support from many experts and policy-makers. Many have specifically endorsed the recommendation to develop an adaptive framework that allows for deliberate and proactive consideration of the risks and implications of the technology, and in doing so, always considers the totality of dimensions from the individual to society and systems.

One key observation made through the feedback was acknowledgement that several of the implications considered in the framework are inherently multifaceted and extend across multiple categories. For example, disinformation could be considered from both the individual and geostrategic lens; thus, the consequences would be wide ranging.

The option to include case studies or exemplars to test the framework was also suggested. This could be used to develop guidelines to demonstrate how it could be used in practice in different contexts. However, this would be a significant undertaking and may confine how different groups perceive the use of this framework. It is best done by policy-makers working with experts in specific jurisdictions or contexts.

Since October 2023, there have been several significant national and multilateral initiatives with further consideration of the ethics and safety of AI. The implications of AI on the integrity of some of our critical systems, including financial, government, legal and education, as well as different knowledge systems (including scientific and indigenous knowledge), are of increasing concern. The revised framework further reflects these aspects.

The feedback received to date is reflected in the revised version of the analytical framework, which is now released as a guide to policy-makers.

While the framework is presented in the context of AI and related technologies, it is immediately transferable to the considerations of other rapidly emerging technologies such as quantum and synthetic biology.

---

[3] Gluckman, P. and Allen, K. 2018. Understanding wellbeing in the context of rapid digital and associated transformations. *INGSA*. https://ingsa.org/wp-content/uploads/2023/01/INGSA-Digital-Wellbeing-Sept18.pdf

[4] OECD. 2022. OECD Framework for the Classification of AI systems. *OECD Digital Economy Papers*, No. 323,#. Paris, OECD Publishing. https://oecd.ai/en/classification

The framework

The following table presents the dimensions of a putative analytic framework. Examples are provided to illustrate why each domain may matter; in context, the framework would require contextually relevant expansion. It is also important to distinguish between the generic issues that arise during platform developments and those that may emerge during specific applications. No single consideration included here should be treated as a priority and, as such, all should be examined.

The issues are broadly grouped into the following categories as outlined below:

- Wellbeing (including that of individuals or self, society, social life and civic life)
- Trade and economy
- Environmental
- Geostrategic and geopolitical
- Technological (system characteristics, design and use)[5]

The table details dimensions that might need to be considered when evaluating a new technology.

Dimensions of impact	Criteria	Examples of how this may be reflected in analysis
Individual / self	Users' AI competency	How competent and aware of the system's properties are the likely users who will interact with the system? How will they be provided with the relevant user information and cautions?
	Impacted stakeholders	Who are the primary stakeholders that will be impacted by the system (individuals, communities, vulnerable, sectoral workers, children, policy-makers, professionals etc.)?
	Optionality	Are users provided with the opportunity to opt out of the system or are they given opportunities to challenge or correct the output?
	Risks to human rights and democratic values	Does the system impact fundamentally on human rights, including but not limited to privacy, freedom of expression, fairness, non-discriminatory etc.?
	Potential effects on people's wellbeing	Do the system impact areas relate to the individual user's wellbeing (job quality, education, social interactions, mental health, identity, environment etc.)?
	Potential for human labour displacement	Is there a potential for the system to automate tasks or functions that were being executed by humans? If so, what are the downstream consequences?
	Potential for identity, values or knowledge manipulation	Is the system designed or potentially able to manipulate the user's identity or values set, or spread disinformation?
	Opportunities for self-expression and self-actualization	Is there a potential for artifice and self-doubt? Is there a potential for false or unverifiable claims of expertise?
	Measures of self-worth	Is there pressure to portray idealized self? Could automation replace a sense of personal fulfilment? Is there pressure to compete with the system in the workplace? Is individual reputation harder to protect against disinformation?
	Privacy	Are there diffused responsibilities for safeguarding privacy and are there any assumptions being made on how personal data are used?
	Autonomy	Could the AI system affect human autonomy by generating over-reliance by end-users?

[5] The technological criteria considered in the framework are specifically for AI and will need to be revised for other technologies as appropriate.

