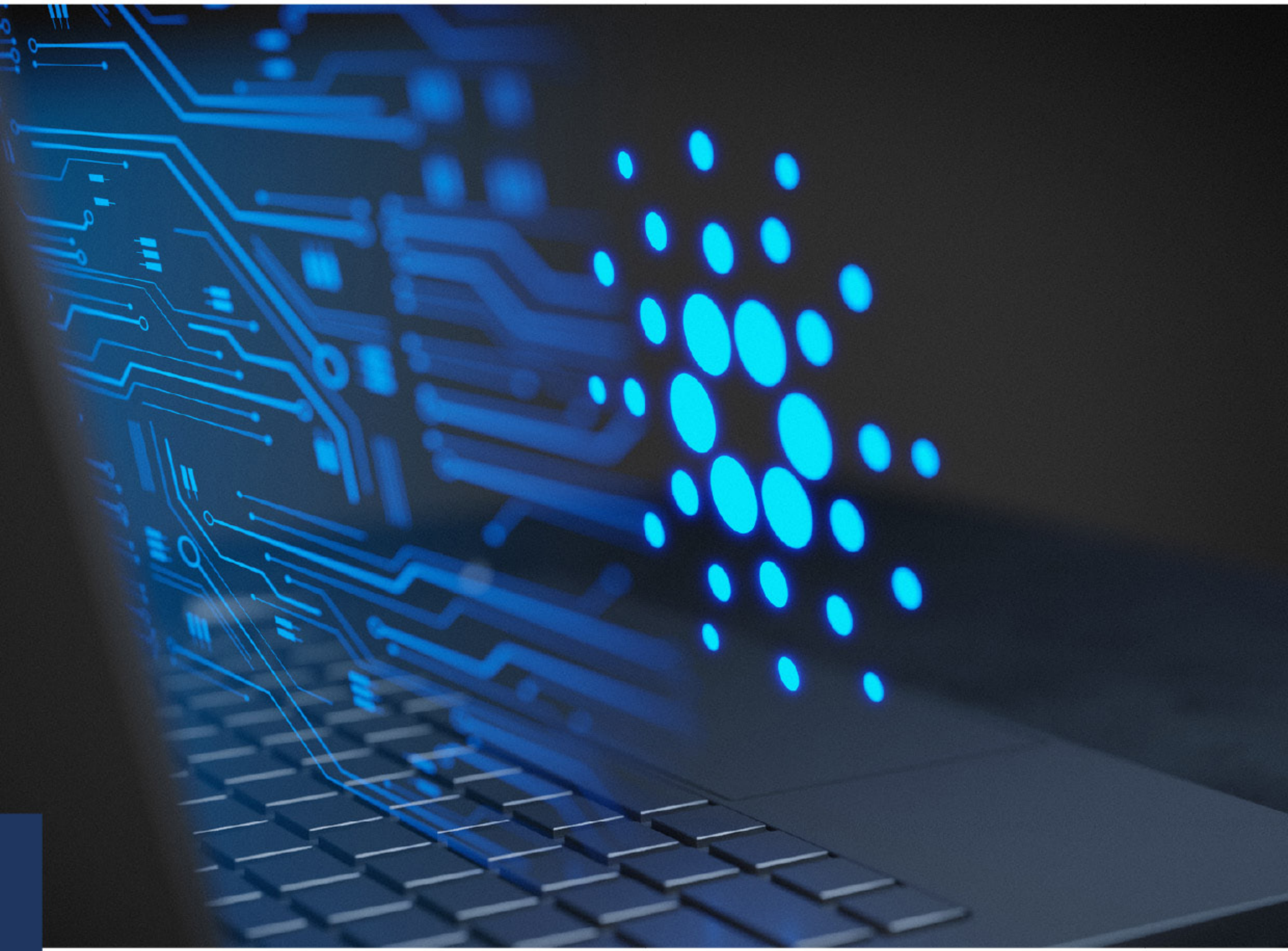
CEBRI

ITS Instituto de Tecnologia & Sociedade do Rio



An *AI Safety and Opportunities Institute for Brazil:*

A Design of Features, Functions, and Potential Challenges

CEBRI & ITS Rio, March/2025

CEBRI

www.cebri.org

TO THINK
TO DIALOG
TO DISSEMINATE
TO INFLUENCE

#2 Think Tank in South and Central America

University of Pennsylvania's Think Tanks and Civil Societies Program 2020 Global Go To Think Tank Index Report

The Brazilian Center for International Relations (CEBRI) is an independent, non-partisan and multidisciplinary entity guided by excellence, ethics and transparency and that focuses on the creation and dissemination of high-quality content about the global scenario and Brazil's international role. CEBRI maintains an open debate with government and non-government entities to influence the construction of an international agenda for Brazil and to support the formulation of forward-looking public policies that can make an impact.

In the course of its twenty-two year history, CEBRI has organized more than 500 events and produced more than 200 publications. CEBRI cooperates with more than 100 high-level entities on every continent. CEBRI stands out for its intellectual capital, for its ability to bring together the multiple views of renowned specialists and for the significance of its Board of Trustees.

CEBRI uses its international connections to identify and analyze major international issues and to foster the interaction between the production of knowledge and political action. CEBRI operates as counterpart to strategic global institutions such as the Council on Foreign Relations in the US and Chatham House in the UK, as well as several other Councils on International Relations on the global arena.

CEBRI's international significance is also attested by the University of Pennsylvania's Global Go to Think Tanks survey, which found CEBRI to be considered one of the most important think tanks in the world.

EXECUTIVE SUMMARY

This paper discusses the characteristics, functions, and challenges for the creation of an AI Institute for Safety and Opportunities in Brazil, or AISO (*AI Safety and Opportunities Institute*).

The first wave of AI governance institutions is known as AISIs, or AI & Safety Institutes. This period refers to the creation of eight AI institutions defined between the UK AI Summit (November 2023) and the Paris AI Summit (February 2025).

This article analyzes in detail the lessons learned and the similarities between the existing AISIs, in addition to proposing a discussion on how Brazil can participate in and influence the international process.

We advocate the idea that the second wave of AI Institutes should focus on balancing technical concerns about risk with the development of scientific and methodological tools that promote opportunities. For this reason, we suggest adopting the terminology *AI Safety and Opportunities Institute (AISO)*

Considering the particularities of Brazil, this reformulation points to an alternative path.

If safety (represented by the concepts of safety and security) was the main driving force during the UK AI Summit, the direction shifted to action at the AI Action Summit in Paris. Part of this shift reflects growing concern that the existential threats of AI overestimate the risks and underestimate the technology's development.

It is worth noting that the United Kingdom's AISI has already replaced the word "Safety" and that in the European Union and the United States there is growing concern about how to strengthen scientific and technical resources in order to expand economic opportunities. This new understanding is not exactly a novelty, but a logical unfolding of the evolution of existing institutes.

We argue that a Brazilian AISO needs to balance opportunities for AI adoption with concerns about safety, considering the focus on "safety" and "opportunity" simultaneously. It is worth noting that the proposed AISO would have an advisory and not a regulatory function.

Advancing comments received during the meetings with experts, two points stand out: the need for participation from different sectors, and the obligation to have good alignment with international organizations and agendas.

The document is divided into two sections: the first reviews the aspects of the first wave of AISIs, and the second presents the characteristics and design that an AISO can have.

INDEX

Instituto
de Tecnologia
& Sociedade
do Rio

EXECUTIVE SUMMARY 03

INTRODUCTION 06

Section 1. The First Wave of AI Safety Institutes 09

1.1 The agendas of the summits 10

1.1.1 UK AI Security Summit 10

1.1.2 South Korea AI Security Summit 11

1.1.3 AI Action Summit in Paris 12

1.2 Reference institutions 13

1.2.1 International Panel on Climate Change (IPCC) 13

1.2.2 International Atomic Energy Agency (IAEA) 14

1.2.3 European Centre for Algorithmic Transparency (ECAT) 15

1.2.4 National Institute of Metrology, Quality and Technology (INMETRO) 16

1.3 Existing AISIs 16

1.3.1 United Kingdom 16

1.3.2 United States of America 18

1.3.3 Singapore 20

1.3.4 Canada 21

1.3.5 Japan 22

1.3.6 France 23

1.3.7 European Union 24

1.3.8 South Korea 25

1.3.9 Brazil 26

1.4 Comparison tables 28

1.4.1 Comparative governance structure 28

1.4.2 Strategies to foster cooperation 29

1.4.3 Organizational Overview of AISI Network Members 30

Section 2. Principles and functions of the AISO 31

2.1 Features 32

2.1.1. Independent governance 33

2.1.2. Technical and evidence-based leadership 34

2.1.3 Advisory body 35

2.1.4 Shared methodologies 36

2.1.5 Promotion of international cooperation and alignment 37

2.1.6 Focus on the common good 38

2.2 Functions 39

2.2.1 Research 39

2.2.2 Targeted practices 39

2.2.3 Cooperation 40

Next steps 41

INTRODUCTION

The AI Safety Institutes (AISIs) are institutional bodies dedicated to the safe development of Artificial Intelligence. They contribute through scientific and technical analysis to the understanding and management of the risks inherent in AI technologies, while promoting safe innovation.

However, AISIs are a relatively new phenomenon, the result of three summits held in a sequence of approximately two years. The first was the UK AI Summit held in November 2023, followed by the Korea AI Summit in May 2024, and thirdly the Paris AI Summit which took place in February 2025.

Two years after the beginning of this first wave, what are the characteristics of this moment?

What can be known so far are common traits of the AISIs created. First, we know that all AISIs are government-run bodies, open to the participation of different stakeholder groups, but governed from within the state. Second, all of them have multi sectoral participation, with emphasis on the scientific and technical community. Third, they all focus primarily on AI safety. And finally, everyone cooperates with each other, especially after the formation of the AISI Network, launched after the Korea AI Summit.

In parallel, another characteristic of the first wave is the growth of the private sector's participation in interacting with AISIs, sometimes creating internal institutions with some compatible functions. Key incentives for the private sector include building reputations, engaging with or complying with existing regulations, and encouraging the development and responsible adoption of AI in their markets.

