APPLIED RESEARCH PROJECT 2025/26

in partnership with

GENEVA GRADUATE INSTITUTE

Globethics

RESEARCH REPORT

MITIGATING CATASTROPHIC AI-RISKS:
THE ROLE OF INCLUSIVE & ANTICIPATORY INFORMAL DIPLOMACY

By Cora Ebner, Eesha Jamal, Greta Franco & Isabela Duleba

Executive Summary

There is a growing gap between accelerating artificial intelligence (AI) capabilities and AI governance. While catastrophic AI risks are increasingly recognised in political, technical and policy debates, existing diplomatic processes and governance mechanisms remain slow, fragmented and constrained by geopolitical competition, private-sector interests, uncertainty and low trust. In this landscape, the relevance of traditional diplomacy is being questioned. The report therefore examines the potential role of informal diplomacy, which moves cooperation beyond state-actors and formal negotiation channels, in filling this gap. Furthermore exploring how inclusive, trust-building dialogue and cooperation mechanisms can be employed to mitigate catastrophic risks from AI technology developments. The report understands catastrophic risks as severe, irreversible or globally disruptive harms to human societies (International AI Safety Report 2025; WEF, 2025). These include risks linked to warfare utilisation, loss of control, critical infrastructure disruption, malicious use, large-scale mis- and disinformation, socioeconomic disruption and the escalation of interstate animosity. The report does not treat informal diplomacy as a substitute for binding regulation, but as a practical bridge that can build trust, align risk perceptions, provide a channel for cooperation and a shared language among actors.

Methodologically, the project combines desk research, expert surveys and semi-structured expert interviews, analysed through a context-mechanism-outcome (CMO) framework. Hereby, the report analyses how governance contexts, such as institutional capacity, political prioritisation, incentive structures and shared risk perceptions, shape the conditions for cooperation, the potential for informal diplomacy mechanisms, and the governance outcomes they can support. Latin American countries are used as the primary regional case study, with African Union member states serving as secondary comparator. This allows the report to assess how regional contexts, particularly the existence of a regional coordination framework, in the case of the African Union, or the lack of it, in the case of Latin America, shape the prospects for informal diplomacy mechanisms.

The findings confirm that experts perceive catastrophic AI risk as a serious and urgent governance concern. However, experts rarely describe these catastrophic risks in isolation, instead they were most often linked to weak and slow governance capacity, lack of control and democratic decision-making. These factors also influence the perceived threat level of risks, for which military and warfare utilisation was ranked as the most severe overall. This was followed by systemic societal risks, whose threat perception was elevated by the lack of governance capacity to manage socioeconomic disruption and threats to information integrity. The findings furthermore confirmed that current conditions for cooperation remain weak. Survey respondents identified lack of political prioritisation, legal authority, technical knowledge and alignment of interests as the key barriers to anticipatory governance. Trust is especially limited in relation to the private sector. Although companies hold essential technical expertise and model development control, they were the lowest ranked actor across governance characteristics, largely due to concerns about profit-incentives, lack of transparency and accountability. By contrast, were academic and research communities, as well as civil society organisations, viewed as the most trustworthy and aligned with public interest, but lacking the agency needed to effectively shape governance outcomes.

The report further found that cooperation in the near future is most feasible in less technically sensitive areas, especially within capacity-building, incident reporting and shared risk taxonomies. These mechanisms are valuable in that they help develop a common language of risk, reduce uncertainty, and help prepare the conditions that make stronger governance possible, without requiring actors to share sensitive details or compute access. In Latin America, the absence of a regional AI governance framework, combined with political uncertainty, institutional fragmentation, resource constraints and dependence on foreign technology firms,

limits the prospects for rapid convergence. Since binding regional agreements are unlikely in the short term, the report recommends an incremental approach based on soft law, political convergence on risk understanding statements and inclusive regional dialogues involving civil society, academia and technical experts.

Based on the report's findings, this project presents five key policy recommendations. First, the creation of non-binding shared risk taxonomies and multilateral incident reporting frameworks, with a human rights integration into the design, should be a key priority. Second, the development of a standardisation process led by an international organisation, building on efforts by civil society and inter-governmental actors, should serve as a key step to move away from the current fragmentation of AI governance. Third, private sector inclusion should be made more technically meaningful. In this, technical representation, through engaging with model development teams, should be prioritised over relying on policy or diplomatic representatives, while additionally being leveraged in capacity-building. Fourth, due to their superior human-rights and transnational expertise, should civil society organisations be given more agency in standardisation and governance evaluation processes. Fifth, regional governance efforts, especially in Latin America, should employ informal diplomacy to include non-state actors, build trust and public legitimacy, and strengthen anticipatory capacity.

Anticipatory governance requires more than technical risk assessment; it demands sustained cooperation and dialogue across states, private actors, civil society, international organisations and research institutions. Ultimately, the report advances inclusive and anticipatory informal diplomacy as a promising pathway for global AI governance to prevent catastrophic risks. Informal diplomacy can help create the conditions for stronger and more politically and practically feasible governance, by building trust, clarifying risks and connecting actors who are often excluded from formal negotiations.

