A framework for evaluating rapidly developing digital and related technologies:

AI, Large Language Models and beyond

International Science Council

This discussion paper is published by The International Science Council,
5 rue Auguste Vacquerie, 75116 Paris, France

To cite this report:
International Science Council 2023. A framework for evaluating rapidly developing digital and related technologies: AI, Large Language Models and beyond, Paris, France, International Science Council.
DOI: 10.24948/2023.11
https://council.science/publications/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 governance and potential regulation. The policy and public debates on artificial intelligence (AI) and its use have brought these issues into acute focus. While broad principles for AI have been promulgated by UNESCO, OECD and others, and there are nascent discussions regarding global or jurisdictional regulation of the technology, there is an ontological gap between the development of high-level principles and their incorporation into regulatory, policy, governance and stewardship approaches. This is where the non-governmental scientific community could have a particular role.

It has been proposed by a number of academics and policy experts that the International Science Council (ISC) – with its pluralistic membership from the social and natural sciences – establish a process to produce and maintain an annotated framework/checklist of the risks, benefits, threats and opportunities associated with rapidly moving digital technologies, including – but not limited to – AI. The purpose of the checklist would be to inform all stakeholders – including governments, trade negotiators, regulators, civil society and industry – of potential future scenarios, and would frame how they might consider the opportunities, benefits, risks and other issues.

The outputs would not act as an assessment body, but as 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. Any analytical framework should ideally be developed independent of governmental and industry claims, given their understandable interests. It must also be maximally pluralistic in its perspectives, thus encompassing all aspects of the technology and its implications.

This discussion paper provides the outline of an initial framework to inform the multiple global and national discussions taking place related to AI.

Background: why an analytical framework?

The rapid emergence of a technology with the complexity and implications of AI is driving many claims of great benefits. However, it also provokes fears of significant risks, from individual to geo-strategic level. Much of the discussion tends to take place at the extreme ends of the spectrum of views, and a more pragmatic approach is needed. 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 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 needed, because AI is a class of technologies with broad use and application by multiple types of users. This means that the full context must be taken into account when considering the implications of AI for individuals, social life, civic life, societal life and in the global context.

Unlike most past technologies, digital and related technologies have a very short period of time from development to release, largely driven by the interests of the production companies or agencies. AI is rapidly pervasive; some properties may only become apparent after release, and the technology could have both malevolent and benevolent applications. Important values dimensions will influence how any use is perceived. Furthermore, there may be geo-strategic interests at play.

To date, the regulation of a virtual technology has largely been seen through the lens of “principles” and voluntary compliance. More recently, however, the discussion has turned to issues of national and multilateral governance, including the use of regulatory and other policy tools. The claims made for or against AI are often hyperbolic and – given the nature of the technology – difficult to assess. Establishing an effective global or national technology regulation system will be challenging, and 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, OECD and the European Commission, amongst others, and various high-level discussions are underway regarding issues of potential regulation, there is a large ontological gap between such principles and a governance or regulatory framework. What is the taxonomy of considerations that a regulator might need to consider? A narrowly focused framing would be unwise, given the broad implications of these technologies. This potential has been the subject of much commentary, both positive and negative.

The development of an analytical framework

The ISC is the primary global NGO 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 major technological powers. Following extensive discussion over recent months, including the consideration of a non-governmental assessment process, the ISC concluded that its most useful contribution would be to produce and maintain an adaptive analytic framework that can be used as the basis for discourse and decision-making by all stakeholders, including during any formal assessment process that emerges.

This framework would take the form of an overarching checklist that could be used by both government and non-governmental institutions. The framework identifies and explores 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 are more likely if all domains are considered. This is the inherent value of a checklist approach.

The proposed framework is derived from previous work and thinking, including the International Network for Governmental Science Advice's (INGSA) digital wellbeing report[1] and the OECD AI Classification Framework[2] 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.