Having understood the dimension of the first wave of AISO, this report asks what should be the characteristics of the next movement, a possible second wave of AI & Safety Institutes.

With this in mind, this report proposes the creation of a Brazilian AI Safety & Opportunities (AISO).

In advance, the suggestion is that the new organization should not be confused with and not replace functions performed by existing organizations. The debate in Congress is moving towards proposing an organization with political powers; Inmetro already performs standardization functions and, through its Directorate of Scientific Metrology, conducts AI research focused on safety and reliability, having recently established its Artificial Intelligence Center at Inmetro (CIAI).; the ANPD already addresses AI issues from the perspective of personal data, among other cases.

That is, this is a design proposed to support current or future organizations that exist. It would be the design of an organization with an incentive to do what the others cannot do (alone or collectively) and that supports the other organizations to fulfill their main function.

Remaining open for debate, this report suggests the need for an AISO in Brazil, composed of the following pillars:

- The AISO needs to be **potentially independent,** that is, not controlled or managed exclusively by the government alone or by the private sector. For this, multisectoral coordination is essential, bringing together specialists and technicians from different sectors of society. This model would help avoid the capture of the institution by a specific actor and, more importantly, create an environment of trust for the sharing of information and the construction of scientific knowledge.

- AISO needs to be **expert-led and evidence-based,** so that it becomes capable of providing technical and scientific expertise to the AI ecosystem. By having technical experts in AI, AISO will be able to address in depth both the risks and opportunities associated with the accelerated development of this technology.

- AISO should function as an **advisory** body, not a regulatory one, offering insights into the risks, opportunities, and impacts of AI to governments and industry actors. AISO's transformative power will not come from a regulatory function, but from the technical and scientific position it will occupy and the influence it will have on different stakeholders.

- The AISO should promote the **reuse and adaptation** of existing **methodologies,** best practices and standards. This implies promoting the use and adaptation of existing methodologies, developing, adjusting, and testing tools used to assess risks and opportunities, as well as good practices that increase the interoperability, sustainability, and efficiency of AI systems.

- AISO should be **a "hub of hubs"**, promoting national and international connections. This entails helping to foster a network of institutes and bodies responsible for AI research and development and engaging with other AISIs around the world, such as the network already established after the Seoul Summit.

- The AISO should focus exclusively on the common good and benefit. This direction motivates governmental and non-governmental actors to work together.

To further this discussion, this report is divided into two main sections:

The first section analyzes the characteristics, functions, and challenges of the "first wave" of AISIs. This "first wave" refers to the set of AISIs that emerged from the United Kingdom and Korea summits, ending with the Paris summit. These institutions are already widely discussed in the existing literature and will serve as a basis for our analysis.

The second section proposes the characteristics and functions that will guide the implementation of the future AISO. In this section, we speculate on the desirable aspects for the creation of a truly independent institute, with multisectoral leadership and a specific focus on the Brazilian context, while remaining open to public consultations and discussions on our conclusions.

NOTES

This is an exploratory document, subject to updating, which aims to stimulate debate among stakeholders and contribute to Brazil's advancement in discussions on the subject, both nationally and internationally. It should be understood as a starting point for structuring and deepening conversations about the creation and role of AI Safety Institutes.

The report was prepared by ITS Rio, with the support of CEBRI. The collaboration between these two institutions aims to ensure that Brazil and the Global South respond creatively and appropriately to the opportunities provided by technology in the digital age, promoting a high-level and well-informed debate on international relations.

This report was partially funded by Microsoft.

SECTION 1. THE FIRST WAVE OF AI SAFETY INSTITUTES

The first wave of AISI (a term proposed by the Institute for AI Policy & Strategy, IAPS) comprises eight institutions created between November 2023 and January 2024. We deal below with the similarities between them, and the lessons learned throughout the process, based on their main characteristics, functions and challenges faced by these newly created institutions.

As the main **characteristics** of first-wave institutions, the IAPS highlights the fact that they are:

(i) administered by the government, but open to technical and scientific participation and to the contribution of other sectors;
(ii) focused on technical and scientific institutions;
(iii) restricted to a mandate focused on advanced AI systems;
(iv) devoid of regulatory powers; and
(v) internationally connected in a network of reciprocal relationships (a point added by our research).

Regarding shared **functions**, IAPS points to:

(i) research to promote fundamental safety methodologies in AI;
(ii) development of standards to influence market practices, inside and outside the government, through guidelines and protocols; and
(iii) cooperation, designed to reduce gaps between governments, industry and society at large.

Finally, with regard to the main **challenges** according to the IAPS, the first wave was criticized for:

(i) focusing excessively on specific sub-areas of AI safety, which makes it difficult to deal uniformly with macro trends such as competitiveness, fairness, and bias – in addition to reducing safety in general to existential threats, in ways that have actually diminished the attention given to other relevant aspects of AI-related public policy;
(ii) create redundancy with existing institutions, in particular with well-established standards development bodies; and
(iii) develop new forms of institutional collaboration with industry.

Another source of inspiration for understanding the first wave is research conducted by the Alan Turing Institute, which analyzes the potential roles that AISIs could play based on the work of two model institutions for the first wave: the International Atomic Energy Agency (IAEA) and the Intergovernmental Panel on Climate Change (IPCC).

The functions of these pre-existing institutions are aligned with the functions of the existing AISIs, namely: (i) to promote technical research and cooperation; (ii) develop safeguards and assessments; and (iii) support policymaking and governance.

These topics will be explored in more detail in the following sections.

Esses tópicos serão aprofundados nas seções a seguir.

1.1 THE AGENDAS OF THE SUMMITS

During the first wave of AISIs, three major events took place between 2023, 2024, and 2025: one in the United Kingdom, the other in Korea, and the last in France. Below, we look at the key factors that drove these events and the impact they had on the first wave of AISIs.

1.1.1 UK AI SAFETY SUMMIT

In November 2023, the UK's AI Safety Summit marked a turning point in the international dialogue on AI safety, by establishing the first AI & Safety Institute in the country. The initiative was created to serve as a model and inspire the creation of similar institutions around the world. The summit ended with the so-called Bletchley Declaration (also endorsed by Brazil), which highlighted the need for rigorous safety standards and international cooperation to manage the risks associated with AI.

Inspired by the so-called "Bletchley Effect", countries such as the United States, Japan, Singapore, and Canada have established their own AI safety institutes. These AISIs are designed as *hubs* to standardize and share safety assessments of AI technologies across borders, fostering a global AI governance network with an initial focus on governance.

1.1.2 AI SUMMIT KOREA (AI SEOUL SUMMIT)

In May 2024, the Korea AI Safety Summit continued the agreements of the UK Summit and prioritized strengthening international cooperation, as well as creating standardized frameworks for safety testing, promoting AI applications that support the SDGs (Sustainable Development Goals), and expanding the role of AI in sectors such as healthcare, education and economic development.

The results of the summit were consolidated in three main documents:

1. SEOUL DECLARATION

Endorsed by participating nations, it reaffirmed the commitment to promoting AI safety and highlighted the need for greater cooperation and sharing of best practices among countries.

2. MINISTERIAL DECLARATION

Focused on a detailed framework for international collaboration on AI governance. It included commitments to support the development of AISIs and the strengthening of research and standardization efforts.

3. SEOUL DECLARATION OF INTENT

Highlighted the importance of scientific and technical cooperation in AI safety, promoting joint research initiatives and standardized safety protocols.

In November 2024, in San Francisco, the International Network of AI *Safety Institutes was founded.* The Seoul summit was essential in laying the groundwork for continued international dialogue and cooperation on the challenges and complexities associated with AI development and governance, as well as a clear framework for the creation of new AISIs.

1.1.3 PARIS AI ACTION SUMMIT

The *AI Action Summit in Paris*, held in early 2025, marked a significant shift in global discussions on AI governance, moving beyond the risk-centric approach that defined the first wave of AISIs, towards a broader agenda that still highlights technical and scientific debates, but now connected to innovation, sustainability and opportunity.

Unlike previous summits in Bletchley Park (UK) and Seoul, which focused primarily on AI safety, the Paris summit emphasized evidence-based strategic leadership in the area of AI that more clearly encompasses aspects of competitiveness, technological development, and infrastructure investment.

French President Emmanuel Macron's keynote speech encapsulated this transformation, framing AI as a geopolitical and economic asset, rather than just a regulatory challenge. The emphasis placed on France's nuclear-powered AI infrastructure (data centers), as well as its potential to attract sustainable AI investment, illustrates a broader trend: countries are prioritizing geopolitical aspects of local technology development as much as the need to respond to risks or security concerns.

This perspective was evident in the international response to the summit, with some countries favoring a balancing approach and others arguing for fewer barriers or limitations. For this reason, Paris marks a moment of transition in the focus of AI governance, in which countries prioritize approaches based on specific strategic issues and advantages, rather than fostering a single global governance.

One of the most notable discussions at the summit was the role of sustainability in AI governance. With AI models being potential major energy consumers, the summit underscored the importance of integrating sustainable energy solutions into AI infrastructure.

France has taken advantage of its leadership position in nuclear energy to present itself as a leader in AI sustainability, while countries with high renewable energy potential — such as Brazil — have been encouraged to capitalize on their clean energy resources to attract investments in AI. This focus on sustainability aligns with Brazil's existing AI strategy. It also suggests a new path for AI governance, based on institutions that not only mitigate risk but also promote AI-driven solutions to climate change and sustainable computing.

In this context, the AI Action Summit in Paris represents a window into a second wave of AI governance institutions, moving beyond safety-focused AISIs.

The summit in France offers a crucial lesson to Brazil: the future of AI leadership belongs to those who create an approach to safety and opportunity in AI.

1.2 REFERENCE INSTITUTIONS

From the earliest days of discussions at the UK Summit, the proposal to establish an AISI was inspired by two successful institutional models: the International Panel on Climate Change (IPCC) and the International Atomic Energy Agency (IAEA).

These above are an original reference of a center along the lines of AISI, but other organizations can serve as a reference.