TABLE OF CONTENTS

1. **Introduction** **5**
2. **Background** **6**
3. **Literature Review** **7**
   3.1 Risks 7
      3.1.1 *Definitions* 7
      3.1.2 *Identifying Catastrophic Risks* 8
   3.2 Governance and Diplomacy 9
      3.2.1 *The Current AI Governance Regime* 9
      3.2.2 *Assessing Regime Quality* 11
   3.3 Non-state Actors and the Role of Informal Diplomacy 11
      3.3.1 *Civil Society* 12
      3.3.2 *Private Sector* 12
   3.4 Regional Case study 14
      3.4.1 *Case Selection* 14
      3.4.2 *Latin America* 14
4. **Methodology and Research Design** **17**
   4.1 Analytical Framework 17
   4.2 Research design 17
   4.3 Limitations 18
5. **Findings** **19**
   5.1 Summary of Key Stakeholders and Governance Capacity 19
   5.2 Perceptions on Catastrophic risks 21
   5.3 Hindering and Enabling Conditions and Mechanisms of Cooperation 23
   5.4 Case Study: Latin American countries 25
6. **Discussion and Implications for Policy** **27**
   6.1 Informal Diplomacy Mechanisms 27
   6.2 Non-state Actors 29
      6.2.1 *Private Sector Inclusion* 29
      6.2.2 *Civil Society and Research Community* 30
   6.3 Regional Insights 30
7. **Final Remarks** **32**
8. **Bibliography** **34**
9. **Appendix** **38**

Annex 1: Qualitative codebook 38
Annex 2: List of abbreviations and key definitions 38
Annex 3: Latin American Countries 38
Annex 4: African Union AI governance 38
Annex 5: Further details on Methodology 38

10. **Remark on the Use of Artificial Intelligence** **38**

1. Introduction

Current global governance mechanisms appear to have been outpaced by the rapid advancement and uncertain trajectory of artificial intelligence (AI) capabilities. While numerous harms from general-purpose AI are already well-established, evidence of additional "catastrophic risks", here defined as the possibility of severe, irreversible, or globally disruptive harm to human societies, are gradually emerging (International AI Safety Report 2025; WEF, 2025). The coming years are identified as a critical juncture in the race between AI capabilities and AI governance (Kokotajlo et al., 2025). While the urgency of efficient anticipatory governance is widely recognised, current attempts reflect uncertainty, mistrust and competition among key actors. In this context, informal diplomacy mechanisms present a valuable opportunity for AI governance, as a way to move beyond the constraints of formal state-to-state negotiations, or state-led multilateral processes, in this report referred to as traditional diplomacy. In contrast, informal diplomacy refers to flexible, less structured channels of cooperation, characterised by non-state actor inclusion and aimed at building trust and dialogue (DiploFoundation, 2021a).

Against this backdrop, this project aims to investigate how informal diplomacy mechanisms can be designed and operationalised, as part of anticipatory governance frameworks, to prevent or mitigate catastrophic risks posed by the advancement of AI. Focusing on the decade centred on 2025 (± five years), the central inquiry of this project asks:

How can inclusive, anticipatory informal diplomacy mechanisms be employed to mitigate catastrophic risks from accelerating AI capabilities ?

The project employs an additional focus on Latin American countries, to explore the question:

What characterises the employment of informal diplomacy mechanisms in anticipatory AI-risk governance in the Latin American countries without a regional governance framework?

To contextualise the landscape in which our inquiry is set, section 2.1 provides an overview of the challenges which define contemporary catastrophic AI-risk governance. Key characteristics are further explored through desk research presented in the literature review in section 3, which begins by exploring "risks" in the context of AI. In section 3.2 the literature review provides further details on diplomacy approaches and governance assessment, followed by 3.3 which focuses on the inclusion of non-state actors.

The latter part of the literature review, section 3.4, is dedicated to Latin America which is utilised as a primary case study, for which African Union (AU) member states will be used as a secondary comparator, allowing the project to assess the current maturity and potential prospect of informal diplomacy mechanisms in the context of Latin American countries without an institutionalised regional AI framework. The methodology behind this, as well as the theoretical frameworks and conduction of expert surveying and interviewing, is described in section 4.

The following sections analyse findings of expert consultations, contextualised by the literature review, employing a context-mechanisms-outcome (CMO) evaluation framework. Section 5 presents the key findings of the conducted expert consultations. The section begins with a summary of key findings on prominent actors in AI governance then moves on to findings on catastrophic risk perceptions, hindering and enabling conditions for cooperation and on feasibility and value of future governance mechanisms.

Finally, subsection 5.4, discusses the qualitative and quantitative data collected on Latin American countries, highlighting overarching themes which characterise the employment of informal diplomacy mechanisms in this case. Final recommendations, for the employment of inclusive, anticipatory informal diplomacy mechanisms to mitigate catastrophic risks from accelerating AI capabilities, are given in section 6, serving as the condensed answer to the central inquiry of the project. The report concludes and gives recommendations for future research in section 7. In the annexes further details can be found, including the qualitative codebook used for data analysis, key definitions and abbreviations, supplemental information on the Latin American and AU contexts and expert consultations.

2. Background

Technical AI developments are rapidly gaining importance on the political agenda, resulting in strategic and geopolitical concerns steering the position of governments towards potential legislative and regulatory measures (Schmitt, 2022). As the World Economic Forum Global Risk Report 2026 identifies the coming decade as the "Age of Competition" (p.15), AI is evolving in an environment that generally lacks shared global norms, trust-building channels or communication frameworks.Technological races towards Artificial General Intelligence (AGI) and potentially superintelligent systems, such as the one between the U.S. and China, illustrate the dangers of competitive environments which may lead actors to deploy unsafe AIs or cede control to AIs (Hendrycks, et al., 2023). Within the landscape of catastrophic risks, which is divided into the three main categories of malicious use, loss-of control, and systemic risks, AI- races are only one of the risks which necessitate, and underline the urgency, of international cooperation on AI governance.