Dimensions of impact	Criteria	Description
Individual / self	Human development	Is there an impact on acquisition of key skills for human development, such as executive functions or interpersonal skills, or changes in attention time affecting learning, personality development, mental health concerns etc.?
	Personal health care	Are there claims of self-diagnosis or personalized health care solutions? If so, are they validated to regulatory standards?
	Mental health	Is there a risk of increased anxiety, loneliness or other mental health issues, or can the technology ameliorate such impacts?
	Human evolution	Could large language models and artificial general intelligence change the course of human evolution?
	Human-machine interaction	Could the use lead to de-skilling and dependency over time for individuals? Are there impacts on human interaction?
Society / social life	Societal values	Does the system fundamentally change the nature of society, enable normalization of ideas previously considered anti-social, or breach societal values of the culture in which it is being applied?
	Social interaction	Is there an effect on meaningful human contact, including emotional relationships?
	Population health	Is there a potential for the system to advance or undermine population health intentions?
	Cultural expression	Is an increase in cultural appropriation or discrimination likely or more difficult to address? Does reliance on the system for decision-making exclude or marginalize culturally relevant sectional ties of society?
	Public education	Is there an effect on teacher roles or education institutions? Does the system emphasize or reduce the digital divide and inequity among students? Is the intrinsic value of knowledge or critical understanding advanced or undermined?
	Distorted realities	Are the methods used to discern what is true still applicable? Is the perception of reality compromised?
Economic context (trade)	Industrial sector	In which industrial sector is the system deployed (finance, agriculture, health care, education, defence etc.)?
	Business model	In which business function is the system employed and in what capacity? Where is the system used (private, public, non-profit)?
	Impact on critical activities	Would a disruption of the system's function or activity affect essential services or critical infrastructures?
	Breadth of deployment	How is the system deployed (narrow use within unit vs. widespread nationally/international)?
	Technical maturity	How technically mature is the system?
	Interoperability	Are there likely to be silos, nationally or globally, that inhibit free trade and impact cooperation with partners?
	Technological sovereignty	Is a desire for technological sovereignty driving behaviours, including control over the entire AI supply chain?
	Income redistribution and national fiscal levers	Could the core roles of the sovereign state be compromised (e.g., reserve banks)? Will the state's ability to meet citizens' expectations and implications (social, economic, political etc.) be advanced or reduced?
	Digital divide (AI divide)	Are existing digital inequalities exacerbated or new ones created?
Civic life	Governance and public service	Could the governance mechanisms and global governance system be affected positively or negatively?
	News media	Is public discourse likely to become polarized and entrenched at a population level? Will there be an effect on the levels of trust in the Fourth Estate? Will conventional journalist ethics and integrity standards be further affected?

The framework

Dimensions of impact	Criteria	Description
Civic life	Rule of law	Will there be an effect on the ability to identify individuals or organizations to hold accountable (e.g., what kind of accountability to assign to an algorithm for adverse outcomes)? Is there a loss of sovereignty created (environmental, fiscal, social policy, ethics etc.)?
	Politics and social cohesion	Is there a possibility of more entrenched political views and less opportunity for consensus building? Is there the possibility of further marginalizing groups? Are adversarial styles of politics made more or less likely?
	Social licence	Are there privacy concerns, trust issues and moral concerns that need to be considered for stakeholder acceptance of the use?
	Indigenous knowledge	Could Indigenous knowledge and data be corrupted or misappropriated? Are there adequate measures to protect against misrepresentation, misinformation and exploitation?
	Scientific system	Is academic and research integrity compromised? Is there a loss of trust in science? Are there possibilities of misuse, overuse or abuse? What is the consequence of the practice of science?
Geostrategic / Geopolitical context	Precision surveillance	Are the systems trained on individual behavioural and biological data and could they be used to exploit individuals or groups?
	Digital colonization	Could state or non-state actors (e.g. large technology companies) harness systems and data to understand and control other countries' populations and ecosystems, or undermine jurisdictional control?
	Geopolitical competition	Could the system stir competition between nations over harnessing individual and group data for economic, medical and security interests?
	Shift in global powers	Is the status of nation-states as the world's primary geopolitical actors under threat? Do technology companies wield power once reserved for nation-states and have they become independent, sovereign actors (emerging technopolar world order)?
	Disinformation	Would the system facilitate the production and dissemination of disinformation by state and non-state actors with an impact on social cohesion, trust and democracy?
	Dual-use applications	Is there a possibility for both military application as well as civilian use?
	Fragmentation of global order	Could silos or clusters of regulation and compliance develop that hinder cooperation, lead to inconsistencies in application and create room for conflict?
Environmental	Energy and resource consumption (carbon footprint)	Do the system and requirements increase uptake of energy and resource consumption over and above the efficiency gains obtained through the application?
	Energy source	Where is the energy sourced from for the system (renewable vs. fossil fuels etc.)?
Data and input	Detection and collection	Are the data and input collected by humans, automated sensors or both?
	Provenance of the data	Are the data and input from experts provided, observed, synthetic or derived? Are there watermark protections to confirm provenance?
	Dynamic nature of the data	Are the data dynamic, static, dynamic updated from time to time or real-time?
	Rights	Are the data proprietary, public or personal (related to identifiable individuals)?
	Identifiability of personal data	If personal, are the data anonymized or pseudonymized?
	Structure of the data	Are the data structured, semi-structured, complex structured or unstructured?
	Format of the data	Is the format of the data and metadata standardized or non-standardized?
	Scale of the data	What is the dataset's scale?
	Appropriateness and quality of the data	Is the dataset fit for purpose? Is the sample size adequate? Is it representative and complete enough? How noisy are the data? Is it error prone?

