



The AI Dialogues

Interim note #1

**Is international AI
governance achievable?**

The AI Dialogues

In 2024, [Renaissance Numérique](#), a leading independent French think tank dedicated to the digital transformation of society and its impacts on citizens, launched the [AI Dialogues](#), a three-day series bringing together European and international experts to discuss international, European and local governance issues.

This note is based on the discussions that took place during the first day of the *AI Dialogues*, on Friday, April 26 2024, at the University of Geneva (UNIGE), on the theme: “[Is international AI governance achievable?](#)”.

This is an interim note. A final report will be published at the end of the program, before the end of the year. Please feel free to [contact us](#) to comment on this note.

The arguments presented in this note do not necessarily reflect the position of the participants (see the list at the end of this document); they remain the editorial decision of Renaissance Numérique.















Funding Partners of the AI Dialogues	Academic Partners of the AI Dialogues
    	        

Table of contents

Introduction	3
What is AI governance?	4
Seven functions of governance	6
Building scientific consensus	6
Building political consensus and norms	7
Coordinating policy and regulations	8
Enforcing standards and restrictions	8
Three additional functions: emergency response, joint research and the distribution of benefits	9
The governance of AI: working with old institutions or building new ones?	10
Challenges with the governance of AI	12
Defining AI	12
Building consensus	15
Making AI systems accountable	16
Existing models for the governance of AI	17
Standardize the technical layers as a first step to explainability and regulate	18
Product safety: testing before selling	18
Content: building accountability mechanisms	19
Conclusion	20

Introduction

The advancement of artificial intelligence (AI) has fostered the development of a new era of innovation and capability. Virtually all actors can benefit from systems that can optimize processes or accomplish specific tasks. But the ubiquitous nature of AI comes with challenges. For example, if unaccounted for, discriminations induced by algorithms can be detrimental to social groups; the hallucinations inherent to Large Language Models (LLMs) can lead to erroneous decisions with potentially severe consequences, especially in critical fields like medicine, finance and social policy. In addition to such short-term risks, others worry (Dafoe, 2020¹ ; Roose, 2023²) about long-term existential risks posed by intelligent systems escaping human control.

The cooperation between companies and regulators in the field of AI is essential to ensure these risks are mitigated. As new AI systems become concentrated in the hands of a few economic actors, in particular through vertical integrations (e.g. partnerships), new governance risks emerge. Regulation aims to distribute decision-making authority to

¹. Dafoe A. (2020). "AI Governance: Opportunity and Theory of Impact", Allandafoe.com, September, 2020. <https://www.allandafoe.com/opportunity>

². Roose K. (2023). "AI Poses 'Risk of Extinction,' Industry Leaders Warn", *The New York Times*, May 30, 2023., /2023/05/30 <https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html>

ensure that no single entity holds excessive control over AI deployment and impact. In parallel, significant efforts are made to hold powerful actors accountable, which requires building transparency to increase public trust and prevent misuse.

Various domestic and international spheres have debated the question of the governance of AI in order to both face the numerous challenges that AI poses and allow actors to make the most of the technology.

This paper, based on the first AI Dialogue organized by Renaissance Numérique on May 26, 2024, asks three questions that can help decision-makers navigate the field of AI governance:

1. What is AI governance?
2. What are some of the challenges it poses?
3. What other fields need governance?

What is AI governance?

Roberts *et al.* define global AI governance as « the process through which diverse interests that transcend borders are accommodated (...) so that cooperative action may be taken in maximizing the benefits and mitigating the risks of AI » (Roberts *et al.*, 2024)³.

Given the widespread use of AI, one of the main challenges is to build a robust, solution-driven, governance framework, while reflecting the various interests of the stakeholders involved. The effects of AI concern multiple actors, including states, public institutions, intergovernmental organizations, businesses, civil society organizations, academics, communities, individuals (citizens, users, etc.), who should all have a say in discussions about its regulation.

Intergovernmental organizations such as the Organisation for Economic Co-operation and Development (OECD), the United Nations Educational, Scientific and Cultural Organization (UNESCO), the International Telecommunication Union (ITU) or the G7 group have endorsed this role. They produced ethical frameworks and guidelines that serve as normative building blocks for the governance of AI. Furthermore, the European Union (EU)⁴ and some states, in particular the United States⁵, have tried to take the lead on this issue, in particular by establishing voluntary commitments on AI.

³. Roberts H., Hine E., Taddeo M. & Floridi L. (2024). Global AI governance: Barriers and pathways forward, *International Affairs*, Volume 100, Issue 3, May 2024, pp. 1275–1286, <https://doi.org/10.1093/ia/iiae073>

⁴. European AI Office: <https://digital-strategy.ec.europa.eu/en/policies/ai-office>

⁵. White House (2023). “Voluntary AI Commitments”. Online (PDF): <https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf>

As computing power advanced in the late 2010s, these discussions gained prominence. With the recent advent of generative AI, they have taken on an increasingly global dimension. On November 1st and 2nd, 2023, the leaders of the world met in the UK for the [AI Safety summit](#). On May 21, 2024, they met again in South Korea for the [AI Seoul Summit](#). The next international summit gathering both political economic leaders will take place during the AI Action Summit in Paris in 2025. There are also other global initiatives such as the AI for Good Global Summit.⁶