While there is broad consensus that AI developments can generate catastrophic risks, there is still considerable debate about how these risks should be understood and governed. Governance here refers to the framework of rules, systems and processes through which actors articulate collective interests, establish rights and obligations, and mediate differences. The AI governance actors referred to in this report include states, intergovernmental organisations, private sector, civil society and research communities. Over the past several years, a growing body of international frameworks have been created to showcase more concrete guidance on what responsible AI governance should actually look like. Debates on AI governance originated within the field of AI ethics, centred on principles such as transparency, fairness, and accountability. These themes prevail in multilateral organisations, such as UNESCO, whose 2022 Recommendation on the Ethics of AI is presented as a human rights-based approach. It is based on the advancement of principles of transparency, fairness and human dignity, emphasising human oversight of AI systems, and advocating for policymakers to translate core principles and values into action. However, as values-based approaches, ethics-by-design, and other principled recommendations are translated into the development of ethical frameworks and technical governance models, governments' implementation of these remains largely voluntary and dependent on individual states (Schmitt, 2022).

Calls for more binding regulation and transnational governance are increasing, as showcased by the 2024 OECD AI Principles,urging governments to build regulatory environments that are interoperable across borders, to hold developers accountable for the real-world outcomes of their systems, and to ensure that AI remains transparent and explainable to those it affects (OECD, 2024b). The Council of Europe's Framework Convention on Artificial Intelligence presented in September 2024 marks the first legally binding international treaty on the subject (CAIPD, 2026), obliging its 44 signatory states (as of December 2025) to ensure that activities within theirAI systems in their jurisdictions are consistent with human rights, democracy and the rule of law. Nonetheless, many of the most crucial decisions are not being made by democratic institutions or

multilateral forums, but rather through bilateral agreements between governments and AI companies (Alais, 2026). As diplomacy is relocated to new arenas, AI introduces a new disruption, where the effectiveness of such agreements and the relevance of traditional diplomacy are being questioned.

The work towards effective AI international governance, which goes beyond voluntary self-commitments, requires a recognition of both the global and sub-national dimensions of AI policy (Schmitt, 2022). With the cross-border nature of the digital ecosystem and the fact that AI research is dominated by transnational technology firms, it is suggested that purely national AI regulation is inefficient. Simultaneously, civil and private sector actors play a crucial role at the sub-national level to manage the global "AI trust deficit" (Sharma, 2025), however, the extent of their effective inclusion remains largely insufficient. While civil society actors have shown a clear willingness to participate in AI governance, the need for a collective effort to break down systemic barriers is highlighted as necessary to their effective contribution (Gálvez Callirgos, 2025). The private sector on the other hand is presented as both a contested diplomatic actor for AI risks and an indispensable technical partner, due to its superior technical knowledge and capacity to accelerate implementation and enhancing the technical relevance of regulation and norms (Medeiros, 2020; Hatano, 2026). Its governance involvement, however, is limited by concerns about transparency, accountability and lack of binding agreements, reaffirming a need to further explore alternative diplomacy mechanisms in which the private sector capacity can be leveraged while mitigating the structural risks tied to their inclusion.

The challenges caused by commercial power in AI regulation and development also manifest in geographical disparities. Latin America showcases how long-standing patterns of dependency and unequal insertion into the global order results in AI becoming a fundamentally political-economic issue (Vivares et al., 2025). While the employment of AI is rapidly increasing in the region, simultaneous regulatory developments remain hindered by foundational weaknesses, including lack of cross-border coordination, computational infrastructure and technical expertise, as well as weak data governance (Sanchez-Pi et al., 2021). As the need for moving beyond reactive AI regulation increases (Contreras, 2024), informal diplomacy mechanisms present a potential avenue for anticipatory governance in the region shaped by institutional diversity, developmental asymmetries and geopolitical influences.

3. Literature review

3.1 Risks

3.1.1 Definitions

This report solely focuses on catastrophic risks. Unlike ordinary risks, which are manageable and localized, catastrophic risks can lead to devastating, unpredictable outcomes and even systemic collapse. Part of this debate emerges in how policy literature conceptualizes these risks. Some focus on risks that are not produced solely by technology, but instead interact with existing political, economic and social systems. In this sense, AI is viewed as an amplifier of pre-existing issues. According to the International AI Safety Report (2025), AI poses three types of risk. First, malicious use risks, including creating harm through fake content such as deepfakes, cyber offences such as finding and exploiting vulnerabilities in open-source software, and the role it can play in biological weapons development, such as providing detailed instructions for producing toxins. Second, reliability issues, such as general AI's hallucinations that can cause harm if users solely rely on them for medical advice. Within this category, there is also the issue of bias and discrimination, whereby General-purpose AI systems amplify social and political inequalities regarding race, gender and disabilities (International AI Safety Report, 2025). Finally, the third type of risk is systemic risks in which AI exemplifies

broader societal hazards, for instance, environmental risks due to the large energy and water consumption of AI. This also includes market concentration risks, since AI development is dominated by a select few companies, exposing societies to vulnerable systemic risks: if organisations in critical sectors such as in finance or healthcare rely on a small number of AI systems, any disruption to these systems could cause greater failures within these sectors (International AI Safety Report, 2025).

Moreover, not all the aforementioned risks outlined by the International AI Safety Report can be identified as a catastrophic risk. The OECD categorizes AI related risks as hybrid threats, whereby they are catastrophic when they interact with flawed systems and operations that can cause political, social and economic destabilisation (Polchar and Alfonzo Santamaria, 2024). According to the OECDs global risk mapping and International AI Safety Reports findings, AI is placed in the same thematic space as nuclear winter, biodiversity collapse and antimicrobial resistance, falling under the malicious use category of catastrophic risks.