For example, the **Financial Action Task Force** (FATF) develops policies to combat money laundering, terrorist financing, and other financial crimes, promoting international standards and collaboration among member countries. Both organizations conduct assessments to identify risks and provide recommendations, without having binding regulatory power, and both are organizations that rely on cross-industry partnerships.

Another reference is the **European Organization for Nuclear Research** (CERN). Like our vision for AISO, it focuses on the cutting edge of scientific and technological research, in this case, conducting cutting-edge research in fundamental physics. Both rely on international, multi-stakeholder collaboration, and have a global impact. CERN is also an intergovernmental organization composed of 23 member states.

It is also worth highlighting the parallel with the European Centre for Algorithmic Transparency (ECAT), which conducts technical analyses and transparency audits of algorithmic systems used by major online platforms. ECAT contributes to the development of methodologies that inform industry best practices and support the establishment of evaluation benchmarks. Furthermore, it serves as a knowledge hub, fostering global dialogue and providing technical expertise to the European Commission in the formulation and implementation of public policy.

In addition, it is important to mention the Brazilian experience of the National Institute of Metrology, Quality and Technology (INMETRO), a reference for standardization and systematization and of scientific research focused on the reliability and safety of AI. And it is worth noting that the Brazilian Plan for Artificial Intelligence (PBIA) proposes the creation of the National Center for Algorithmic Transparency and Trustworthy AI, aimed at promoting explainability, auditability and supervision of AI systems in use in the country.

1.2.1 INTERNATIONAL PANEL ON CLIMATE CHANGE (IPCC)

Established in 1998 by the United Nations Environment Programme (UNEP) and the World Meteorological Organization (WMO), the Intergovernmental Panel on Climate Change (IPCC) is an intergovernmental institution that brings together 195 member states and several observers.

The IPCC's primary mission is to provide governments with regular and unbiased assessments of the science behind climate change and to support informed policymaking.

The IPCC operates through its highest body, the plenary, where representatives decide on strategic aspects such as the work program, the structure and mandate of its working groups, and the scope of its assessment reports.

The model presented by the IPCC facilitates broad international cooperation and consensus-building, and is seen as the model for the proposed International Panel on AI Safety (IPAIS). Inspired by the IPCC, IPAIS would dedicate itself to the mission of harmonizing views and policies on AI safety on a global scale. Without adopting a prescriptive stance, the panel would provide a structured platform for scientific evaluation and policy guidance, offering insights that could significantly shape international AI safety standards and governance.

However, it is necessary to recognize the fundamental differences between the challenges addressed by the IPCC and those posed by AI. While climate change policy focuses on adapting to external environmental changes, AI safety requires the development of safe and responsible technology by corporations and companies. This distinction underscores the need for AI-specific governance structures, which may diverge significantly from those of the IPCC.

In response to these complexities, the proposal for an IPCC-inspired IPAIS emphasizes the need for an independent, expert-led body that can provide governments with an in-depth understanding of the capabilities, risks, and potential impacts of AI. AISIs must not only evaluate scientific evidence, but also develop global methodologies for evaluating AI systems, in a manner similar to the IPCC's approach to climate data. In addition, by fostering an international network of AI safety institutes, just as the IPCC does, AISIs cultivate a community of AI experts to facilitate research collaborations.

1.2.2 INTERNATIONAL ATOMIC ENERGY AGENCY (IAEA)

The International Atomic Energy Agency (IAEA) was established in 1957 amid concerns about the proliferation of nuclear technology and its potential for misuse. The IAEA promotes the peaceful use of nuclear technology, while regulating and verifying compliance to ensure that nuclear materials are not used for military purposes.

The concept of an international body that offers expert guidance, auditing capacity and standardization is seen as an inspiration for the global model of AISIs.

The IAEA operates on three pillars: safety and security, science and technology, and safeguards and verification. In turn, these pillars align with the goals of the AI & Safety Institutes (AISIs). For example, the IAEA's role as a verification body under the Treaty on the Non-Proliferation of Nuclear Weapons (NPT) underscores its authority in risk mitigation and international cooperation.

However, adapting this model to AI governance involves addressing distinct challenges. Unlike nuclear technology, AI systems pose less of a threat than nuclear weapons. They present a wide spectrum of social risks and harms, such as discrimination and unequal treatment, but AI is not a weapon in itself. This requires a governance structure that addresses high-impact scenarios and everyday applications — but it is a very different need than that involving nuclear technology.

More importantly, leaving any margin of doubt about the scope of the institute, as well as conceptual ambiguities around safety and security, opens space for securitization processes that easily convert general debates on AI governance into national and international security issues, in a way that anticipates and inhibits public debates on more relevant issues from a development perspective.

AI systems are also widely and rapidly deployed in a variety of industries, unlike nuclear technologies, which are more concentrated and therefore easier to monitor. This diffusion makes the development of a central regulatory body similar to the IAEA unlikely. In addition, AI technologies are primarily developed by private entities, which creates potential conflicts of interest and raises concerns about institutional independence.

1.2.3 EUROPEAN CENTRE FOR ALGORITHMIC TRANSPARENCY (ECAT)

The European Centre for Algorithmic Transparency (ECAT) is a center of excellence created by the European Commission in April 2023 with the mission of providing technical and scientific support for the implementation of the Digital Services Act (DSA), as well as promoting research on the impact of large-scale algorithmic systems used by digital platforms and search engines.

ECAT's operating structure is interdisciplinary and involves experts in data science, artificial intelligence, social sciences, law, ethics, and public policy. Its activities are organized into three main pillars: technical and operational support for regulation, scientific and prospective research, and network and community building. The center conducts detailed technical investigations on algorithmic systems —including transparency and risk inspections, testing, and analysis—to support enforcement of DSA compliance by Very Large Online Platforms (VLOPs) and *Very Large Online Search Engines* (VLOSEs). These operators are required to identify and mitigate systemic risks associated with the algorithms, such as the spread of misinformation, algorithmic discrimination, and impacts on mental health, especially of young people

In the field of governance, ECAT acts as a technical center under the responsibility of the JRC, reporting directly to the European Commission and interacting with academic institutions, civil society, regulators and the private sector. In addition to providing guidance on data access for researchers, the center also contributes to the development of practical methodologies to ensure that algorithms are fair, auditable, explainable, and accountable.

1.2.4 NATIONAL INSTITUTE OF METROLOGY, QUALITY AND TECHNOLOGY (INMETRO)

The conformity assessment framework of the National Institute of Metrology, Quality and Technology (INMETRO) aims to ensure that products, processes and services sold in Brazil meet minimum requirements for safety, quality and performance. Over the years, especially from the 1990s onwards, INMETRO began to take a more active role in the formulation and management of conformity assessment schemes, with a focus on ensuring technical standards that favored both consumer protection and the competitiveness of the national industry.

INMETRO (Law No. 9,933 of 1999) now has the competence to prepare technical regulations and conduct conformity assessment activities with regard to safety, human, animal and plant health, environmental protection and the prevention of misleading business practices.

INMETRO is responsible for the Brazilian Conformity Assessment System (SBAC), composed of a network of accredited bodies, laboratories, and inspection entities, with the Institute acting as the central coordinating body. The main instruments of this system are the Technical Regulations (RTs), which define the mandatory requirements for products and services, and the Conformity Assessment Programs (PACs), which operationalize the verification of compliance with these requirements.

The inspection is carried out through the Brazilian Network of Legal Metrology and Quality (RBMLQ-I), which collects samples of products already sold, sends them to accredited laboratories for technical tests and, in case of non-compliance, initiates administrative proceedings as established by law.

In addition to INMETRO, Brazil has the National Quality Infrastructure Strategy (ENIQ), with the objective of strengthening and modernizing the national quality infrastructure system in Brazil. Quality infrastructure (QI) is understood as an integrated set of institutions, policies, practices, and regulatory frameworks that ensure the safety, compliance, competitiveness, and sustainability of products, processes, and services. Its fundamental components are: metrology, technical regulation, standardization, conformity assessment, accreditation and market surveillance.

1.3 EXISTING AISIS

1.3.1 UNITED KINGDOM

DATE CREATED:

November 2023

MISSION

The mission of the UK's AISI is to provide the government with an empirical understanding about the safety of advanced AI systems. As a research organization linked to the UK's Department of Science, Innovation and Technology, the institute seeks to minimize surprises arising from rapid and unexpected advances in AI, thereby enhancing public safety and human well-being.

FUNCTIONS

The UK's AISI is dedicated to understanding the risks associated with advanced AI technologies. This includes testing advanced AI systems, informing policymakers about their risks, and fostering cross-industry collaboration to mitigate those risks. In addition, AISI is responsible for promoting research in the common good and strengthening AI development practices and policies on a global scale.

STRUCTURAL MODEL AND STAKEHOLDERS

In a way, the UK's AISI operates as a startup within government, combining government authority with the expertise and agility of the private sector. This approach allows the institute to operate efficiently and respond quickly to the dynamic field of AI safety. In effect, the institute has recruited high-level talent from both the public and private sectors, increasing its capacity to conduct impactful projects. In 2024, the United Kingdom and the United States signed a Memorandum of Understanding (MOU) that establishes collaboration between the two countries for the development of tests for the most advanced AI models. Effective immediately, the U.S. AND U.K. AISIs have outlined plans to build a common approach to AI safety testing and share their capabilities to ensure that these risks can be addressed effectively.

REGULATORY CONTEXT

AI regulation in the UK is structured to balance encouraging innovation with managing the risks associated with such technologies. This balanced approach is articulated in the *white paper* entitled *"Pro-Innovation Approach to AI Regulation"*, presented to Parliament in 2023. The UK's regulatory framework, in which the institute operates, is designed to encourage innovation without imposing overly prescriptive rules. This framework allows AISI to self-regulate according to established guidelines, ensuring both the advancement and safe deployment of AI in various sectors. It should be noted, though, that the UK's AISI is not a regulatory body, but rather a research organisation linked to the UK Government's Department of Science, Innovation and Technology.