	AI Safety Summit (London, november 2023)	AI Seoul Summit (Seoul, may 2024)	AI Action Summit (Paris, february 2025)
Focus	Emphasis on frontier AI safety	Emphasis on frontier AI safety	Emphasis on AI opportunities
Outcomes	Bletchley Declaration , signed by 28 countries	Frontier AI Safety Commitments , signed by 16 companies Seoul Statement of Intent toward International Cooperation on AI Safety Science , signed by 10 countries and the EU Seoul Ministerial Statement , signed by 27 countries and the EU	N/A
Key takeaways	The next generation of AI models should be tested before they are released AI offices should help Voluntary commitments should be put on a legal or regulatory footing International standards for safety are needed	Companies commit to: - develop methods to identify and mitigate risks - not develop AI tools when risk is too high - be transparent, “except insofar as doing so would increase risk or divulge sensitive commercial information to a degree disproportionate to the societal benefit” Statement of Intent stresses need for cooperation between new AI institutes	N/A

Figure 1: Comparison of three AI Summits

⁶. AI for Good Global Summit: <https://aiforgood.itu.int>

Although these initiatives attempt to set important norms to promote a responsible usage of AI, they are only one part (arguably the most visible) of AI governance.

This section identifies the multiple facets of governance based on a review of literature. It does so by citing examples from other fields whilst highlighting the specificities of AI. Although it is drawn from the scientific literature, this kind of typology can be criticized for various reasons so it must be considered with caution. Nevertheless, we believe that it provides an heuristic entry point for a fairly simple exploration and understanding of the goals, roles and possible actions of the different actors in an international governance of AI. These functions and the actors are meant to be discussed during the second day of the AI Dialogues.

Seven functions of governance

A number of institutions come to mind when thinking of global governance models: the European Commission, the International Telecommunication Union (ITU), the World Health Organization (WHO), the World Trade Organization (WTO), the International Energy Agency (IEA), the International Atomic Energy Agency (IAEA) and the International Panel on Climate Change (IPCC) are all examples of international institutions that play a governance role. However what they do differs significantly. Overall, we highlight seven functions of governance, put forward in the report *International AI Institutions: a literature review of models, examples, and proposals* (Maas, Villalobos, 2023)⁷.

Building scientific consensus

Emerging technologies impact society in various ways. On March 2023, the Future of life Institute published an open letter calling all AI labs to postpone for a short delay the training of AI systems with high capabilities⁸. The letter, which was also signed by academic AI researchers, cited risks such as AI-generated propaganda, extreme automation of jobs, human obsolescence, and a society-wide loss of control. Although difficult in some cases, an objective assessment of these impacts is a necessary step for effective regulation.

It took several decades for the scientific community on climate change to standardize indicators and methods to evaluate the environmental footprint of human activity. Taking inspiration from the IPCC, several organizations now seek to establish a scientific

⁷. Maas M.M, Villalobos J.J. (2023). International AI Institutions: A Literature Review of Models, Examples, and Proposals, *AI Foundations Report 1*, Available at SSRN: <https://ssrn.com/abstract=4579773>

⁸. Future of life Institute (2023). "Pause Giant AI Experiments: An Open Letter". Online: <https://futureoflife.org/open-letter/pause-giant-ai-experiments>

consensus on the impact of information and communication technologies on society, such as the International Panel on the Information Environment and the Observatory on Information and Democracy.

AI could be a good candidate for such an assessment and the French AI Commission has recommended that a new organization work towards this goal in its 2024 report (Aghion, Bouverot, 2024).⁹ The widespread use of AI in society raises questions that the scientific community can answer. Amongst many examples, such questions could be how to build more transparent and accountable AI models, how to measure the effects of the use of deepfakes on election cycles, or what are the effects of the malign use of LLMs on the level of cybersecurity of companies.

Building political consensus and norms

Political consensus between states is necessary to align national policymaking responses or to initiate further regulations. It can involve negotiations around the best practices to address a specific issue or the set of values that all actors engaging in an activity should adhere to. Political consensus can be transcribed in hard laws that come with sanctions, or take the form of soft laws that indicate a preferred point of view. The UN¹⁰, the OECD¹¹, the Council of Europe¹² and the G7, but also the International Standardization Organisations (ISO)¹³ and the International Telecommunications Union (ITU) are all examples of international institutions seeking to build political consensus.

In this context, some private players have already implemented a set of principles that guide their development activities (e.g. Google, Microsoft, etc.). A political consensus is therefore needed to establish the best common set of values that respect human and fundamental rights, to which developers and providers of AI systems should adhere, and to make each actor in the chain accountable for.

As technology rapidly evolves, frameworks will continue to change and new challenges will emerge for policymakers to consider. As an example, states and international organizations are increasingly adopting new principles on the responsible military and defense-related use of artificial intelligence (AI). While most take the form of policies, States and international organizations tend to call them principles to reflect that they are a set of guiding criteria. Often, they are referred to as principles on responsible AI (RAI),

⁹. Aghion P., Bouverot A. (2024). IA : notre ambition pour la France, *Commission de l'Intelligence artificielle*. PDF: https://www.economie.gouv.fr/files/files/directions_services/cge/commission-IA.pdf

¹⁰. UN (2024). "Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development", General Assembly of 11 March 2024, A/78/L.49. PDF: <https://documents.un.org/doc/undoc/ltd/n24/065/92/pdf/n2406592.pdf?token=o9c5RBjCz02TeWB9qS&fe=true>

¹¹. OECD (2024). "Recommendation of the Council on Artificial Intelligence", Revised document of 3 May 2024, OECD/LEGAL/0449. PDF: <https://legalinstruments.oecd.org/api/print?ids=648&lang=en>

¹². Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law [Vilnius, 5.IX.2024] - <https://rm.coe.int/1680afae3c>

¹³. Among the prominent ISO standards relevant to AI are ISO/IEC 42001 and ISO/IEC 27001.

a term that originates from ethical guidance on the civilian development and use of AI. At the outset, the various principles on military and defense-related RAI are relatively homogeneous. States and international organizations RAI principles typically contain elements comparable to what NATO calls “responsibility and accountability,” “explainability and traceability,” “reliability,” “governability,” and “bias mitigation.” The principles and their elements may be called differently but tend to consist of similar substance. Some States have additional criteria. Switzerland, for instance, includes “agility” as a distinct principle. Interestingly, NATO includes the unique principle of “lawfulness” in addition to the typical principles. This principle states: “AI applications will be developed and used in accordance with national and international law, including international humanitarian law and human rights law, as applicable.”