Dimensions of impact	Criteria	Description
Model	Information availability	Is any information available about the system's model?
	Type of AI model	Is the model symbolic (human-generated rules), statistical (uses data) or hybrid?
	Rights associated with model	Is the model open-source or proprietary, self- or third-party managed?
	Single or multiple models	Is the system composed of one model or several interlinked models?
	Generative or discriminative	Is the model generative, discriminative or both?
	Model building	Does the system learn based on human-written rules, from data, through supervised learning or through reinforcement learning?
	Model evolution (AI drift)	Does the model evolve and/or acquire abilities from interacting with data in the field?
	Federated or central learning	Is the model trained centrally or in several local servers or 'edge' devices?
	Development/ maintenance	Is the model universal, customizable or tailored to the AI actor's data?
	Deterministic or probabilistic	Is the model used in a deterministic or probabilistic manner?
	Model transparency	Is information available to users to allow them to understand model outputs and limitations or use constraints?
	Computational limitation	Are there computational limitations to the system? Is it possible to predict capability jumps or scaling laws?
Task and output	Task(s) performed by system	What tasks does the system perform (recognition, event detection, forecasting etc.)?
	Combining tasks and actions	Does the system combine several tasks and actions (content generation systems, autonomous systems, control systems etc.)?
	System's level of autonomy	How autonomous are the system's actions and what role do humans play?
	Degree of human involvement	Is there some human involvement to oversee the overall activity of the AI system and the ability to decide when and how to use the AI system in any situation?
	Core application	Does the system belong to a core application area such as human language technologies, computer vision, automation and/or optimization or robotics?
	Evaluation	Are standards or methods available for evaluating system output?

Sources of the descriptors:

1. INGSA. 2018. Understanding wellbeing in the context of rapid digital and associated transformations. https://ingsa.org/wp-content/uploads/2023/01/INGSA-Digital-Wellbeing-Sept18.pdf

2. OECD Framework for the Classification of AI Systems: a tool for effective AI policies. https://oecd.ai/en/classification

3. New descriptors (sourced through the extensive consultation and feedback and literature review).

How could this framework be used?

This framework could be used in many ways, including:

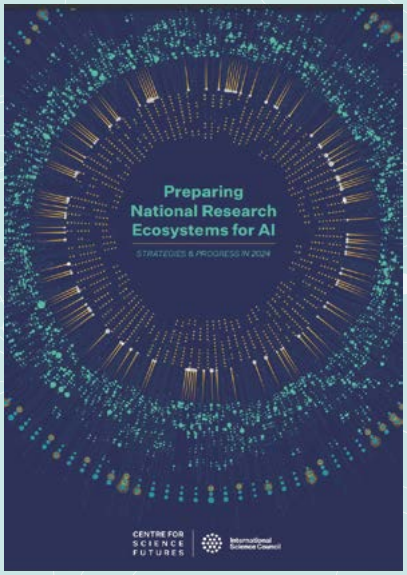
- To bridge the gap between high-level principles and assessment for regulatory or governance purposes. The framework can support this by establishing a validated common taxonomy of the range of issues that merit consideration by relevant stakeholders as a basis to inform and shape further thinking. For example, at a national level, the framework could be used as a tool by the government as it develops a national AI strategy and policies to establish a common basis of the risks and opportunities across stakeholder groups.
- To inform impact assessments. The EU AI Act requires organizations that provide AI tools or adopt AI in their processes to undertake an impact assessment to identify the risk of their initiatives and apply an appropriate risk management approach. The framework presented here could be used as a foundation for this.
- To inform horizon scanning for risks and future scenarios. The categorization of risks in the UN AI Advisory Body's Interim Report[6] is broadly aligned to the framing presented in the framework here. There is an opportunity for the framework to be used to build consensus and test the severity of emerging risks as well as pre-empt these.
- To enhance the ethical principles needed to guide and govern the use of AI. The framework can do this by providing a flexible foundation upon which trustworthy systems can be developed and ensuring the lawful, ethical, robust and responsible use of the technology. These principles could be tested against the full range of impacts presented in this framework.
- To facilitate a stocktake of existing and evolving measures (regulatory, legislative, policy, standards, governance etc.) and identify gaps that need further consideration. These could be mapped against the framework categories at a national or multinational level to determine gaps and identify suitable measures to mitigate the risks.
- To support government use of AI. As many governments determine their respective strategies for the use of AI within agencies and systems, the framework could be used to define appropriate risk thresholds and identify key stakeholders and responsibilities.
- To support public discourse and establish social licence on how AI is used and the underpinning data that will be used across government services or more broadly in society.