On the other hand, when looking at academic literature, authors focus directly on the technological dimensions, such as the idea that AI models could develop faster than human ability to control or align them with human values. This comprises scenarios of loss-of-control or failures in alignment, especially as AI models are incorporated into sensitive fields such as cybersecurity, biotechnology and the military (Bostrom, 2002). With insights taken from both perspectives, it seems unlikely that these risks stem from a single source. Instead, they most likely arise at the intersection of the technical sector and fragile governance environments (International AI Safety Report, 2025), making them difficult to isolate and manage. Consequently, creating mitigation solutions for these risks cannot rely on technical solutions alone, rather it demands coordination across different sectors and countries, as each of these sectors does not possess the full knowledge, authority or capacity needed to respond effectively (WEF, 2025). While the need for coordination is recognized, there is little known about the conditions under which such coordination takes place in practice, particularly in contexts marked by competition or institutional fragmentation.

3.1.2 Identifying Catastrophic Risks

Across the literature, risk identification commonly begins with structured expert assessments. For instance, a multinational OECD pilot utilised a two-stage process in which experts first generated and shortlisted emerging risks, and then rated each on impact, scope, five-year probability, national capacity and confidence using standardised scales (Polchar & Santamaria, 2024). This demonstrates the measurability of expert perceptions, and that perceived capacity gaps play an essential role in how risks are evaluated.

As illustrated by the OECD, foresight methodologies offer complementary techniques for understanding why experts prioritise certain risks (Polchar, 2024). Practices such as “horizon scanning” support the systematic identification of “weak signals" (developments that are not yet widely recognised), through expert consultations, unconventional data sources and generative AI. Futures and foresight are constant iterative processes, which systematically explore multiple possible futures to inform decision-making and public policy (OECD OPSI, 2024). Exploratory methods, such as cross-impact analysis and contextual risks, help in understanding how geopolitical, institutional, and technical assumptions shape expert reasoning (Polchar, 2024). Thus, foresight strategies allow actors to “perceive, make sense of, and act upon” anticipated futures in the absence of an “absolute future” (OECD OPSI, 2024). This method aligns with anticipatory governance, which similarly relies on early warnings, scenario exploration, and structured stakeholder input to address uncertainty.

Scenario-based exercises further illustrate how analysing sequential "what-if" developments can reveal escalation dynamics, governance failures, and the points where cooperation mechanisms are most urgent and most strained (Kokotajlo et al., 2025). AI 2027, written by former Open-AI researcher Daniel Kokotajlo and several other technical experts, gained major attention in AI discourse globally, reading like a sci-fi thriller. While advancing the idea that people losing their job is the best-case scenario, given that current AI developments open up to the possibility of human life as we know it being effectively ended by the end of the decade, Kokotajlo reaffirms that the scenario was published to bring attention to the urgency of governance (The New York Times, 2025). While the scenario was met with varying degrees of belief among the sequence of events highlights how catastrophic risks do not stem primarily from AI's independent capabilities, but rather from what humans instruct the technologies to do.

Across these perspectives, there is an important but underexplored theme: the conditions under which actors are able to recognize and coordinate responses to emerging threats. While foresight tools are of the utmost importance for predicting emerging risks, they do not fully foretell how cooperation emerges in practice.

3.2 Governance and Diplomacy

3.2.1 The Current AI Governance Regime

AI governance today can be described as a regime complex, i.e., "a collective of partially overlapping and nonhierarchical regimes" governing the same issue (Raustiala and Victor 2004, p. 1). This is illustrated by dialogue and regulation being debated within multiple institutions, such as the ITU, OECD, UNESCO, G7, G20, as well as several regional and national forums, without any central authority tying them together. This regime complexity may be perceived as a weakness when compared to more structured regimes, However, as Keohane and Victor (2011) posit, the flexible structure of these regimes can have its benefits, arguing that regimes are created to fit states' interests, and due to the ever-changing nature of the anarchic system, preferences, information, beliefs and power dynamics are subject to change.

This is especially relevant in the face of current geopolitical competition. Due to its importance in national security, AI can be seen as being directly tied to state survival, with relative gains of cooperation becoming less important, and thus states become less willing to collaborate (Waltz, 1979). Marino et al. (2025) argue that AI governance is not simply a coordination problem, but rather a competition over who defines the norms and infrastructures that will shape future global influence. The competition becomes not only related to technological leadership, but also over who sets the rules of the game, which can create power asymmetries for countries that are not rule-setters. This competition also involves corporations that possess control over technical expertise, giving them a central role in the governance landscape. Therefore, AI governance demands a different approach to diplomacy, as traditional intrastate diplomacy may not take into account situations where private actors are active decision makers.

In environments marked by uncertainty and competing interests, the flexibility of complex regimes can bring some benefits, as it allows for the gradual construction of norms and adaptation to them, something that can be more difficult to achieve in more rigid regimes (Keohane and Victor 2011, p. 16). However, this ends up creating a fragmented environment with ambiguity in the rules, which discourages compliance due to increased uncertainty and transaction costs (Pratt 2018, pp. 4-6). Furthermore, Alter and Meunier (2009) warn of three main consequences. First, this fragmentation creates an environment in which actors engage in choosing the most convenient forum, selecting institutions and rules that are aligned with their interests. Second, this enables

chessboard politics, where actions in one forum are used to influence outcomes in another. Third, regime shifting, where actors shift discussions to other spaces to shape broader governance outcomes. In this sense, regime complexity does not simply describe the architecture of governance, but actively structures how governance is practiced.