Coordinating policy and regulations

Governance aims to align and coordinate policies, standards, and norms to ensure a coherent international approach to common problems. This role includes various functions: directly regulating technology use and requiring countries to follow certain rules, assisting countries in policy implementation, focusing on harmonizing and coordinating policies, certifying industries or regions to meet certain standards, and sometimes monitoring and enforcing compliance with these norms. Examples of institutions endorsing this role are the World Trade Organization or the International Monetary Fund.

In the case of AI, the proliferation of national regulations calls for greater coordination between countries, as embodied in the AI Act. Furthermore, the widespread use of AI across all industries demands careful attention to the interoperability of regulations across sectors. Adherence to AI principles also necessitates verification to ensure these principles are respected, requiring dedicated resources in international institutions for monitoring.

Enforcing standards and restrictions

Governance can also have to deal with the enforcement of rules and standards, the imposition of a threshold above which usage is prohibited, and straightforward bans. For example, institutions can be in charge of ensuring that standards developed by the ITU or ISO are respected. Moratoria, non-proliferation regimes, export control lists, monitoring and verification mechanisms, licensing regimes, and tracking key resources are other mechanisms that fall in this category.

Common examples are in the nuclear and military domains, with the International Atomic Energy Association and the Missile Technology Control Regime. In the case of AI, the use of the technology in the military sector with autonomous weapons, or in

surveillance with facial recognition, can justify the need for strong enforcement and restrictions.

It should be noted that incentives for compliance are key to successful enforcement. Codes of practice (for example, under the AI Act) or more generally codes of conducts may be plausible solutions toward international governance. The challenge, however, is that actors take them seriously. Previous experience in the field of content regulation has shown that this was not the case (Hernández-Echevarría, 2024).¹⁴ Under this light, imposing sanctions may become a necessity. Labels could be another way, the benefits that a company would gain from labeling its product creating a plausible incentive.

Three additional functions: emergency response, joint research and the distribution of benefits

Three additional functions are worth mentioning. Some issues may necessitate close surveillance of its effects on social stability and, in case of imminent emergency, the ability to react at a global scale. The World Health Organisation plays this role in the health sector. Emergency response may be necessary with AI in order to address existential threats.

Governance can also imply the development of joint research programmes to make significant progress in a given sector. The European Organization for Nuclear Research (CERN) is one such example. Given the costs associated with increasing computing power demands in the field, research in AI may benefit from joint research projects that mutualise resources. And indeed, a growing proposal among some academics and policymakers^{15 16 17} suggests forming an international coalition for AI research, modeled after the collaborative and large-scale nature of CERN.

Finally, governance can also seek to distribute the benefits induced by a given technology. With AI, it may be needed to ensure that AI is made accessible in all parts of the world, for example for use cases in the health industry.

¹⁴. Hernández-Echevarría C. (2024). Major Tech Platforms Fail to Deliver on EU Fact-Checking Commitments, Risking DSA Compliance, *Tech Policy Press*, available at: <https://www.techpolicy.press/major-tech-platforms-fail-to-deliver-on-eu-factchecking-commitments-risking-dsa-compliance/>

¹⁵. “Securing Our Digital Future: A CERN for Open Source large-scale AI Research and its Safety” (2023). Online petition submitted via openPetition on 1 June 2023, <https://www.openpetition.eu/petition/online/securing-our-digital-future-a-cern-for-open-source-large-scale-ai-research-and-its-safety#petition-main>;

¹⁶. Scholl G. (2022). “We need a CERN for AI in Europe”, *Humboldt Kosmos magazine*, Alexander von Humboldt Stiftung, interview with Professor Holger Hoos, 1 August 2022. Online: <https://www.humboldt-foundation.de/en/explore/magazine-humboldt-kosmos/by-courtesy-of-how-artificial-intelligence-is-changing-our-lives/we-need-a-cern-for-ai-in-europe>

¹⁷. Kaspersen A. (2021), “Time for an Honest Scientific Discourse on AI & Deep Learning, with Gary Marcus”, Carnegie Council for Ethics in International Affairs, 3 November 2021, <https://www.carnegiecouncil.org/media/series/aiei/20211103-honest-scientific-discourse-ai-deep-learning-gary-marcus>.

The governance of AI: working with old institutions or building new ones?

In their December 2023 interim report, "Governing AI for Humanity"¹⁸, the United Nations offered a categorisation of seven functions of AI governance. These functions are in essence the same as the ones highlighted above. Some are identical, such as building scientific consensus, others are repackaged, for example the design and the enforcement of norms which are grouped together under one function in the UN categorisation.

The UN report ranks these functions according to their institutional “hardness” (see figure 1). Some of these functions can be conducted more or less formally. For example, scientists can work on building scientific consensus independently of States and without a clear mandate to do so. Others, like enforcing standards and building norms, require “harder” institutional support.