The recent name change from the AI Safety Institute to the AI Security Institute in the UK signals a fundamental shift in the country's focus on artificial intelligence. This change, announced by Technology Secretary Peter Kyle at the Munich Security Conference in February 2025, emphasizes a more security-oriented approach, with a specific focus on preventing the misuse of AI in cyber threats and related security risks. This is not a complete change in its mandate or structure – the focus on security was already a matter of concern for the institute, but it has now become a more explicit focus.

1.3.2 UNITED STATES OF AMERICA

DATE CREATED:

November 2023

MISSION

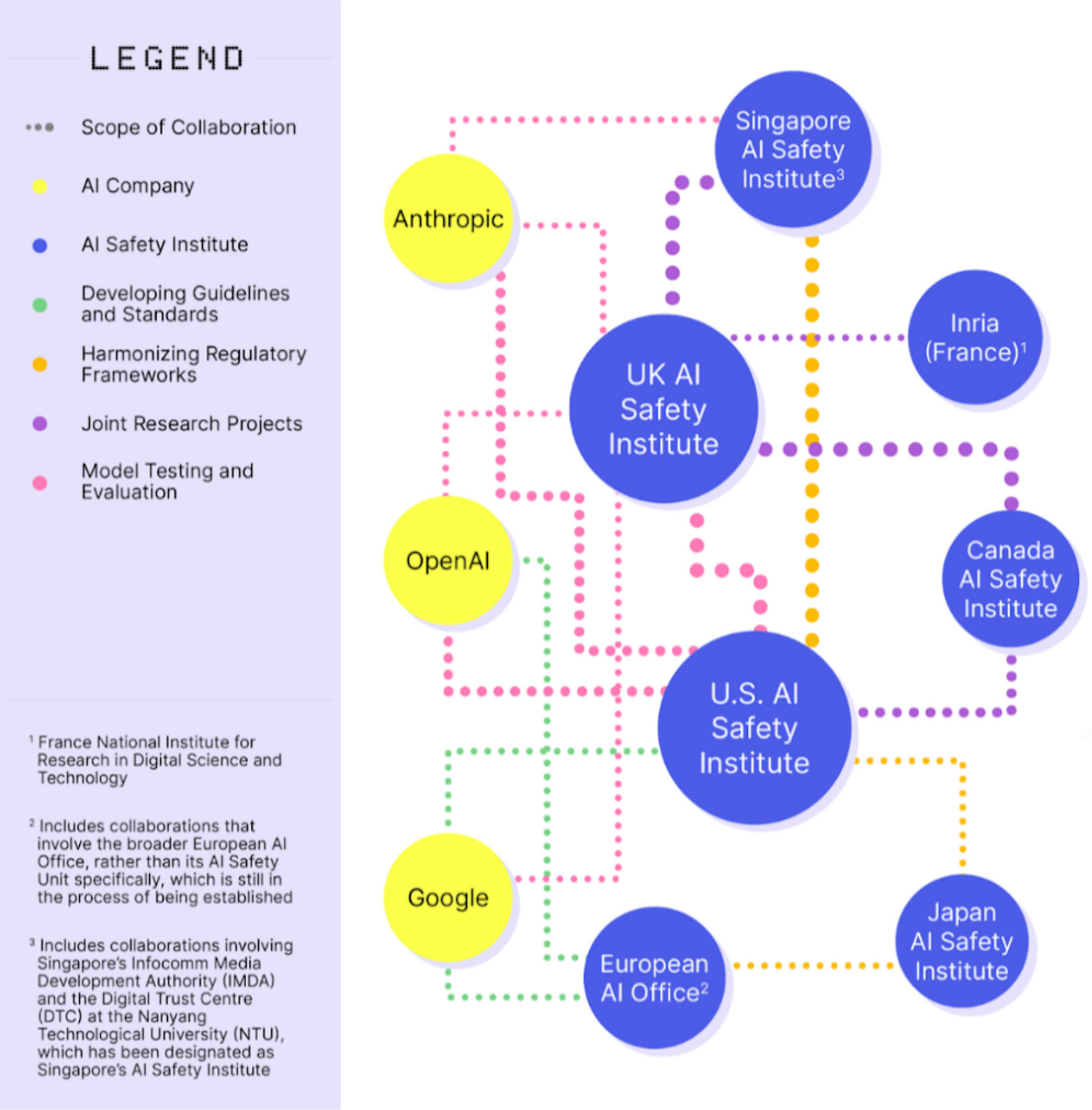
The U.S. AISI operates under the direction of the National Institute of Standards and Technology *(NIST)*. Its mission is to help advance, understand, and mitigate the risks associated with advanced AI.

FUNCTIONS

The core functions of the U.S. AISI focus on enhancing AI safety. They include specialized research to better understand the capabilities and risks of AI and development of testing standards and protocols to ensure safe AI practices and promoting national and international cooperation on AI safety initiatives. The U.S. AISI places significant emphasis on community engagement by publishing usable tools, benchmarks, and guidance to support the development and safe adoption of AI.

STRUCTURAL MODEL AND STAKEHOLDERS

The U.S. AISI has been structured to operate at the "speed of relevance" while maintaining a flexible and responsive approach to the rapid evolution of AI technologies. He focuses on short- and long-term research projects and empowers his team to drive critical initiatives for AI security, including the creation of interim bodies such as the Testing Risks of AI for National Security (TRAINS) Task Force. Along with its UK counterpart, the U.S. AISI is one of the most influential institutes in the current ecosystem, maintaining strategic connections with other institutions and technology companies. These connections encompass the development of guidelines, the harmonization of structures, the testing of models, and the advancement of joint research initiatives (see the diagram developed by the OECD.AI below, which highlights existing collaborations between AISIs and companies working on AI, showing how they have created a complex web of interactions around AI governance).


LEGEND
••• Scope of Collaboration
● AI Company
● AI Safety Institute
● Developing Guidelines and Standards
● Harmonizing Regulatory Frameworks
● Joint Research Projects
● Model Testing and Evaluation
[1] France National Institute for Research in Digital Science and Technology
[2] Includes collaborations that involve the broader European AI Office, rather than its AI Safety Unit specifically, which is still in the process of being established
[3] Includes collaborations involving Singapore's Infocomm Media Development Authority (IMDA) and the Digital Trust Centre (DTC) at the Nanyang Technological University (NTU), which has been designated as Singapore's AI Safety Institute
Singapore AI Safety Institute[3]
Anthropic
UK AI Safety Institute
Inria (France)[1]
Canada AI Safety Institute
OpenAI
U.S. AI Safety Institute
Google
European AI Office[2]
Japan AI Safety Institute

Source: OECD. AI, 2024

REGULATORY CONTEXT

Although the U.S. IASI does not possess regulatory powers, it seeks to significantly influence industry, government, and societal practices through the development of scientific and technical standards and safety protocols.

It should be noted that the U.S. AISI is currently under review, following the repeal of the Executive Order on AI risks, signed by former President Joe Biden, and the U.S. participation in the Paris AI Action Summit, which introduced a broader strategy to reduce government oversight over AI governance[1].

[1] The United States will serve as the inaugural chair of the International Network of AI Safety Institutes as stated by the U.S. Department of Commerce, in November 2024. Source: https://www.commerce.gov/news/fact-sheets/2024/11/fact-sheet-us-department-commerce-us-department-state-launch-international.

1.3.3 SINGAPORE

DATE CREATED:

April 2024

MISSION

Singapore's AISI was established as part of the country's Digital Trust Centre (DTC) and is linked to Nanyang Technological University (NTU). The DTC's mission is to drive innovation in AI to enable companies to effectively incorporate innovative solutions into their operations, as well as foster a local pool of digital trust experts.

FUNCTIONS

Singapore's AISI functions include detailed testing and evaluation of AI systems to assess their safety and performance. The institute also develops models to ensure the safe design, deployment, and operational integration of AI. Additionally, it establishes strict content assurance protocols to ensure the integrity of AI-generated results. Its goal is to develop significant expertise in local language and content security. Finally, it provides essential technical analysis and recommendations to guide governance and policymaking in AI.

STRUCTURAL MODEL AND STAKEHOLDERS

Singapore's AISI governance model conforms to the country's broader digital and AI strategies. It operates as part of the DTC, leveraging the research ecosystem in Singapore and collaborating internally with different stakeholders and internationally to advance AI safety sciences. In 2024, Singapore established a partnership with the European Union, which aims to strengthen collaboration on AI safety through joint research, exchange of expertise, and development of standardized testing and evaluation protocols.

REGULATORY CONTEXT

Singapore's regulatory framework for AI emphasizes an industry-specific approach rather than general and rigid regulations. This framework supports AISI's mission by enabling flexible and dynamic adaptation to the ever-evolving AI landscape.

1.3.4 CANADA

DATE CREATED:

November 2024

MISSION

The mission of the AI Safety Institute of Canada (CAISI) is to ensure that the Canadian government is well-prepared to understand and mitigate the risks associated with advanced AI systems, with a particular focus on risks such as synthetic content and misrepresentation fraud.

FUNCTIONS

CAISI works through two main research streams. The first, led by Canadian Institute for Advanced Research (*Canadian Institute for Advanced Research*, CIFAR), funds multidisciplinary research on AI safety, with a focus on immediate and long-term risks. The second stream, operated by the National Research Council (National Research Council), focuses on issues of cybersecurity and international AI security, potentially involving classified projects.

STRUCTURAL MODEL AND STAKEHOLDERS

Overseen by Innovation, Science and Economic Development Canada (ISED), CAISI is guided by a board of experts from Canada's three leading AI research institutes – Mila, Amii and Vector Institute – ensuring that the institute's efforts are deeply integrated into the national AI strategy and linked to global AI safety initiatives.

REGULATORY CONTEXT

The Artificial Data and Intelligence Act (AIDA) shapes the regulatory landscape in which CAISI operates. This framework emphasizes the responsible design, development, and deployment of AI systems, ensuring that AI deployments in Canada are safe and non-discriminatory. AIDA establishes a Chief Data and AI Officer to monitor compliance and conduct audits, thereby providing a structured and actionable framework for CAISI's initiatives.