Unsurprisingly, the most powerful states are better equipped to navigate this landscape. As Pratt (2018) argues, even with the lack of a formal hierarchical structure, hierarchies still emerge. Institutions often defer authority to one another, and this deference tends to reflect underlying power dynamics. Powerful states with broad institutional access and resources gradually end up influencing governance, as others start relying or copying their standards. In AI governance, this becomes clear when analyzing how the EU, the U.S. and China have started projecting their preferred models across the international community through diffusion (Bradford 2020; Bradford 2023). Nonetheless, this hierarchy is also shaped by private actors. In AI, corporations control many of the resources necessary for governance, and as such powerful states benefit from their proximity to dominant technology companies, while these firms gain access to public decision making. This reveals a deeper political economy where public policy can become more aligned with the interests of private actors rather than accountable governments.

Disagreement in governance goes deeper than compliance but questions the underlying meanings behind such rules (Wendt,1992). In AI, there is no consensus on what the system should be; some actors have a preference for safety and rights whereas others for competitiveness. Additionally, actors can tailor their discourse depending on the strategic context, and private companies may invoke rights-based arguments, such as privacy or safety, when these also further commercial interests. Without a minimum convergence on this matter, governance will remain largely unstable, as agreements on rule are near impossible when actors have different definitions for what these rules mean.

Timelines and the scale of potential impacts remain unclear, further contributing to the uncertainty. This is where anticipatory governance becomes critical. Anticipatory governance is referred to as an approach based in the exploration, anticipation, and preparation for change and uncertainty (UNESCO, 2023; Muggah, 2025). Within AI governance, anticipatory measures are aimed at identifying and mitigating potential risks before they materialise (UNESCO, 2025). By addressing risks early in the AI lifecycle, it can be ensured that systems are employed with safeguards in place which minimise harms for individuals and society, which in turn fosters greater public trust and accountability.

Applied to AI, the question becomes not whether current institutions are equipped to govern, but whether they will be equipped to deal with systems that will emerge in five or ten years. Current evidence suggests that anticipatory AI governance remains limited, as AI safety work often relies on retrospective model testing rather than prospective risk assessment, which can leave emerging or high-priority risks insufficiently addressed (International AI Safety Report, 2025). In a governance landscape defined by a deep trust deficit and the lack of binding coordination mechanisms, informal diplomacy mechanisms present a promising pathway for AI governance, especially as formal institutions are constrained in ways that prevent them from addressing the pressing issues in an effective way. Informal diplomacy can be defined as the convening of diplomatic activities through non-official channels, with more flexible and less structured methods than traditional diplomacy (DiploFoundation, 2021a). The report also refers to “multistakeholder governance" to emphasise the act of “regulating the global commons” (The Stanley Foundation, 2016, p.2) beyond state-to-state or bilateral agreements. While, in theory, all governance can be understood as being negotiated among multiple actors, the term highlights cases where the authority and regulatory power of states is shared with non-state actors, which

are involved in the negotiations of rule-making and coordination (Vallejo & Hauselmann, 2004; The Stanley Foundation, 2016). Informal diplomacy may also facilitate the inclusion of a wider range of actors including experts and voices from the Global South, making them a place where power imbalances can potentially be questioned.

3.2.2 Assessing Regime Quality

When analysing the climate change regime complex, Keohane and Victor (2011, p. 17) established five criteria to evaluate regime complexes. In order to assess efficient policy implications, it is useful to apply these to AI governance.

Criteria	Definition (Keohane and Victor 2011)	Application to AI governance
Coherence	A regime complex should have a coherent set of rules.	The current AI governance landscape includes several institutions promoting diverging regulatory approaches, including risk-based governance and innovation-oriented models. This creates tension rather than alignment.
Accountability	A regime complex should be accountable to its audience, including state and non-state actors.	Many AI governance arrangements occur through bilateral agreements between private firms and national governments. These often operate with limited transparency and weak accountability to wider affected audiences.
Determinacy	A regime complex should have rules whose meaning is clearly defined.	AI governance rules often remain unclear and lack consistent normative content. Key concepts such as "risk" and "safety" are interpreted differently by different actors, making long-term coordination difficult.
Sustainability	A regime complex should be stable and not easily disrupted by small shocks.	Ongoing geopolitical competition and rapid technological change make AI governance arrangements highly fluid and unstable. This creates a weak equilibrium among actors and regulatory approaches.
Epistemic quality	A regime complex's rules should be aligned with scientific knowledge.	Expert communities play a crucial role in AI governance. However, rapid technological development and the concentration of expertise within a small, often private-sector-based community create barriers for academia and civil society, raising concerns about the robustness of knowledge integration.

Table 1. AI Regime Complex

3.3 Non-state Actors and the Role of Informal Diplomacy

The discourse on AI governance reveals a fundamental tension between the objectives and methods of private corporations and those of civil society actors. Private actors usually prioritise commercial advantage, innovation and market leadership; while civil actors usually rely on public interest, fairness and inclusivity. This dichotomy leads to an opposite approach to regulation, with private actors preferring self-regulation, "soft law," and voluntary guidelines that allow for flexibility and minimise liability; and civil actors demanding for robust oversight, transparency and public participation in regulation (Taeihagh, 2021).

3.3.1 Civil Society

Civil society organisations (CSOs), which include NGOs, independent research and higher education institutions, advocates and other autonomous actors, play an important role as a counterweight to state and corporate priorities. One of civil society's main strengths lies precisely in its diversity, as these actors represent a wide range of communities, expertise and perspectives. At the same time, this heterogeneity can make it difficult for civil society to speak with a single voice or reach consensus on governance priorities, particularly compared to the more concentrated influence of powerful corporate actors. Nevertheless, CSOs continue to voice concerns of affected communities, bring attention to overlooked risks such as discrimination and environmental degradation, and provide valuable technical, legal and human rights expertise (Lenoir et al., 2025).