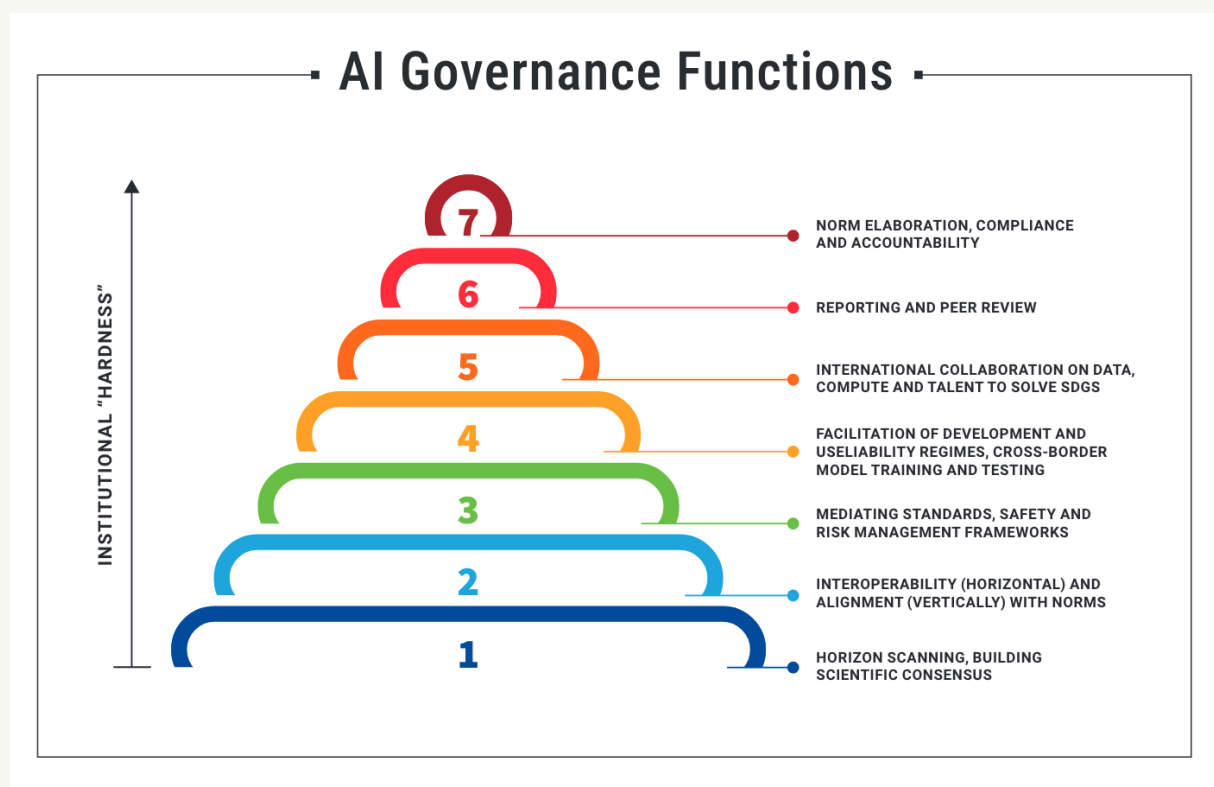


Figure 2: AI governance functions distributed by institutional “hardness”
(Source: United Nations, *Interim Report: Governing AI for Humanity*, 2023, p.16)

¹⁸. United Nations (2023). “Interim Report: Governing AI for Humanity”. Online: <https://www.un.org/en/ai-advisory-body>

Whether AI requires specific institutions remains an open question in public debates and the scientific literature.¹⁹

On the one hand, new institutions are not necessarily required for global governance. For example, the extraterritoriality of existing regulations could lead to an international governance regime without the need for a new international organization in charge of AI governance. The EU is aiming for the AI Act to have the same “Brussels effect” (Bradford, 2019)²⁰ as the GDPR – in other words, to become a standard if other states adopt it or if private companies implement it internationally to avoid compliance costs with multiple regimes. However, it is believed the EU approach toward regulating AI may not have the same global reach (Engler, 2020)²¹. Even if the AI Act is stringent enough, the fact that major powers in the field are also pursuing AI sovereignty²² could affect the successfulness of modeling other jurisdictions' rules on the EU approach (*de jure* Brussels effect).²³ In any case, several AI legislations could emerge without there being an international body dedicated to AI. Additionally, the governance functions could be undertaken by existing organizations. For example, the OECD and the UN have already performed important functions for AI governance by setting norms through their ethical frameworks.

In this scenario without new institutions, coordinating efforts between institutions and states becomes a priority. States, like companies, compete with one another. The perceived competitive advantage of AI has led states to enact policies to strengthen their international position. Without a sovereign authority to dictate rules and impose sanctions, states may therefore attempt to game the system for their benefit. Policies supporting national competitiveness are not necessarily detrimental for cooperation. Cooperation between existing institutions can however be dysfunctional, with institutions potentially imposing conflicting standards rather than working toward a harmonious framework.

On the other hand, some argue that AI does need new institutions (Roberts *et al.*, 2024), for example modeled on the functioning of existing organizations. Existing institutions alone may not cover all the functions required for AI governance. For instance, it is unclear which current institution could create an emergency mechanism in response to an imminent threat posed by an AI system. For the time being, emergency mechanism

¹⁹.For proponents of the view that AI governance needs specific institutions, see Hausenloy and Denis (2023) and Maas and Villalobos (2023), for those who argue for the coordination between existing institutions instead, see Roberts *et al.* (2024)

²⁰. Bradford A. (2019). *The Brussels Effect: How the European Union Rules the World*, Oxford University Press.