1.3.5 JAPAN

DATE CREATED:

February 2024

MISSION

AISI Japan is dedicated to enhancing the safety and security of AI systems through the development and dissemination of safety standards, conducting safety assessments, and promoting international cooperation. The institute seeks to support the efforts of the public and private sectors in understanding and mitigating AI risks throughout its lifecycle.

FUNCTIONS

The core functions of Japan's AISI involve supporting both the government and the private sector by conducting research on AI safety and examining various evaluation methods. The institute also plays a crucial role in creating standards for AI safety and acts as a hub for consolidating the latest information and promoting collaboration among relevant stakeholders. In addition, he actively engages in national, cross-sectoral, and international collaborations to align with global AI safety initiatives and standards.

STRUCTURAL MODEL AND STAKEHOLDERS

The governance structure of Japan's AISI involves coordination among various government ministries and agencies, reflecting a broad approach to AI safety. This includes partnerships with Japan's Ministry of Internal Affairs and Communications and the Ministry of Economy, Trade and Industry. The institute has ongoing collaborations with the U.S. AISI and the EU AI Office to discuss harmonizing regulatory frameworks around AI risk assessment and mitigation.

REGULATORY CONTEXT

Japan's AI regulatory framework emphasizes agile governance, with the AI Guidelines for Business providing clear, accessible, and voluntary guidance for AI developers, vendors, and users. While these guidelines are not legally binding, they influence legal outcomes related to AI and encourage adherence to principles that balance innovation with social and individual rights.

1.3.6 FRANCE

DATE CREATED:

TBD (Announced January 2025)

MISSION

The French research institutions *Laboratoire National de Métrologie et d'Essais* (LNE) and the National Institute for Research in Digital Science and Technology (INRIA) announced a partnership to create an "AI Evaluation" program that will promote the research and development of testing and evaluation methods for general-purpose AI models at the national level. Although this program has not yet been named as an official AISI for France, nor is there certainty that it will be structured as an institute, detailed information is still pending (an announcement was expected during the Paris Action Summit on AI).

FUNCTIONS

Distancing itself from the safety-centric approach emphasized by the U.K. and U.S. at previous summits, France seeks to promote a vision centered on openness and innovation around AI governance. The country advocates for an AI governance model that fosters an open innovation ecosystem, countering the dominance of major technology players and encouraging a more equitable technological development landscape.

STRUCTURAL MODEL AND STAKEHOLDERS

Although it is not yet known how France will structure its AISI, the country has taken significant steps that indicate a predisposition for international cooperation in this area and tends to reinforce an academic/scientific approach, given the nature of the organizations discussing the issues so far. In February 2024, France and the UK announced a historic new partnership between the UK's AI Safety Institute and INRIA (National Research Institute for Digital Science and Technology) to jointly support the safe and responsible development of AI technology.

REGULATORY CONTEXT

France is subject to the EU AI Law and the French approach to AI regulation is in line with President Macron's advocacy for a multipolar world and a distinct European path in global AI policy. An influential report by the AI commission appointed by the French government offers a window into France's concerns. The report largely dismisses the key concerns surrounding the prototypical risks of AI and instead argues that this narrative is being used to legitimize barriers to entry that would lead to the concentration of AI development on the already dominant players.

1.3.7 EUROPEAN UNION

DATE CREATED:

January 2024

MISSION

The European Union was the first to propose the creation of a governance body for artificial intelligence, starting the EU AI Bureau project in 2021 and officially launching it in early 2024, following the enactment of the EU AI Act. Tasked with implementing the AI Act, the EU AI Office has been given a broader and more robust mandate compared to the first wave of AISIs. This includes maintaining a unit dedicated to AI safety, focusing on evaluating general-purpose AI models and establishing technical partnerships with other AISIs. In addition, the Office represents the EU at AISI-related events and is part of the international network of AISIs.

FUNCTIONS

The EU AI Office is engaged in a comprehensive set of activities to strengthen the safety and governance of AI across the European Union. It conducts extensive research on AI safety and evaluates general-purpose AI models to ensure they meet rigorous safety standards. In addition, the Office develops tools and methodologies to assess AI capabilities and systemic risks, actively ensuring that these assessments are compliant with the AI Act. In terms of policy and strategic development, the Office provides essential guidance for AI policies and practices within the EU, with a special focus on promoting trustworthy AI and optimizing its benefits for society and the economy. The Cabinet is also committed to enhancing international collaboration by working with global partners to influence AI governance and standards.

STRUCTURAL MODEL AND STAKEHOLDERS

The EU AI Cabinet is structured around five units, each focused on specific aspects of AI governance: Excellence in AI and Robotics, Regulation and Compliance, AI Safety, AI Innovation and Policy Coordination, and AI for Social Welfare. In addition, the Cabinet includes advisory roles such as the Lead Scientific Advisor and the Advisor for International Affairs, ensuring that the Cabinet remains at the forefront of AI research and policy.

In order to make well-informed decisions, the AI Office collaborates with Member States and the wider expert community through dedicated forums and expert groups. Recently, the EU and Japan announced their intention to strengthen cooperation between the EU AI Bureau and the Japan AI Safety Institute. The Office also maintains ongoing collaborations with private companies, such as Google and OpenAI, to develop guidelines and standards related to risk and safety assessment in AI. It should be noted, however, that there is an overlap of scientific and regulatory functions.

REGULATORY CONTEXT

The EU AI Office operates under the EU AI Act, the first comprehensive legal framework on AI in the world. This law focuses on ensuring the health, safety, and fundamental rights of people, while providing legal certainty to businesses. The AI Office supports the enforcement of this law by providing guidance, developing compliance tools, and facilitating cooperation between member states.

1.3.8 SOUTH KOREA

DATE CREATED:

November 2024

MISSION

South Korea's AI & Safety Institute was officially launched by the Ministry of Science and ICT (MSIT) at the Pangyo Global R&D Center. As Korea's center for AI safety research, AISI facilitates collaborative research and information sharing between industry, academia, and research institutes in the field of AI safety. Through these efforts, AISI seeks to develop competitive technologies, train qualified professionals in the AI safety industry, and promote AI safety policies.

FUNCTIONS

The main objective of South Korea's AISI is to promote research and advance risk assessment methodologies through collaboration with industry, support the development of safety requirements for AI, and foster international cooperation to harmonize global AI safety standards. According to its inaugural director, AISI is not a regulatory body, but rather a collaborative organization dedicated to supporting Korean AI companies by reducing the risk factors that hinder their global competitiveness.

STRUCTURAL MODEL AND STAKEHOLDERS

South Korea's AISI is structured as a consortium formed by Korean organizations from the industry, academia, and research sectors to promote mutual cooperation in research, evaluation, and R&D in AI safety policy. Member organizations should jointly focus on key initiatives, including the development and validation of an AI safety framework.

REGULATORY CONTEXT

South Korea's AI governance is aligned with the newly established Unified AI Act, which consolidates nineteen separate AI-related proposals. This act integrates ethical standards, transparency, and rigorous risk management into national policy, aiming to foster a trust-based environment for AI innovation while ensuring public safety and well-being. The act also houses the AISI initiative.

1.3.9 BRAZIL

Although Brazil does not have an Artificial Intelligence Security Institute (AISI), the country is advancing in the regulatory discussion on AI governance. In December 2024, the Brazilian Senate approved Bill No. 2,338/2023, which proposes the establishment of the National System for Regulation and Governance of Artificial Intelligence (SIA). This Bill is currently under review in the Chamber of Deputies.

The SIA was designed to oversee AI activities in Brazil, ensuring that they follow established guidelines and respect fundamental rights. A central component to this system is the "competent authority", which is responsible for several critical functions, such as:

OVERSIGHT AND ENFORCEMENT

Monitor compliance with the framework defined by the Bill, conducting audits and imposing sanctions in case of non-compliance. This includes evaluating AI systems to ensure that they do not pose excessive risks.

GUIDANCE AND STANDARDIZATION

Establish guidelines and standards for the development and deployment of AI systems, promoting practices that are ethical, transparent, and aligned with national interests.

COLLABORATION AND COORDINATION

The authority works together with other government bodies, private sector entities, and international organizations to promote a cohesive approach to AI governance. This collaborative effort aims to harmonize regulations and encourage the sharing of best practices.

Given this evolving regulatory landscape, the establishment of an AISO in Brazil must carefully consider the structure established by the SIA, and it is important to focus on the separation of the two institutions and their alignment. By working together with the SIA, an AISO could help bridge the gap between regulation and opportunity, ensuring that AI governance in Brazil is not only protective but also proactive in enabling responsible and strategic technological growth.

1.4 COMPARISON TABLES

1.4.1 COMPARATIVE GOVERNANCE STRUCTURE

Country	Act of Creation	MGovernance & Stakeholder Model	Government Department or Agency
United States	Executive Order	Governed by the National Institute of Standards and Technology (NIST), AISI includes provisions for the creation of temporary bodies, such as the AI Risk Testing Task Force for National Security (TRAINS).	Department of Commerce
United Kingdom	Presented to Parliament by the Secretary of State for Science, Innovation and Technology, following the UK AI Summit.	It works as a "startup" within the government, combining government authority with the expertise and agility of the private sector	Department of Science, Innovation and Technology
Singapore	It is not clear	Part of the Digital Trust Centre, linked to Nanyang Technological University	Infocomm Media Development Authority (IMDA) and National Research Foundation (NRF)
Canada	It is not clear	Overseen by the Department of Innovation, Science and Economic Development (ISED) and working in collaboration with the National Research Council of Canada (NRC)	Innovation, Science and Economic Development Canada (ISED)
Japan	Created by ten ministries and five government-affiliated organizations	Organized by the Information Technology Promotion Agency (IPA), in collaboration with relevant ministries and agencies, including the Government Office	Ministry of Economy, Trade and Industry
France	National Institute for Assessment and Safety of Artificial Intelligence (INESIA) announced ahead of Paris AI Action Summit, February 2025	The objective of INESIA is to bring together the main national actors in assessment and security, such as the French National Agency for the Security of Information Systems (ANSSI), the French Research Institute for Computer Science and Automation (INRIA), the National Laboratory of Metrology and Testing (LNE) and the Center of Experts in Digital Regulation (PEREN), without creating a new legal structure	General Secretariat for Defence and National Security (SGDSN), on behalf of the Prime Minister, and the Directorate-General for Enterprise (DGE)