While developed for the consideration of AI, this analytical framework could be applied to any rapidly emerging technology. The issues are broadly grouped into the following categories for further examination:

- Wellbeing (including that of individuals or self, society and social life, and civic life)
- Trade and economy
- Environmental
- Geo-strategic and geo-political
- Technological (system characteristics, design and use)

[1] 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

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

A list of considerations for each of the above categories is included along with their respective opportunities and consequences. Some are relevant for specific instances or applications of AI while others are generic and agnostic of platform or use. No single consideration included here should be treated as a priority and, as such, all should be examined.

How could this framework be used?

This framework could be utilized in, but not limited to, the following ways:

- To bridge the gap between principles and assessment by establishing a validated common taxonomy of the range of considerations that could be utilized by relevant stakeholders as a basis to inform and shape further thinking, including any assessment framework that might be developed by authorities.
- 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 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 stock take of existing measures (i.e., regulatory, legislative, policy) and identify any gaps that needs further consideration.
- The framework is agnostic to the technology and its use. It could therefore be used in quite distinct areas such as synthetic biology.

The following table is an early shaping of the dimensions of an analytic framework. Depending on the technology and its use, some components will be more relevant than others. The 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 platform developments and the generic issues that may emerge during specific applications.

Dimensions to consider when evaluating a new technology

Initial draft of the 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 (i.e., individuals, communities, vulnerable, sectoral workers, children, policy-makers, professionals)?
	Optionality	Are users provided with an option to opt-out of the system; should they be given opportunities to challenge or correct the output?
	Risks to human rights and democratic values	Could the system impact (and in what direction) on human rights, including, but not limited to, privacy, freedom of expression, fairness, risk of discrimination, etc.?
	Potential effects on people's wellbeing	Could the system impact (and in what direction) the individual user's wellbeing (i.e., job quality, education, social interactions, mental health, identity, environment)?
	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 to or potentially able to manipulate the user's identity or values set, or spread disinformation? Is there a potential for false or unverifiable claims of expertise?
	Measures of self-worth	Is there pressure to portray an idealized self? Could automation replace a sense of personal fulfilment? Is there pressure to compete with the system in the workplace? Is individual reputation made harder to protect against disinformation?
	Privacy	Are there diffused responsibilities for safeguarding privacy and are there any assumptions being made on how personal data is utilized?
	Autonomy	Could the system affect human autonomy by generating over-reliance on the technology by end-users?
	Human development	Is there an impact on acquisition of key skills for human development such as executive functions, interpersonal skills, changes in attention time affecting learning, personality development, mental health concerns, etc.?
	Personal health care	Are there claims of 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 mitigate such impacts?
	Human evolution	Could the technology lead to changes in human evolution?
Society / social life	Societal values	Does the system fundamentally change the nature of society or enable the normalization of ideas previously considered anti-social, or does it breach the societal values of the culture in which it is being applied?
	Social interaction	Is there an effect on meaningful human contact, including emotional relationships?
	Equity	Is the application/technology likely to reduce or enhance inequalities (i.e., economic, social, educational, geographical)?

Dimensions of impact	Criteria	Description
Society / social life	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 potentially exclude or marginalize sections of society?
	Public education	Is there an effect on teacher roles or education institutions? Does the system emphasize or reduce inequity among students and the digital divide? Is the intrinsic value of knowledge or critical understanding advanced or undermined?
	Distorted realities	Are the methods we use to discern what is true still applicable? Is the perception of reality compromised?
Economic context (trade)	Industrial sector	Which industrial sector is the system deployed in (i.e., finance, agriculture, health care, education, defence)?
	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?
	Breath of deployment	How is the system deployed (narrowly within an organization vs widespread nationally/internationally)?
	Technical maturity (TRL)	How technically mature is the system?
	Technological sovereignty	Does the technology drive greater concentration of technological sovereignty.
	Income redistribution and national fiscal levers	Could the core roles of the sovereign state be compromised (i.e., Reserve Banks)? Will the state's ability to meet citizens' expectations and implications (i.e., social, economic, political) be advanced or reduced?
Civic life	Governance and public service	Could governance mechanisms and global governance systems be affected positively or negatively?
	News media	Is public discourse likely to become more or less polarized and entrenched at a population level? Will there be an effect on the levels of trust in the media? Will conventional journalism ethics and integrity standards be further affected?
	Rule of law	Will there be an effect on the ability to identify individuals or organizations to hold accountable (i.e., what kind of accountability to assign to an algorithm for adverse outcomes)? Does this create a loss of sovereignty (i.e., environmental, fiscal, social policy, ethics)?
	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?
Geo-strategic / geo-political context	Precision surveillance	Are the systems trained on individual behavioural and biological data, and if so, could they be used to exploit individuals or groups?
	Digital colonization	Are state or non-state actors able to harness systems and data to understand and control other countries' populations and ecosystems, or to undermine jurisdictional control?