However, despite its value and necessity, civil society also remains the least powerful actor in practice, as systemic obstacles hinder substantive participation, including limited financial capacity, technical complexity, gaps in expertise and a lack of a cohesive civil society voice. While Civil Society from the Global Majority has shown a clear willingness to participate in AI governance, without a collective effort to break down systemic barriers, it will not be able to offer its valuable perspective in discussions on governance (Gálvez Callirgos, 2025). Despite these challenges, civil society is already making a tangible impact through multiple entry points: producing policy research, developing practical tools for ethical AI, conducting independent audits, building collaborative networks, and engaging directly in policy consultations. To fully unlock this potential, a set of actionable recommendations has been proposed for CSOs, policymakers in the Global Majority and major powers, and international organizations. These include institutionalizing meaningful consultation, establishing dedicated funding mechanisms, strengthening technical capacity, and creating formal seats for Global Majority CSOs within international governance bodies (Gálvez Callirgos, 2025).

Building on these proposals, three fundamental reforms stand out as particularly important for empowering civil society: first, the implementation of open and transparent selection processes for participation in expert groups; second, the introduction of a "duty to respond," ensuring that contributions receive formal feedback; and third, the creation of dedicated funding mechanisms to enable independent and sustainable engagement (Lenoir et al, 2025).

3.3.2 Private Sector

The private sector is presented as both an indispensable technical partner and a contested diplomatic actor for AI risks. Policy and industry literature generally converge in the belief that neither governments nor companies can address governance gaps alone. However, the key drivers of AI development tend to be companies that have extensive resources and technology capabilities also in other areas, such as computing and data access (Koenig, 2025). This has led to a concentration of economic power and policy influence among a small number of big technology companies, and this trend, occurring beyond the sphere of states, reasserts AI as a broader political issue. Additionally, the rapid developments of AI technologies require an adaptable governance framework, which can keep pace with the developments, and ensure these occur in a safe and ethical setting. When regulation lags behind AI developments, it leaves a governance vacancy which is oftentimes filled by a private company's decisions (Alais, 2026).

While AI development generates risks that private firms cannot adequately govern alone (Hendrycks et al., 2023), Alais (2026) argues that regulatory weakness can allow such firms to become de facto AI governors. The early 2026 standoff between the U.S. Department of War and Anthropic illustrates the ambiguous role of the private sector in AI governance. Anthropic's decision to retain safeguards against fully autonomous weapons and mass domestic surveillance may appear as a responsible act of corporate risk mitigation. However, it also exposes a governance failure, where protection against catastrophic AI risks depended on discretionary judgement of a private actor, rather than enforceable public regulation. As Alais warns, "Rights that depend on the goodwill of a technology executive […] that could be withdrawn tomorrow, are not rights" (2026).

Anthropic itself has acknowledged this gap, in October 2024 calling for government action within eighteen months and warning that the "window for proactive risk prevention is closing fast" (Anthropic, 2024). The, since updated, Responsible Scaling Policy, introduced in 2023 establishes an adaptive "if–then" framework in which safety obligations increase with capability thresholds, presented as a prototype for "judicious, narrowly-targeted regulation" (Anthropic, 2024; 2026). Anthropic's standpoint, which is explicitly aimed at identifying and mitigating "catastrophic risks", aligns with broader calls for public–private cooperation. CIGI argues that joint governance is necessary because firms often possess a better understanding of AI tools and the unintended consequences of regulation (Medeiros, 2020). Similarly, the World Economic Forum frames public–private partnerships as a way to address the global "AI trust deficit" by combining "government legitimacy, industry capability and civic oversight" (Sharma, 2025). However, this risks reducing trust to a problem of technical coordination. While firms may possess technical expertise and move faster than states, and PPPs may improve implementation, assurance, and inclusion, they do not resolve a deeper problem of accountability and transparency.

As firms lack democratic authority and remain commercially tied to the systems they aim to govern, the risk of corporate capture also motivates Alais's (2026) concern about the relocation of governance itself. It is highlighted that crucial decisions on AI deployment and regulation are increasingly relocated from multilateral institutions. Instead, "governance by procurement" (Alais, 2026) moves decisions to bilateral agreements between governments and the same companies whose power, incentives, and systems require governing. The structure is framed as unevenly exclusionary, as it marginalises the agency of those affected by its outcomes, and excludes large parts of the world from the conversations in which terms are set. As countries across the Global South and Europe build their digital futures while acting as tenants on someone else's digital infrastructure which they cannot control. The resulting fragile and discretionary form of governance reinforces the need for stronger and inclusive international governance.

Concerns about the legitimacy, accountability, and effectiveness of involving the private sector in AI governance are also raised in a review of UNESCO's collaboration with the private sector through the Business Council for the Ethics of AI (Hatano, 2026). Deficits in representation, binding mechanisms, and transparency were identified in implementation of the human-rights based approach presented by UNESCO in 2021 through corporate participation in capacity building, policy guidance and norm-setting. Nonetheless, it is noted that the incorporation of corporate expertise in the development and implementation of international AI governance instruments holds considerable value for enhancing technical relevance, accelerating the uptake of ethical norms and fostering policy innovation. These advantages may directly improve the capacity of governance to keep up with AI developments, highlighted as crucial by Alais (2026), and the review further proposes that the credibility of UNESCO's model could be enhanced through mandatory transparency, broader stakeholder representation and independent oversight.