European Union	EU AI Act	The European Union AI Office engages with Member States and the wider expert community through dedicated expert forums and groups	EU AI Office
South Korea	AI Basic Law	It functions as a government-supervised consortium made up of industry, academia, and research sectors to promote mutual cooperation	Ministry of Science and ICT

1.4.2 STRATEGIES TO FOSTER COOPERATION

Country	Does it cooperate with the private sector?	Do you cooperate with other AISIs?
United States	Yes, with technology companies and industry groups.	Yes, has partnered with AISI UK
United Kingdom		Yes, has partnered with the U.S. AISI for AI safety testing standards
Singapore		Yes, it has partnered with the European Union's AISI
Canada		Yes, it collaborates internationally, but the specific AISIs are not detailed
Japan		Yes, it has partnered with the U.S. AISI and the EU AI Office
France		Yes, has partnered with AISI UK
European Union		Yes, has partnered with AISI Japan
South Korea		Yes, it has partnered with the European Union

1.4.3 ORGANIZATIONAL OVERVIEW OF AISI NETWORK MEMBERS

	United States	United Kingdom	European Union	Japan	Singapore	South Korea	Canada	France	Kenya	Australia
Establishment	February 2024	November 2023	May 2024	February 2024	May 2024	May 2024 (Announced)	April 2024 (announced)	-	-	-
Organization Name	US AISI	UK AISI	EU AI Office	Japan AISI	Singapore AISI	Korea AISI	Canada AISI	-	-	-
Linked to the Organ	National Institute of Standards and Technology (NIST)	Department of Science, Innovation and Technology (DSIT)	Directorate-General for Communication Networks, Content and Technology	Information Technology Promotion Agency	Digital Trust Centre (Centro de Confiança Digital)	Tele communications and Electronics Research Institute				
Financing (USD and local currency)	$10 million (Fiscal Year 2024)	> $65 million/ year (>£50 million/year, 2024–2030)	US$51 million (€46.5 million) (Funding period unknown)		US$7.5M/Year (S$10M/Year), 2023–2027	US$7.2–14.4 million/year (●10–20 billion/year), expected to start in 2025	$36.5 million (C$50 million) (funding period unknown)			
Team	About 20 (current core team)	About 20 (current core team)	About 50 (planned, AI safety unit)	About 23 (current team)		Minimum of 30 people (planned, budget pending)	Not disclosed			
Public List of Functions	US AISI Vision, Mission and Strategic Goals	Introducing the AI Safety Institute	Tasks of the EU AI Bureau	Japan's AISI Functions	Early Research Areas	Not disclosed	Not disclosed			
Published Research or Guidelines	Misuse Risk Management for Foundational Dual-Use Models	View website		View website	AI Governance Model for Generative AI	Not disclosed	Não divulgado			

Caption ▨ No public statement □ Public information

This table is an unofficial translation. The original table was prepared by the Center For Strategic & International Studies in the document The AI Safety Institute International Network - Next Steps and Recommendations by Gregory C. Allen and Georgia Adamson in October 2024. Available at this link.

SECTION 2. PRINCIPLES AND FUNCTIONS OF AISO

NOTE:

this is an open section for discussion. Its objective is to stimulate a debate around the pillars and functions that a Brazilian AISO institution should focus on.

This section describes the functions and provides examples of activities that AISO should cover. We took into account the limitations and criticisms (see IAPS and Alan Turing Institute, for example) directed at first-wave institutions, and also the particular needs of Brazil as an AI ecosystem.

As mentioned above, most AISIs focus more on risks than opportunities. This is evident when reviewing their mandates (which are predominantly focused on security and risk). Inspired by the lessons learned from the first wave of AISI (see IAPS report, in particular in the "Challenges" section), this chapter describes the characteristics and potential functions of an AISO, mainly from a Brazilian perspective.

The design model addresses the following challenges identified in the first wave of AISIs:

SPECIALIZATION

Due to the fact that AISIs specialize in a sub-area of AI governance (such as bias and fairness), they can be seen as a way to "deprioritize" other AI-related issues (such as innovations and competitiveness).

OVERLAPS WITH EXISTING INSTITUTIONS

Considering that the regulation of technologies and products is not a "green field," it remains challenging to determine how new institutions interact with existing regulatory bodies. One concern is to create redundancies with sector-specific international organizations, particularly with Standards Development Organizations (SDOs) and with already established and well-established methodologies, such as the potential overlap with INMETRO, in the Brazilian case.

RELATIONSHIP WITH INDUSTRY AND GOVERNMENT

The challenge identified is to balance the relationship between industry and government. A close relationship between government and business is necessary to ensure technological compensation and capacity development, but it can bring complexities arising from information asymmetries.

2.1 FEATURES

Based on the current scenario, an AISO should have the following characteristics:

(1) have open and independent governance with the participation of different stakeholders,
(2) be led by specialists,
(3) have an advisory (and not regulatory) role,
(4) promote shared methodologies,
(5) to promote national and international connections and
(6) be oriented towards the common good.

It is important to note that AISO is not intended to be a regulatory institution, which means that it should not be conceived with normative powers.

2.1.1. INDEPENDENT GOVERNANCE

PROBLEM

Discussions during the AI Action Summit in Paris demonstrated that organizations run by a single agent can be quickly reshaped and lose their independence.

Here are some lessons learned from the first wave of AISIs:

- **In times of rapid policy change, multistakeholder institutions can present a stable institutional arrangement.** Internet governance has witnessed various forms of multistakeholderism, as is the case of ICANN and the Brazilian Internet Steering Committee (CGI.br). While multi-stakeholder governance brings with it certain drawbacks, such as longer response times and a more conservative approach, it leads to organizational structures that are less likely to be captured by a single agent and also facilitates information sharing and the establishment of a trust-based environment.

POSSIBLE SOLUTIONS

With this in mind, ideally an AISO should have an open and independent form of governance, which means it should be participatory and go beyond relying solely on the government or the private sector.

- One solution is to establish an independent center, with the adoption of a multisectoral approach in governance, bringing together experts from different sectors of society. This format would prevent a single agent from taking control over the entire organization without the clear support of other stakeholders. On the other hand, it makes it difficult for members of the government to formally participate.

- **Another solution is to create a government-run organization based on collaboration with other stakeholders.** Open government mechanisms for non-governmental actors in government-led initiatives are often based on low-level participation mechanisms, such as providing information and consultation. However, more effective participatory mechanisms should at least include collaboration tools. And one of the difficulties is to finance the organization with resources outside those of the State itself.

2.1.2. TECHNICAL AND EVIDENCE-BASED LEADERSHIP

PROBLEM

Due to the rapid pace of AI development and the wide range of implementation areas, there is a high expectation for expert-led, evidence-based information and technical guidance. This leads to the need to increase the supply of scientific contributions and technical results to the ecosystem.

Here are some lessons learned from the first wave of AISIs:

- **The expertise provided by AISIs focuses solely on AI safety.** If we look at the design decisions of the first wave, all agencies were conceived with a high degree of expertise in areas of advanced AI. However, in all cases, the body of experts, the forms of collaboration between experts, and the results of projects largely focus on AI safety.

- **Limited availability of skilled labor.** Due to the shortage of talent and the rapid pace of innovations, there is a high cost for all actors, including the government, to produce technical and scientific standards led by experts. The end result is a general shortage of experts to provide evidence on both safety and opportunities in AI, if each stakeholder acts in isolation.

POSSIBLE SOLUTIONS

One desired solution for AISO is to reduce the overall cost of accessing advanced, expert-led, evidence-based information about AI technologies. The need for information is driven by both the safety of AI (which is handled by first-wave institutions) and opportunities in AI. By focusing on technical experts in AI, AISO will be able to address in depth both the risks and opportunities associated with the accelerated development of this technology.

- **One possible solution is to structure expert-led governance.** One of the criteria for being part of the organization's governance structure should be not only diversity of representation, but also expertise in AI, which would allow it to address the low availability of knowledge about AI opportunities in the ecosystem.

- **Another solution is the creation of routines led by specialists, which must be fast and of high technical interest.** Delivery routines can be based on permanent opportunities for experts to contribute to the results, overseen by a coordinated and centralized effort, which would increase the attractiveness for AI experts to contribute to AISO.

2.1.3 ADVISORY BODY

PROBLEM

Many countries are discussing how to structure their AI governance systems and how they can be integrated into a global conversation on the topic. Part of these efforts involves the establishment of AI regulatory agencies. In all cases, these are organizations that are highly responsible for ensuring compliance with policies. This creates an opportunity for organizations that offer the opposite: exclusively consultative contributions, based on meaningful, actionable insights into the state of AI advancement and its impact on governments, industries, and society at large.

Here are some lessons learned from the first wave of AISIs:

- **ISSIs are focused on technical and scientific issues that complement or support regulatory and enforcement activities.** This includes setting standards and debating testing and methodologies for AI agents.

POSSIBLE SOLUTIONS

One desired solution for AISO is to focus on advisory roles **and inputs, without the intention of exercising regulatory or enforcement powers.** The technical and scientific nature of the institutes is best suited to provide evidence-based advice on standards, methodologies and testing opportunities. AISO could provide guidance on possible pathways for all stakeholders to address AI risks and seize opportunities for AI development.

2.1.4 SHARED METHODOLOGIES

PROBLEM

The first wave of AISIs focused on the production of innovative structures and guidelines, which requires a lot of resources and time. By responding to the need to act quickly and deliver immediate results, the proliferation of ISSI can lead to redundancy of methodologies. This can also overlap with initiatives led by academic institutions and the private sector.

Here are some lessons learned from the first wave of AISIs:

- **It is difficult to develop advanced methodologies and a reciprocity arrangement can be beneficial.** Because AI technologies require advanced technical knowledge, developing methodologies is a costly process. This means that it is important to find ways to create alignment and international cooperation and promote reciprocity.