Dimensions to consider when evaluating a new technology

Dimensions of impact	Criteria	Description
Geo-strategic / geo-political context	Geo-political competition	Does the system affect competition between nations and technology platforms for access to individual and collective data for economic or strategic purposes?
	Trade and trade agreements	Does the system have implications for international trade agreements?
	Shift in global powers	Is the status of nation-states as the world's primary geo-political actors under threat? Will technology companies wield power once reserved for nation-states and are they becoming independent sovereign actors?
	Disinformation	Is it easier for state and non-state actors to produce and disseminate disinformation that impacts social cohesion, trust and democracy?
Environmental	Energy and resource consumption (carbon footprint)	Does the system and requirements increase uptake of energy and resource consumption over and above the efficiency gains obtained through the application?
Data and input	Detection and collection	Are the data and input collected by humans, automated sensors or both?
	Provenance of the data	With regards to the data are these provided, observed, synthetic or derived? Are there watermark protections to confirm provenance?
	Dynamic nature of the data	Are the data dynamic, static, updated from time to time or updated in real-time?
	Rights	Are data proprietary, public or personal (i.e., related to identifiable individuals)?
	Identifiability of personal data	If personal data, are they 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 is the data? Is it error prone?
Model	Information availability	Is information about the system's model available?
	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?

Dimensions of impact	Criteria	Description
Model	Development and 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? Can we predict capability jumps or scaling laws?
Task and output	Task(s) performed by system	What tasks does the system perform (i.e., recognition, event detection, forecasting)?
	Combining tasks and actions	Does the system combine several tasks and actions (i.e., content generation systems, autonomous systems, control systems)?
	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 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 to evaluate system output or deal with unforeseen emergent properties?

Sources of the descriptors:

1. 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

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

3. New descriptors (from multiple sources)

A way forward

Depending on the response to this discussion paper, an expert working group would be formed by the ISC to further develop or amend the above analytical framework by which stakeholders might comprehensively look at any significant developments either of platforms or of use dimensions. The working group would be disciplinarily, geographically and demographically diverse, with expertise spanning from technology assessment to public policy, from human development to sociology and futures and technology studies.

To engage with this discussion paper, please visit council.science/publications/framework-digital-technologies/

Acknowledgements

Many people have been consulted in the development of this paper, which was drafted by Sir Peter Gluckman, President, ISC and Hema Sridhar, former Chief Scientist, Ministry of Defence, and now senior research fellow, University of Auckland, New Zealand.

In particular we thank Lord 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 VijayRaghavan, former principal scientific adviser to the Government of India; Amandeep Singh Gill, UN Secretary General's Envoy on Technology; Dr Seán Óh Éigeartaigh, Executive Director, Centre for the Study of Existential Risks, University of Cambridge; Amanda-June Brawner, Senior Policy Advisor, and Ian Wiggins, Director of International Affairs, Royal Society UK; Dr Jerome Duberry, Dr Marie-Laure Salles, Director, Geneva Graduate Institute; Mr Chor Pharn Lee, Centre for Strategic Futures, Prime Minister's Office, Singapore; Barend Mons and Dr Simon Hodson, the Committee on Data (CoDATA); Prof Yuko Harayama, Japan; Professor Rémi Quirion, President, INGSA; Dr Claire Craig, University of Oxford and Former Head of Foresight, Government Office of Science; and Prof Yoshua Bengio, UN Secretary General's Scientific Advisory Board and at Université de Montréal. The checklist approach was generally endorsed and the timeliness of any action by the ISC was emphasized.

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