²¹. Engler A. (2022). “The EU AI Act will have global impact, but a limited Brussels Effect, Brookings, June 8, 2022. Online:

<https://www.brookings.edu/articles/the-eu-ai-act-will-have-global-impact-but-a-limited-brussels-effect/>

²². Larsen B. (2022). “The geopolitics of AI and the rise of digital sovereignty”, Brookings, December 8, 2022. Online: <https://www.brookings.edu/articles/the-geopolitics-of-ai-and-the-rise-of-digital-sovereignty/>

²³. Karathanasis T. (2024). Defining AI Systems in the EU and Beyond: An Assessment of AI Act’s Norms Global Outreach, *Under Review*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4834179

strategies are discussed through bilateral forums of cooperation (e.g., U.S-EU Trade and Technology Council)²⁴.

Building new institutions does not eliminate coordination challenges. It is unlikely that a single new institution will encompass all seven functions of AI governance. It may be necessary to delegate norm-building to existing institutions while focusing on enforcing regulations and coordinating policies. The new institution would need to engage with a broad range of stakeholders or risk being inefficient.

At a national level, or transnational level in the case of the European Union, even if a new institution in charge of AI is created, it will need to work with existing regulators. In accordance with the EU AI Act, an AI Office has been established in June 2024 within the European Commission to contribute to the supervision of AI systems, along with national competent authorities, and bodies, offices and agencies of the Union. The AI office and national competent authorities can endorse more or less governance functions. For example, they may build norms but leave enforcement to existing regulators. AI Governance at European level will also be supplemented by a European AI Board, which will be tasked to provide a platform for cooperation and exchange among market surveillance authorities and notifying authorities about issues related to market surveillance and notified bodies respectively.

Challenges with the governance of AI

In order to address the range of issues that AI raises, governance should address as many of the seven functions highlighted above: build scientific consensus, build norms, coordinate policy, enforce standards, foster joint research, anticipate emergency response and distribute resources. It can do so by creating new institutions or using existing ones. Whether one chooses to improve interactions between existing institutions or to create new ones, issues of coordination abound. In this section, we highlight some of the challenges that were addressed by participants during the first AI Dialogue.

Defining AI

Defining artificial intelligence (AI) for policy and governance purposes presents a challenge due to the rapidly evolving nature of the technology. Governance structures must define AI clearly to make enforcement possible; yet the definition must remain flexible to accommodate technology evolutions, ensuring that regulations can adapt alongside technological advancements. For example, in its recommendation on the

²⁴. White House (2024). "U.S-EU Joint Statement of the Trade and Technology Council". 5 April 2024. Online: <https://www.whitehouse.gov/briefing-room/statements-releases/2024/04/05/u-s-eu-joint-statement-of-the-trade-and-technology-council-3/>

ethics of artificial intelligence, UNESCO chooses not to provide one single definition of an AI system, but several elements that help seize its impact on human rights. Recently, the emergence of large language models (LLMs) has incited European institutions to include a definition of general-purpose AI in the AI Act (see Table 1).

Historically, definitions of AI have changed significantly, reflecting the technological progress from the 1970s to the present day. Initially, AI was often described simply as software. Over time, this definition has expanded to encompass entire systems, illustrating the shift from a narrow to a broader understanding of what constitutes AI. This evolution is evident in discussions surrounding legislative efforts such as the AI Act, where definitions over time attempted to capture the complexities of modern AI systems, moving from a focus on software to one on machine-based systems. Internationally, some differences remain. In their approach to regulating AI, Brazil, Canada and Chile define an AI system as a computer system, a technological system or software, respectively.

Given the difficulties in defining AI as a technology - largely due to divergent national policies on AI sovereignty -, and in order to move on with regulation, one participant in our group suggested an alternative approach: focusing on the effects of AI rather than its intrinsic nature. This perspective emphasizes the importance of regulating the impacts and applications of AI. For instance, rather than attempting to define the technicalities of AI technologies that create deepfakes, regulations could focus on the misuse of AI to spread deepfakes with the intent to influence elections. Another option could be to regulate AI based on the seriousness of an incident implicating an AI. The OECD released a report recently on defining AI incidents.²⁵ In developing a common AI incident reporting framework, the report distinguishes between the actual and the potential harm of AI (incident vs hazard). This approach prioritizes the consequences of AI applications, addressing the real-world impacts that require governance.

Despite their inherent incompleteness, definitions of AI still hold significant value. They enable users to exercise their rights, ensuring that the protection of human rights remains a priority. Even if a definition does not capture every aspect of AI, it can still offer a foundation for understanding and regulating the technology's impacts on society.

AI definitions are most useful when they are adopted by a variety of actors. Table 1 shows how the OECD definition of AI systems has been influential, being adopted by the Council of Europe or the European Union's AI Act. This widespread adoption illustrates the role the OECD plays in accomplishing one of the functions of AI governance, namely building political consensus and norms.