- **The ecosystem of AI methodologies is growing rapidly.** Due to the high level of investments, both from the private and public sectors, what was once a field with few methodologies has become an ecosystem of multiple initiatives. This leads to a greater need to share methodologies rather than producing them from scratch.

POSSIBLE SOLUTIONS

A desired solution for AISO is to promote more specialization, encouraging the sharing of methodologies and international alignment. This alternative highlights the growing need to promote the reuse and extension of existing resources, either through the development, adjustment or testing of methodologies, or by increasing interoperability, sustainability and efficiency. Shared methodologies and reciprocity arrangements favor shorter delivery times and provide a wide range of opportunities for consultation and participation of external actors, as well as promoting international alignment.

2.1.5 PROMOTION OF INTERNATIONAL COOPERATION AND ALIGNMENT

PROBLEM

The first wave of AISIs emerged in a relatively new but rapidly growing ecosystem. This required organizations to focus on cooperation, promoted both at the international level (e.g., network of AISIs, a multilateral collaboration practice) and in association with major players (both from AI companies and universities specializing in AI).

Due to the complexity of the emerging ecosystem, the need to foster collaborations has increased significantly and requires even more structured opportunities for engagement with third parties. The need for evidence and concrete cases leads to a significant demand for cooperation between stakeholders and sectors. Also to be noted is the increasing number of industry-led collaboration efforts around AI — including collaborations with existing AISIs, as mentioned above — with which an AISO must interact to achieve optimal results throughout its initiatives.

Here are some lessons learned from the first wave of AISIs:

- **AISIs emerge in a "consolidated field" of agencies.** Countries that have established AISIs are known for their ecosystem. This includes other types of safety institutes (e.g., for food, medical devices, etc.) and standardization bodies. For this reason, AISIs must be highly connected to other initiatives to avoid duplication of efforts.

POSSIBLE SOLUTIONS

A desirable solution for an AISO is to function as a "hub of hubs" for national and international collaboration and alignment on AI. This means taking a holistic view and a "government as a platform" approach, where peer-to-peer collaborations are enhanced when there are interactions with AISO. Nationally, AISIs need to focus on diverse collaborations, while internationally they can focus on alignment and interactions with other similar institutes. In addition, AISO would promote two-way international cooperation, allowing Brazil not only to keep up with and align with global AI safety standards, but also to share its own solutions, milestones, and perspectives with other countries. This approach strengthens Brazil's role as an active actor in international discussions, expanding its capacity to influence the global agenda and at the same time incorporate good external practices, in a model based on reciprocity and constructive exchange..

2.1.6 FOCUS ON THE COMMON GOOD

PROBLEM

The first wave of AISIs emerged as government-run agencies and, as such, focused on issues of public concern. Private sector institutions can also be oriented towards the common good, but they also have significant incentives to focus on private interests and market dynamics. As such, we can assume that a focus on public benefit is a common feature for both private and public stakeholders, which can lead to mutually advantageous collaboration, especially when focusing on mitigating risks and promoting opportunities for AI technologies.

Here are some lessons learned from the first wave of AISIs:

- **The main topic of common good addressed by AISIs is safety.** However, there is a growing criticism of the exclusive focus on safety and opportunity, as well as a growing skepticism of some forms of safety debate (e.g., the difference between safety and security, or the focus on specific uses of AI, such as content moderation).

- **There is a shortage of structured tools to foster innovation.** Two common strategies to foster innovation are providing funding directly or cutting red tape for new developments (e.g., sandboxes). These are important strategies, but when we look at the diversity of instruments designed to address security (e.g., benchmarks, readiness reviews, test team building, security testing, etc.), we notice an imbalance between known strategies for dealing with risk and promoting innovation.

POSSIBLE SOLUTIONS

A desirable solution for an AISO is to be exclusively oriented to the common good. Public benefit represents the shared interest between the two types of institutions that emerged in the first wave of ISISs and is also what drives multistakeholder governance in existing experiences (e.g., CGI.br). The focus on public benefit also repre sents an opportunity to more strategically address how regulatory institutions balance security and innovation.

2.2 FUNCTIONS

Based on the lessons learned from the first wave of AISIs, as well as understanding future trends, an AISO organization must provide three essential functions:

- research, which means the advancement of scientific research;
- guided practices, which influence the development of AI through guidelines and protocols;
- and cooperation, which seeks to overcome interaction gaps in the ecosystem.

2.2.1 SEARCH

The research function aims to advance the "science of safety and AI opportunities." The research approach is primarily empirical and problem-oriented, and aims to broaden technical knowledge to improve understanding of how to make AI systems safer and broaden opportunities for AI development.

- ●Examples of research roles in the first wave include conducting safety assessments of advanced (frontier) AI models, assessing potential and emerging risks of the most critical AI systems, providing evidence for policy-making processes, advancing fundamental research in AI safety, and more. AISIs have also been influential in defining testing mechanisms and establishing basic methodologies for the development and safe implementation of AI models, systems, and applications.

- Examples of how AISIs promote research include collaborative research efforts with industry, academia, civil society, and other governments; establishment of formal partnerships; opening of calls for research proposals; and providing funding for scholarships.

2.2.2 PRACTICAL APPROACHES

The guided practices function aims to influence practices through applied inputs such as technical guidelines and standard protocols. The guided practices approach is to increase harmonization of the adoption of AI technologies by providing easy access to common and repeatable rules, guidelines, or characteristics for activities or their outcomes.

- Examples of first-wave-oriented practice roles include the development of technical safety guidelines and tools/techniques; harmonization of standards; methodologies for risk assessment in AI; AI model evaluations; definition of principles for AI safety standards and best practices; and software libraries, among others.

- Examples of how AISIs promote guided practices include the publication of guidelines and technical principles, joint testing exercises, conducting stress tests, facilitating the development and standardization of methodologies, and promoting cooperation and alignment with international, regional, national, and industry standards.

2.2.3 COOPERATION

The cooperation function aims to fill interaction gaps in the ecosystem and promote international alignment. Opportunities for cooperation include those that are multisectoral (e.g., policy contributions), facilitated interactions (e.g., dialogues between market actors), and international exchanges (e.g., international network of ISISs).

- Examples of the first wave of collaboration include promoting AI opportunities based on local needs and requirements, facilitating multi-stakeholder cooperation and knowledge sharing, raising awareness, conducting regular dialogues between countries and actors, and acting as an information hub. There is a significant opportunity to facilitate international alignment and reciprocity in terms of technical standards and scientific methodologies.

NEXT STEPS

This is an **exploratory** document that aims to stimulate debate among stakeholders and contribute to Brazil advancing discussions on the subject, both nationally and internationally.

In February 2025, the document was discussed with more than 40 peers from the *Network of Centers*, a worldwide network of research centers on technology and society. The meeting took place in Paris, during the AI Action Summit. The following month, the document was discussed with more than 25 Brazilian actors, two-thirds coming from representatives of government organizations, and the rest from the private sector and academia.

Both meetings were closed and held based on the Chatham House Rules, in which mention of content is allowed but without attribution to the source.

We systematize below the main points learned through the consultations:

FINANCING

It is desirable that the organization is properly linked to a governing body, but flexibility is needed both in governance and in operation and financing. In this sense, three paths were suggested: public financing, private financing, and mixed. One of the possibilities, for mixed financing, is the creation of an ICT (Institute of Science and Technology) or to rely on an existing STI (Science, Technology, and Innovation) institution related to the topic – INMETRO, for example, is such an institution.

SECURITY THEME

It is desirable that the AISO stick to the macro theme of safety and risks, in the figure of "*AI red lines*". The term refers to international documents that suggest areas of impossibility of using the technology, such as the UNESCO Recommendation on Ethics in AI, which banned the use of AI for mass surveillance and social scoring, the EU AI Act model, which lists prohibitive uses of the technology, the Hiroshima AI Process statement, among others.

POTENTIAL INTERACTIONS AND REDUNDANCY

There are clear concerns of potential conflicts between existing and future institutions. AISO is not designed as a regulatory institution, but as a support body for the others, support that is given in a voluntary and technical way. In this sense, some organizations were suggested to keep on the radar, such as CONMETRO (an interministerial collegiate that performs the function of Sinmetro's normative body and has Inmetro as its executive secretariat), focused even on the industrial sectors, and action 51 of the PBIA (Brazilian AI Plan), with the future creation of the National Center for Algorithmic Transparency and Reliable AI, and innovation laboratories, such as InovInmetro.

INTERNATIONAL ALIGNMENT

The argument that Brazil should not be isolated from other countries, in particular from the other AISOs of the first wave, was reinforced. The importance of different bodies connecting to international organizations was suggested, to the extent of their focus and purpose. For AISO, the suggestion is to complement and facilitate alignment with the international AISO network.

SANDBOXES

The report did not analyze the potential of AISO to act with sandboxes in particular, especially in sectoral support. It is an interesting opportunity for research functions (best practices, learning), guiding practices (such as harmonization of existing methodologies) and cooperation (fostering the exchange of experiences).

MAPPING THE ECOSYSTEM

It was suggested to carry out a mapping of institutions in the ecosystem, as there are maps of internet governance and other technologies.

ADVOCACY OR KNOWLEDGE

The desire to supply the ecosystem with a technical, knowledge, and research solution, rather than advocacy, was reflected. The rationale is that by adopting the position of advocacy for unresolved problems of other organizations, there is potential interest in reducing the technical aspect in favor of the direct impact on specific public policies.

NATIONAL SOVEREIGNTY

Different speeches reinforced the need to position Brazil as a producer of technology, rather than just a consumer of solutions. This was emphasized with the profile of structuring the promotion of opportunities for the use of technology in the country.

ROLE OF RESEARCH AS A BASIS FOR PUBLIC DECISIONS AND SOLID INNOVATION

Several participants highlighted the centrality of scientific research to support public decisions and ensure innovation and technological growth safely. It was suggested that the AISO function as a space to mitigate "political noise" and produce scientific evidence that guides both public policies and the actions of companies.