This literature implies that diplomacy should neither romanticise corporate self-governance and bilateral state-firm agreements nor ignore firms' crucial technical capacities. Crucially, the value of AI governance is significantly undermined when it is unable to anticipate the needs for risk-mitigation, leaving room for private actors to decide independently. Hybrid governance or PPPs are shown to have both transformative potential and significant structural risk, and ensuring accountability and transparency therefore become crucial to mitigate corporate capture.

3.4 Regional Case Study

3.4.1 Case Selection

This study focuses on Latin American countries, as part of the Global South actors which remain significantly underrepresented in current AI governance scholarship. Current policy debates and the majority of the existing literature focuses disproportionately on contexts where regulatory capacity and institutional visibility are more advanced, particularly in the United States, Europe and China. Consequently, the conditions under which AI governance emerges in the Global South remains insufficiently explored. Additionally, Latin America provides a rich landscape of institutional diversity, developmental asymmetries and geopolitical dynamics that are critical for understanding how AI governance is unfolding beyond the Global North. To that end, an investigation of four Latin American countries (Brazil, Argentina, Colombia and Uruguay) has been developed, which is compared to African Union member states due to their contrasting institutional architectures and levels of policy coordination.

The countries selected were chosen based on the relative maturity of AI strategies regionally and the availability of up-to-date scholarship. Brazil, Argentina, Colombia and Uruguay currently exhibit the highest levels of AI policy development, digital governance capacity and institutional readiness within Latin America (ILIA, 2025). Examining these countries together allows us to analyse how AI governance emerges in the absence of a regional framework at the Mercosur level, and how national initiatives compensate for, or reproduce, this structural gap. The study aims to understand how the local and geopolitical contexts of Latin American countries impact the current and potential use of informal diplomacy mechanisms to mitigate AI risk. In contrast, the African Union has developed an emerging continental-level framework for AI governance, including initiatives on data policy.

The study adopts a comparative design based on diverse case selection and structured comparison, enabling the systematic analysis of how conditions for informal diplomatic cooperation vary across governance contexts.

3.4.2 Latin America

Vivares et al. (2025) map the main challenges faced by the region. They argue that AI is fundamentally a political-economic issue shaped by long-standing patterns of dependence and unequal insertion into the global order. The authors show that AI governance is driven less by internal strategic choices and more by external pressures stemming from Global North models. AI adoption in the region remains concentrated in a few niches of the private sector, such as fintech, marketing, and logistics, while public institutions struggle with limited digital infrastructure, low investment, weak technical capacity, and a strong dependence on extra-regional cloud providers. Although the region generates vast amounts of data, it lacks local infrastructure and R&D ecosystems capable of capturing value. As a result, most of the economic and technological gains from AI development flow to foreign companies, reinforcing existing center-periphery hierarchies.

At the same time, AI use is expanding across the region, particularly in public administration, health and digital public services. However, the literature underscores that this development remains hindered by foundational weaknesses. Sanchez-Pi et al. (2021) identify weak data governance, lack of infrastructure for computational power, low supply of qualified practitioners, many countries working on creating regulations but not sufficiently connecting them across borders and a high degree of public distrust towards technology.

The Roadmap provides an overview of how the current state of AI impacts the way that countries in Latin America are able to utilize AI solutions. The authors also describe the major differences between those countries with established digital ecosystems in place versus those just beginning to create their own policies, and discuss how much reliance on third-party technologies and data from outside of the region negatively impacts bias, data sovereignty and to what degree Latin America controls its own digital future.

The Roadmap, moreover, outlines governance limitations produced by these constraints and provides recommendations to address them from an anticipatory governance perspective. These include enhancing preparedness by increasing institutional capability, developing Good Practice Criteria for Responsible and Explainable AI, creating open-data ecosystems, establishing regulatory sandboxes and implementing engineering techniques such as MLOps to allow continued monitoring and adjustment over time. These recommendations align with the value of tools developed to address both long-term and catastrophic risk management, as identified through the ARP project.

Furthermore, Contreras (2024) argues that Latin America should not merely regulate AI reactively but should also build the institutional capacity to foresee and mitigate future risks associated with advanced AI technology. The need for ethical awareness during development, inclusive governance processes, and ongoing monitoring is critical, while noting that regional regulatory fragmentation and institutional asymmetries hinder preparedness. This is evident across the four countries in our study: Brazil’s comparatively strong frameworks, Colombia’s notable progress, Argentina’s uneven yet developing governance and Uruguay’s early leadership in digital public services illustrate different levels of readiness. The literature describes that the current state of AI governance in Latin America has been, and continues to be shaped, through the existence of differing levels of capacity within the region, and that it is essential to take a cooperative and anticipatory approach to find paths forward for the Latin American region in association to the development of new global governance frameworks for AI.

Country	Maturity level (ILIA, 2025)	AI Governance Framework	Key Strengths	Main Constraints	Analytical Value
Colombia	Adopter (intermediate maturity).	National AI Policy (transparency, accountability, rights; risk assessments).	Strong policy development; progress in human talent (notably self-directed learning); improving infrastructure and data capacity.	Limited institutional capacity; fragmented governance; insufficient investment; weak long-term ecosystem integration.	Illustrates gap between normative governance ambition and implementation capacity in mid-tier ecosystems.
Argentina	Adopter (intermediate maturity).	2023 AI Recommendations (human oversight, transparency, non-discrimination, accountability) + subnational initiatives.	Clear ethical governance principles; emerging multi-level governance (Buenos Aires).	Limited implementation; structural economic constraints; fragmented coordination across levels.	Demonstrates normative advancement without equivalent execution capacity.
Brazil	Pioneer (high maturity).	Alignment with global frameworks (e.g. UNESCO AI Ethics); national research and policy ecosystem.	Strong research and innovation ecosystem; high infrastructure capacity; leadership in regional AI development; significant computing power concentration.	Regional inequality; weak coordination between actors; uneven adoption; risks of exclusion and bias.	Shows high-capacity ecosystem with internal asymmetries and coordination challenges.
Uruguay	Pioneer (high maturity).	Strong data protection regime; digital government and transparency frameworks.	Robust institutional capacity; high infrastructure performance; strong governance and public-sector AI adoption; digital inclusion.	Lack of evaluation metrics; limited private-sector governance reach; emerging risks insufficiently regulated.	Represents high-quality governance with remaining monitoring and scaling gaps.