A way forward

In summary, the analytical framework is provided as the basis of a toolkit that could be used by stakeholders to comprehensively look at the any significant developments either of platforms or of use in a consistent and systematic manner. The dimensions presented in this framework have relevance from technology assessment to public policy, from human development to sociology, and futures and technology studies. While developed for AI, this analytical framework has much broader application to any other emerging technology.

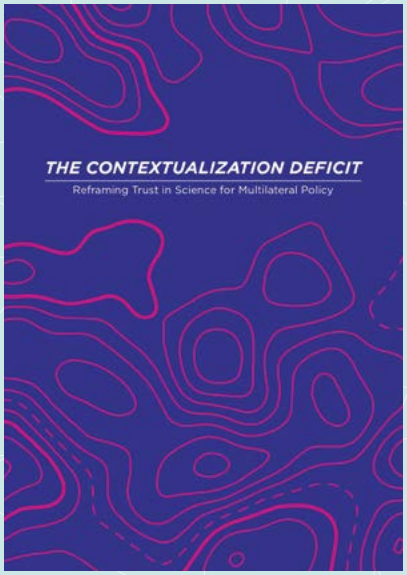
[6] UN AI Advisory Board. 2023. *Interim Report: Governing AI for Humanity*. https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf

You might be interested in



Preparing National Research Ecosystems for AI: Strategies and progress in 2024

This report from the ISC's Centre for Science Futures offers a comprehensive analysis of the integration of artificial intelligence in science and research across various countries. It addresses both the advancements made and the challenges faced in this field, making it a valuable read for science leaders, policy-makers, AI professionals, and academics. Read online in the language of your choice and engage with us to be part of the next edition.



The Contextualization Deficit: Reframing Trust in Science for Multilateral Policy

How can the multilateral policy interface effectively engage with science in ways that are trusted by populations? In order to build trust, this working paper, presented by the ISC's think tank, the Centre for Science Futures, in partnership with the UNESCO Unitwin Chair on Communication for Science as a Public Good, suggests the necessity of updating the science-policy interface model.

Drawing upon empirical evidence from the past 15 years, it explores new frameworks for envisioning what a healthy science-policy interface might entail and how it can effectively address the concerns that drive various communities to either support or oppose science-policy interventions.

Acknowledgements

Many people have been consulted and provided feedback in the development of both the initial discussion paper and feedback following its release. Both papers were drafted by Sir Peter Gluckman, President, the ISC and Hema Sridhar, former Chief Science Adviser, Ministry of Defence, New Zealand and now Senior Research Fellow, University of Auckland, New Zealand.

In particular, the ISC tLord Martin Rees, former President of the Royal Society and Co-Founder of the Centre for the Study of Existential Risks, University of Cambridge; Professor Shivaji Sondhi, Professor of Physics, University of Oxford; Professor K Vijay Raghavan, former Principal Scientific Adviser to the Government of India; Amandeep Singh Gill, the UN Secretary General's Envoy on Technology; Seán Ó hÉigeartaigh, Executive Director, Centre for the Study of Existential Risks, University of Cambridge; Sir David Spiegelhalter, Winton Professor of the Public Understanding of Risk, University of Cambridge; Amanda-June Brawner, Senior Policy Adviser and Ian Wiggins, Director of International Affairs, Royal Society, United Kingdom; Dr Jerome Duberry, Managing Director and Dr Marie-Laure Salles, Director, Geneva Graduate Institute; Chor Pharn Lee, Centre for Strategic Futures, Prime Minister's Office, Singapore; Barend Mons and Dr Simon Hodson, the Committee on Data (CoDATA); Professor Yuko Harayama, Former Executive Director, RIKEN; Professor Rémi Quirion, President, INGSA; Dr Claire Craig, University of Oxford and Former Head of Foresight, Government Office of Science; Prof Yoshua Bengio, UN Secretary General's Scientific Advisory Board and at Université de Montréal; and the many others who provided feedback to the ISC on the initial discussion paper.

International Science Council

Work with the ISC to advance science as a global public good.

Connect with us at:

council.science
secretariat@council.science
International Science Council
5 rue Auguste Vacquerie
75116 Paris, France

twitter.com/ISC
facebook.com/InternationalScience
instagram.com/council.science
linkedin.com/company/international-science-council