BIFURCATION IN INTERNATIONAL APPROACHES

The discussion pointed out that the "second wave" of institutes is dividing into two strands: one focused on security (with priority for national security, as in the United Kingdom), and another more focused on opportunities and national development (such as France and India). In this context, it was suggested that Brazil has a strategic window to build a hybrid model, which balances these trends based on its priorities of inclusion, innovation, and sovereignty.

EXPLORE EXISTING INSTITUTIONS

There was mention of SINMETRO, CONMETRO and INMETRO as structures already consolidated in the country that can serve as a basis for AISO. The idea is to take advantage of the existing infrastructure — which works with conformity assessment, standardization, and metrology — to facilitate the creation of AI assessment methodologies. In addition, it was suggested that this experience be valued as a Brazilian differential in the international scenario.

CEBRI Board of Trustees

International Advisory Board

Albert Fishlow
Alfredo Valladão
Amanda Klabin
Antonio Patriota
Clarissa Lins
Felix Peña
Flávio Damico
Hussein Kalout
Ivan Sandrea
Izabella Teixeira
Jackson Schneider
João Cravinho
Joaquim Levy
José Luiz Alquéres
Leslie Bethell
Marcos Caramuru
Monica de Bolle
Paolo Bruni
Philippe Joubert
Sebastião Salgado
Tom Shannon
Victor do Prado

Senior Researchers

André Nassif
Antonio Lavareda
Danilo Marcondes
Diego Werneck Arguelhes
Ernani Torres
Ernesto Mané
Gabriel Galípolo
Gregório Cruz Araújo Maciel
Guilherme Dantas
José Juni Neto
Marcio Kahn
Marcus André Melo
Pedro Vormittag
Raphael Gustavo Frischgesell

Senior Fellows

Abrão Neto
Adriano Proença
Ana Célia Castro
Ana Paula Tostes
André Soares
Andrea Hoffmann
Antonio Augusto Martins Cesar
Aspásia Camargo
Benoni Belli
Carlos Milani
Carlos Pereira
Daniela Lerda
Dawisson Belém Lopes
Denise Nogueira Gregory
Diego Bonomo
Evangelina Seiler
Fabrizio Sardelli Panzini
Fernanda Cimini
Fernanda Magnotta
Francisco Gaetani
Guilherme Casarões
Igor Rocha
José Mário Antunes
José Roberto Afonso
Kai Lehmann
Larissa Wachholz
Leandro Rothmuller
Leonardo Burlamaqui
Leila Sterenberg
Lia Valls Pereira
Lourival Sant'Anna
Maria Hermínia Tavares
Marianna Albuquerque
Mário Ripper
Matias Spektor
Mathias Alencastro
Miguel Correa do Lago
Monica Herz
Mônica Sodré
Oliver Stuenkel
Patrícia Campos Mello
Paulo Sergio Melo de Carvalho
Pedro da Motta Veiga
Philip Yang
Rafaela Guedes
Ricardo Ramos
Ricardo Sennes
Rogério Studart
Ronaldo Carmona
Sandra Rios
Sergio Gusmão Suchodolski
Tatiana Rosito
Vera Thorstensen

Member entities

ABEEólica - Elbia Gannoum
Aegea - Radamés Casseb
Altera - José Carlos Elias Junior
Amazon Web Services (AWS) - Nayana Rizzo Sampaio
apexBrasil - Jorge Ney Viana Macedo Neves
Banco Bocom BBM - Alexandre Lowenkron
BASF - Manfredo Dieter Rubens
BCG - Arthur Ramos
BHP Billiton - Fernanda Lavarello
BNDES - Gabriel Aidar
BRF - Bruno Machado Ferla
BMA Advogados - Francisco Müssnich
bp - Angélica Ruiz Celis
Brookfield Brasil - Luiz Ildefonso Simões Lopes
Caju - Eduardo Braz del Giglio
CAF - Guilherme Cardoso
Cargill - Lígia Dutra Silva
Consulado Geral da Bélgica no Rio de Janeiro - Caroline Mouchart
Consulado Geral da Irlanda em São Paulo - Fiona Flood e Eoin Bennis
Consulado Geral do México no Rio de Janeiro - Héctor Humberto Valezzi Zafra
Consulado Geral da Noruega no Rio de Janeiro - Mette Tangen
Consulado Geral dos Países Baixos no Rio de Janeiro - Sara Cohen
Dynamo - Luiz Felipe Campos
EDP - João Marques da Cruz
Embaixada da China no Brasil - Qu Yuhui
Embaixada da Suíça - Sergio Bardaro
Embaixada do Reino Unido - Stephanie Al-Qaq
Embraer - Jackson Schneider
ENEVA - Marcos Cintra
ENGIE Brasil - Maurício Bähr
Etel Design - Lissa Carmona
ExxonMobil - Valeria Rossi
Equinor - Veronica Coelho
Finep - Ronaldo Gomes Carmona
Galp - Daniel Elias
Google - Juliana Moura Bueno
Grupo Ultra - Marcelo Araujo
Klabin - Amanda Klabin
Lorinvest - Haakon Lorentzen
LTS Investments - Luiz Eduardo Osorio
Machado Meyer - José Virgílio Lopes Enei
Microsoft - Alessandra Del Debbio
Museu do Amanhã - Ricardo Piquet
Neoenergia - Solange Ribeiro
Origem Energia - Luiz Felipe Coutinho Martins Filho
PATRI - Carlos Eduardo Lins da Silva
Petrobras - Pedro Henrique Bandeira Brancante Machado
Pinheiro Neto Advogados - Ricardo Coelho
Prefeitura do Rio de Janeiro - Lucas Felipe Wosgrau Padilha
Prefeitura de São Paulo - Marta Suplicy
Promon Engenharia - Antonio Bardella Caparelli
Prumo Logística - Bárbara Bortolin
PUC-Rio - Padre Anderson Antonio Pedroso S.J.
Sanofi - Felix Scott
Shell - Flávio Ofugi Rodrigues
Siemens - Luis Felipe Gatto Mosquera
Siemens Energy - André Clark Juliano
SPIC Brasil - Adriana Waltrick
Huawei - Steven Shenjiangfeng
IBÁ - José Carlos da Fonseca Junior
IBRAM - Wilfred Bruijn
Suzano - Mariana Lisbôa
Instituto Arapyaú - Renata Soares Piazzon
Instituto Clima e Sociedade - Rodrigo Fiães
Syngenta - Grazielle Parenti
Itaú Unibanco - Luciana Nicola
JBS - Carlos Alberto Macedo Cidade
Total Energies - Ulisses Martins
UNICA - Patricia Audi
Vale - Gustavo Niskier
Veirano Advogados - Alberto de Orleans e Bragança
Vinci Partners - Alessandro Monteiro Morgado Horta
Volkswagen Caminhões e Ônibus - Antônio Roberto Cortes
TIM - Mario Girasole

CEBRI Team

CEO
Julia Dias Leite

Strategic Planning Director
Luciana Gama Muniz

Academic Director
Feliciano de Sá Guimarães

Director of HR, Operations and Finance
Flavia Theophilo

Director of Corporate Relations
Henrique Villela

Director of Events, Communication and Marketing
Renata Bellozi

Projects

Projects Deputy Director
Léa Reichert

Deputy Director Specialist in Geopolitics and International Trade
Ariane Costa

Energy and Climate Change Specialist
Julia Paletta

Projects Coordinator
Gustavo Bezerra

Projects Coordinator
Laís de Oliveira Ramalho

Projects Analyst
Daniel Fontes

Partnerships and International Cooperation Assistant
Fabricio de Martino

Projects Trainee
Francesca Tozzi

Partnerships and International Cooperation Intern
Marcelo Gribel

Project Intern
José Ricardo Araujo

Project Intern
Marianne Moreira

Deputy Director of Partnerships and International Cooperation
Teresa Rossi

Projects Manager
Thais Jesinski Batista

Partnerships and International Cooperation Manager
Beatriz Pfeifer

Projects Coordinator
Isabella Ávila

Projects Coordinator
Laura Escudeiro

Junior Projects Analyst
Catarina Werlang

Projects Assistant
Felipe Cristovam

Projects Intern
Leonardo David

Projects Intern
Maria Fernanda Ferreira

Executive Advisor
Gustavo Heluane

Project Intern
Júlia Soares

Government and Institutional Relations

Institutional and Government Relations Advisory Board
Antonio Souza e Silva

Corporate Relations

Corporate Relations Manager
Paula Lottenberg

Institutional Relations Manager
Nana Villa Verde

IT

IT Coordinator
Eduardo Pich

IT Intern
João Paulo de Carvalho Pereira

Audiovisual and IT Support Technician
Vagner Oliveira

Communications and Events

Communication and Marketing Manager
Gabriella Cavalcanti

CEBRI-Journal Editorial Coordinator
Bruno Zilli

Courses Coordinator
Davi Bonela

Events Coordinator
Julia Cordeiro

Communication and Marketing Analyst
Beatriz Andrade

Events Analyst
Mariana Carluccio

Communication and Marketing Analyst
Laura Motta

Communication and Marketing Analyst
Lucas Machado

Events Analyst
Vitória de Faria Ribeiro

Events Analyst
Maria Eduarda Cerca

Communication and Marketing Assistant
Alice Nascimento

CEBRI-Journal Editorial Assistant
Victoria Corrêa do Lago

Institutional Communication Consultant
Lydia Medeiros

Events Assistant
Rainara Costa

Events Assistant
Julia Nani

Administrative and Financial

Deputy Administrative Financial Director
Juliana Halas

Deputy Financial Director
Fernanda Sancier

Finance Coordinator
Gustavo Leal

Human Resources Administrative Coordinator
Marcele Reis

Financial Analyst
Miguel Junior

Executive Secretary
Patricia Burlamaqui

General Services Assistant
Vânia Souza

General Services Assistant
Joilson Ribeiro

CEBRI

Centro Brasileiro de Relações Internacionais

Rua Marquês de São Vicente, 336
Gávea, Rio de Janeiro - RJ - Brasil
22451-044

Tel: +55 (21) 2206-4400
cebri@cebri.org.br

@cebrionline

cebri.org