²⁵. OECD (2024). "Defining AI Incidents And Related Terms, *OECD Artificial Intelligence Papers*, n°16. Online: https://www.oecd-ilibrary.org/fr/science-and-technology/defining-ai-incidents-and-related-terms_d1a8d965-en

	Definition of AI system
OECD (2019 and 2023) ²⁶	<p>2019 version: “An AI system is a machine-based system that can, for a given set of human defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.”</p> <p>2023 version: “An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.”</p>
NATO Science and Technology Organization (2020, p. 14) ²⁷	Artificial Intelligence (AI) refers to the ability of machines to perform tasks that normally require human intelligence – for example, recognising patterns, learning from experience, drawing conclusions, making predictions, or taking action –whether digitally or as the smart software behind autonomous physical systems”
UNESCO ²⁸ (2021, p. 10)	<p>UNESCO focuses on the ethical dimension of AI and stresses three aspects of AI systems that must be accounted for to understand their impact: they are information-processing technologies, they are seized through their life cycle, they raise new ethical issues.</p> <p>It approaches AI systems as “systems which have the capacity to process data and information in a way that resembles intelligent behaviour, and typically includes aspects of reasoning, learning, perception, prediction, planning or control.”</p>
AI Act (2024, article 3.1 and 3.63)	<p>AI system: “a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments” (article 3.1).</p> <p>General-purpose AI model: “an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market” (article 3.63)</p>
Council of	AI system: “a machine-based system that for explicit or implicit objectives,

²⁶. For a discussion on the evolution of the OECD definition, see “Updates to the OECD’s definition of an AI system explained”: <https://oecd.ai/en/wonk/ai-system-definition-update>

²⁷. NATO (2020), “Science & Technology Trends 2020-2040 Exploring the S&T Edge”. PDF: https://www.nato.int/nato_static_fl2014/assets/pdf/2020/4/pdf/190422-ST_Tech_Trends_Report_2020-2040.pdf

²⁸. UNESCO (2021). “Recommendation on the Ethics of Artificial Intelligence”. Online: <https://unesdoc.unesco.org/ark:/48223/pf0000381137>

<p>Europe²⁹ (2024, article 2)</p>	<p>infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that may influence physical or virtual environments. Different artificial intelligence systems vary in their levels of autonomy and adaptiveness after deployment.”</p>
---	---

Table 1: Definition of AI systems in selected regulatory frameworks

Building consensus

Ensuring interoperability between several national AI texts is key. In addition, each legislative text on AI must also interact with other sectoral legal frameworks. For example, in the EU, the AI Act must be compatible with texts such as the General Data Protection Regulation (GDPR) and the Directive on Copyright in the Digital Single Market. Achieving interoperability between AI regulations and between AI regulations and legislation in adjacent sectors requires consensus on definitions and methodologies. At the international level, as there are various global initiatives, ensuring interoperability, or as a minimum coordination between these initiatives is key as well.

Reaching a consensus is a significant challenge due to the diverse range of stakeholders involved. These stakeholders have more or less technical backgrounds. They come from various sectors, such as health, education, supply chain management, and work in different types of organizations, such as private companies, public institutions, and non-profit associations. They also represent value systems anchored in different regions. The more participants involved in discussions, the harder it is to agree on a precise set of goals. Even within smaller organizations, such as OpenAI or Twitter, aligning values and principles can be difficult. Internal conflicts, such as the firing of safety teams, highlight the challenges of establishing a cohesive set of values on a smaller scale. Inclusivity, while essential for democratic governance, complicates the decision-making process. It should be stressed that during the process for finalizing the Convention on AI at the Council of Europe, the treaty faced a diplomatic blockade by non-voting observers (the US, Canada, Japan and the UK), who wanted to exclude private actors from its scope, thus limiting the initiative’s impact (Volpicelli , 2024) ³⁰.

Human rights often provide a common framework to harmonize values on an international level. This shared value system can facilitate cooperation among different regions. However, the increasing polarization of global politics makes harmonization difficult. Different regions are pushing their own visions for AI governance. For example, during the AI Dialogue, one participant noted that attempts to set rules for off-limit targets in warfare policy failed due to disagreements among Russia, China, and the

²⁹. Council of Europe (2024). “Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law”, Council of Europe Treaty Series - No. [225]. PDF: <https://rm.coe.int/1680afae3c>

³⁰. Volpicelli G. (2024). International AI rights treaty hangs by a thread, *Politico*, March 11, 2024. Online: <https://www.politico.eu/article/council-europe-make-mockery-international-ai-rights-treaty>

United States. Additionally, states may publicly adhere to a set of norms while pursuing interests that contradict these norms. This makes the enforcement and accountability of both private companies and states a priority.

Tensions between universal values can also lead to inefficiencies. For example, while increasing transparency is generally seen as positive, it can lead to unintended consequences such as fostering cybersecurity breaches or unveiling commercial secrets. Similarly, promoting open-source AI can result in the technology being misused for harmful purposes (dual-use risk). These examples underscore the delicate balance that must be struck between promoting ethical AI practices and mitigating potential risks. In its Recommendation on the Ethics of Artificial Intelligence, UNESCO states, “In any given situation, a contextual assessment will be necessary to manage potential tensions, taking into account the principle of proportionality and in compliance with human rights and fundamental freedoms” (UNESCO, 2021, 18).

Given these complexities, some question the very possibility of effective AI governance. The difficulty of finding agreement on definitions and values suggests that a one-size-fits-all approach may not be feasible. Instead, adaptive and context-specific strategies may be required to navigate the evolving landscape of AI governance.

Making AI systems accountable

Transparency, fairness, and accountability are fundamental principles in most AI ethical frameworks. These principles aim to ensure that AI systems are developed and used in ways that are understandable, just, and responsible. Among these principles, transparency is often seen as a prerequisite for achieving the others. Without transparency, ensuring fairness and holding entities accountable becomes challenging.

Achieving transparency in AI, however, is complex. Unlike tangible products, the functioning of AI systems is largely invisible. For instance, identifying chaos parameters and the potential for AI to generate unpredictable outputs, or "hallucinations," requires new methods. Traditional safety testing methodologies often fall short in capturing these unique aspects of AI behavior.

Beyond technical difficulties in explaining some AI models, public authorities often have limited information to verify compliance with norms. Typically, public actors must rely on the goodwill of private companies to share information. This situation is similar to platform regulation concerning the risks social media pose to society, where European regulation has sought to create mechanisms for transparency and accountability (see next section).

This raises critical questions about accountability and auditing. While transparency is crucial, it requires methods and organizations to verify the accuracy of shared

information. Governance should therefore establish indicators on information that private actors can publicly disclose, along with methods to measure these indicators. For example, testing datasets for potential bias requires agreement on criteria (such as gender, sexual orientation, and ethnicity) and thresholds above which a dataset is considered biased.

A plurality of standards can provide a more comprehensive approach to AI governance. The European AI Act proposes a version of these standards. On the one hand, it can foster consensus and accelerate the testing of AI systems. On the other hand, it might stifle the development of competing standards that could offer better solutions. Balancing the urgency of AI regulation with the need for multiple perspectives is key to creating the best governance approach.

Finally, questions arise about who should conduct these audits. The independence of auditors is crucial. However, large auditing firms often provide consulting services to the companies they audit, raising concerns about their neutrality. While preventing them from working with clients might seem reasonable, it could lead to a shortage of qualified auditors.

Ensuring the accountability of AI actors therefore involves navigating several complexities. Detailed guidelines and standards for verifying information are necessary for ensuring transparency³¹. Effective auditing is central to this effort, yet ensuring the independence of auditors remains a challenge.

Existing models for the governance of AI

AI is not the first technology that raises questions about its international governance. Communication infrastructures, atomic energy and aviation are amongst the examples of innovation that have led to the creation of new governance frameworks. In what follows, we draw comparison between AI and governance approaches in other sectors³². The purpose is to reflect on the process through which governance structures were built in other fields. Product safety and content/Internet were chosen for the similarities they share with AI. As with product safety, AI may require that it be tested before entering the market. As with content, AI companies must make complex systems accountable to public agencies.

³¹. On April 16, 2024, the French Minister of Culture submitted two reports to the Conseil Supérieur de la Propriété Littéraire et Artistique (CSPLA). The first aims to assess the scope of the transparency obligation provided for in Article 53 of the draft European regulation on AI and to draw up a list of information that AI providers must make public, according to the cultural sectors concerned, in order to enable authors and holders of related rights to exercise their rights.

³². For a detailed presentation of governance structures in other sectors, see Maas and Villalobos (2023) and Microsoft, *Global Governance : Goals and Lessons for AI* (2024). Online: <https://blogs.microsoft.com/on-the-issues/2024/05/17/global-governance-goals-and-lessons-for-ai/>

Standardize the technical layers as a first step to explainability and regulate

One model for the governance of AI could be to learn from the past, in particular from the regulation of the web (which led primarily to the standardization of the technical layers, e.g. via ICANN, and ITU, ISO) and to standardize the technical layers of AI models. This could be done by fostering and endorsing existing organizations and their work (e.g. ITU, ISO).

This would force AI models not only to be interoperable worldwide, but also to be understandable, auditable and ready for further regulations and definitions. Consequently, this could be seen as a first necessary step, that does not exclude further models and regulations.

Product safety: testing before selling

Artificial intelligence (AI) shares several similarities with traditional products. AI systems, much like traditional consumer goods, are commercialized and distributed in various markets. The sale and deployment of these technologies necessitate a careful evaluation of safety risks to protect end users. This commonality may justify the need for robust regulatory frameworks to mitigate any potential harm.

There are, however, key differences with physical products. The first is the way AI is tested (see section above). The second is that general purpose AI systems are designed to perform multiple tasks, which means they do not adhere to a constant level of risk. This variability complicates risk assessment and management, necessitating a more dynamic and adaptable approach to regulation. In other words, **there is not one way to test an AI's safety. An AI can pass a safety test when used in one way (e.g. facial recognition to unlock a phone), and fail when used in another (e.g. facial recognition to monitor minority groups).**

The question product safety poses is how to minimize risks before a technology is sold on a market. One significant drawback of this approach is its potential impact on innovation. Product safety regulations are designed to protect consumers, but they often come at a cost. During the AI Dialogue, one speaker remarked that, in the aeronautic industry, stringent safety regulations have slowed down the ability of actors to innovate. In this regard, the European AI Act proposes a risk-based approach to enforce stricter rules for and testing of high-risk AI systems before they are put on the market.

Avoiding excessive burdens for micro-enterprises in complying with such standards should also be considered. The AI Act provides the establishment of national regulatory

sandboxes. These controlled environments intend to foster innovation and facilitate the development, training, testing and validation of innovative AI systems, while reducing the burdens of compliance with the provisions of the AI Act, since successful participants will be deemed to be compliant without the need for further audits.

Content: building accountability mechanisms

Regulating actors developing AI systems and social media companies share several commonalities. Both AI companies and social media platforms raise questions about their impact on fundamental rights, including privacy, security, freedom of expression, and access to information. One of the aims of the Digital Services Act, the European legislation regulating social media, is to remain evolutive in order to accompany future technological innovations. Its focus on systemic risks helps identify categories of risks while not naming specific technologies that could quickly become outdated. Additionally, both developers of general-purpose AI and social media companies operate on a scale that transcends national boundaries. In both cases, the market is dominated by large actors, creating a dependency on the transparency provided by major companies. Consequently, regulators face the challenge of creating mechanisms to ensure transparency and, ultimately, accountability.

There are also differences. AI systems can shape a wide range of sectors, including healthcare, finance, and critical infrastructure, leading to far-reaching consequences for society. Additionally, unlike content, which is not sold on the European market, AI products are actively marketed and sold within Europe.

The regulation of social media companies offers important insights for the governance of AI. Whilst product safety focuses on a product, content regulation focuses on the company developing complex information sharing systems. Policymakers in this space aimed to create a more transparent and accountable framework for regulating digital platforms. Their objective has been to help both regulators and private companies understand the potential harms caused by social media.

To address rapid technological evolutions, policymakers chose to focus on a co-regulatory approach, where private actors must collaborate with regulators in identifying systemic risks and implementing methods to mitigate them. Lawmakers recognized the insufficiencies of previous codes of practice, in particular the inability of regulators to verify the veracity of the companies' publications. They also recognized the asymmetry of knowledge between regulatory agencies and platforms, understanding the need to make this knowledge public in the interest of transparency. Thus, the DSA mandates transparency on specific indicators, initially allowing for diverse

methodologies in developing these indicators. Ultimately, these methodologies should converge, leading to more uniform and comparable standards.

In the case of AI, particularly general-purpose AI, similar questions arise. An international governance framework for AI could seek to break the knowledge asymmetry between companies and regulators to ensure that fundamental rights are effectively protected. This could be achieved by building on existing codes of practice to identify indicators and foster the development of shared methods to measure them, thus making information comparable and contributing to accountability.

Conclusion

In conclusion, the rapidly evolving field of AI raises important questions about governance. Effective AI governance must perform various functions, from building political consensus to enforcing standards and coordinating policy. However, the competing interests of national states make it difficult to reach a consensus on the best approach. It remains to be seen which existing institutions will assume some of these roles and whether new institutions will be needed for others.

As we move forward, the next Dialogue will explore which institutions are best suited to take on these responsibilities, aiming to establish a cohesive and effective governance framework for AI.

AUTHORS

RAPPORTEURS

Théophile LENOIR, Main Rapporteur, Renaissance Numérique
Julian MAREL, Rapporteur/Project Officer, Renaissance Numérique

PUBLICATION DIRECTOR

Jean-François LUCAS, General Manager, Renaissance Numérique

CONTRIBUTORS

Yaniv BENHAMOU, University of Geneva (UNIGE), Associate Professor (Digital Law)
Sarah CLÉDY, Government Affairs and Public Policy Senior Analyst, Google
Theodoros KARATHANASIS, Grenoble Alpes University, Research Fellow on Cybersecurity

THE EXPERTS WHO PARTICIPATED IN THE FIRST DAY OF THE *AI DIALOGUES*

Alexander BARCLAY, State of Geneva, Deputy for Digital Policy.

Arthur BARICHARD, Ministry of Foreign Affairs (France), Deputy to the Ambassador for Digital Affairs.

Yaniv BENHAMOU, University of Geneva (UNIGE), Associate Professor (Digital Law).

Sarah BÉRUBÉ, Organisation for Economic Co-operation and Development (OECD), Deputy Director & Head of Digital Economy Policy Division.

Roland BOUFFANAIS, University of Geneva (UNIGE), Associate Professor (Global Studies & Computer Science).

Jean-Marie BOUTIN, Google France, Director of Institutional Relations.

Agustina CALLEGARI, World Economic Forum (WEF), Lead, Digital safety initiative.

Carl GAHNBERG, Internet Society (ISOC), Director of Policy Development & Research.

Samira GAZZANE, World Economic Forum (WEF), Policy Lead, Artificial Intelligence and Machine Learning.

Amin HASBINI, Kaspersky France and North, West & Central Africa, Head of Global Research and Analysis Team (META).

Theodoros KARATHANASIS, Grenoble Alpes University, Research Fellow on Cybersecurity and AI Regulation.

Michael KENDE, Datasphere Initiative / Analysys Mason, Chair of The Board of Trustees / Senior Advisor.

Nour KHAYAT, Renaissance Numérique, Rapporteur / Project Officer.

Théophile LENOIR, Renaissance Numérique, Rapporteur.

Jean-François LUCAS, Renaissance Numérique, General Manager.

Julian MAREL, Renaissance Numérique, Rapporteur/Project Officer.

Anthony MASURE, HEAD – Genève, Dean of Research.

Jean-Marc RICKLI, Geneva Center for Security Policy (GCSP), Head of global and emerging risks.

Laurent SCIBOZ, University of Applied Sciences and Arts Western Switzerland (HES-SO), Head of applied research Institute.

Nicolas VANBREMEERSCH, Renaissance Numérique, Chair.

Gladys YIADOM, Kaspersky France and North, West & Central Africa, Public Affairs Manager.

About

[Renaissance Numérique](#) is an independent think tank dedicated to the digital transformation of our society. Its purpose is to shine a light on the changes brought about by this transformation, and to provide everyone with the tools to master it.

Renaissance Numérique is a not-for-profit association governed by the French law of 1901. It is fully independent, i.e. not affiliated to any party, company or structure. The digital transformation is profoundly impacting our social, economic and political interactions and structures. To grasp and understand its complexity, which is itself ambiguous and changing, Renaissance Numérique brings together [members](#) from a wide range of backgrounds (political, economic, legal, communications, technical, sociological, etc.) and structures (independent experts, consultancies, law firms, non-governmental organizations, universities, institutions, businesses, etc.).

This diversity of actors and points of view makes Renaissance Numérique a place for debate, a space to enjoy a positive confrontation of ideas, which is unique in the landscape of think tanks and digital players in France and Europe.

www.renaissancenumerique.org



**Renaissance
Numérique**