Table 2. AI in Latin American countries (more information in Annex 3)

4. Methodology and Research Design

4.1 Analytical Framework

This research is anchored by a framework that distinguishes between conditions, mechanisms and outcomes regarding the effectiveness of informal diplomatic processes in anticipatory governance of AI catastrophic risks. While existing literature focuses on foresight tools and the importance of multilateral cooperation, it remains unclear under what conditions such cooperation can occur in practice. For the purposes of this research, five dimensions are highlighted.

First, this study defines the structural factors that determine whether cooperation emerges among states. In this context, the following conditions are evaluated: (1) trust among states, (2) institutional capacity, (3) political prioritisation, (4) incentive structures and (5) alignment of risk perception (Risse, 2000). These conditions are anchored in Risse's (2000) preconditions for argumentative rationality, for instance, trust and risk perception among states is justified as a necessary condition especially in validating authenticity which serves as a precondition for cooperation. Moreover, informal diplomacy falls under non-hierarchical network-like institutions thus institutional capacity is a key condition in facilitating discursive and argumentative processes between states (Risse, 2000). Political prioritisation and incentive structures are supported by the logic of consequentialism, as informal governance arrangements lack sovereign authority and strong enforcement mechanisms. Under such conditions, actors are more likely to cooperate when they perceive strategic or political benefits in doing so. At the same time, persuasion and discursive processes remain important for shaping shared understandings of risk and encouraging voluntary coordination among actors (Risse, 2000).

Second, mechanisms of informal diplomacy are mapped as cooperation practices that can potentially emerge in practice. These mechanisms will mainly be presented as recommendations in section 6. As previously established, due to the absence of binding agreements, informal diplomacy is crucial in enabling coordination among states. Therefore, it is assumed that these factors create favourable outcomes, which in turn can shape levels of anticipatory governance capacity. However, it is important to stress that this relationship is not always linear. Informal mechanisms may be able to partially compensate for weak institutions, but the overall effectiveness is still dependent on institutional and system capacities.

By applying this framework to the comparative cases of the African Union and Latin America, this study aims to identify how different governance contexts influence the emergence and effectiveness of informal diplomacy in addressing catastrophic AI risks.

4.2 Research Design

This research adopts a mixed-methods design to identify the conditions under which informal diplomacy can mitigate potential catastrophic risks arising from advances in AI. These include expert surveys and interviews combined with desk research. For more information on how excerpts were selected and the processes of surveys and interviews that were carried out, please refer to annex 5. Thus, the aim is to identify mechanisms of anticipatory governance, including early warning systems, cross-actor coordination and preparedness capacities. The analysis combines survey data, semi-structured elite interviews, as well as secondary literature to capture the underlying patterns and governance mechanisms.

This research adopts a mixed-methods design to identify the conditions under which informal diplomacy can mitigate potential catastrophic risks arising from advances in AI. These include expert surveys and interviews combined with desk research. For more information on how excerpts were selected and the processes of surveys and interviews that were carried out, please refer to annex 5. Thus, the aim is to identify mechanisms of anticipatory governance, including early warning systems, cross-actor coordination and preparedness capacities. The analysis combines survey data, semi-structured elite interviews, as well as secondary literature to capture the underlying patterns and governance mechanisms.

Rather than establishing causal relationships, the aim of this study is to identify the conditions and mechanisms that can potentially determine the effectiveness of informal diplomatic mechanisms. The outcomes are presented as potential positive effects these mechanisms may have in the anticaptory governance space. Given its focus on anticipatory governance, employing a CMO approach is particularly well-suited to risk-prone environments. The framework will be evaluated through qualitative indicators derived from interviews, survey responses, policy documents and institutional analysis, which will be discussed in detail in the findings section. Trust will be assessed through evidence of sustained cooperation and information-sharing, while institutional capacity will be evaluated through governance structures, expertise and coordination mechanisms. Political prioritization will be measured through policy attention and international engagement, and alignment of risk perception through shared understandings of AI risks and governance priorities. Informal diplomatic mechanisms will be analysed through evidence of recurring coordination and collaborative initiatives, while outcomes will be assessed in terms of anticipatory governance capacity.

4.3 Limitations

This study encounters several limitations that shape the scope and reliability of its findings. First, as AI governance continues to evolve in many developing regions of the world, including Latin America and Africa, continuous updates of the digital strategy and policy regulatory frameworks by the governments could create the risk of outdated analysis. The limited availability of academic literature on the role of the research community in AI governance in these regions constrains the depth of analysis on this dimension. Subsequently, due to the rapid pace at which technology advances as well as changing regulations, the long-term stability of certain conclusions might be fragile. Second, the access granted to senior policymakers, private-sector representatives and specialised experts can vary significantly among countries. The differences in the accessibility of elite interviewees may introduce selection biases, since the individuals willing or able to participate in elite interviews could be unrepresentative.. This challenge was particularly evident in the African context, where significant difficulties were encountered in identifying and securing participation from relevant experts: in several cases, potential respondents were either not reachable or did not respond to interview requests, further limiting the diversity and representativeness of perspectives from the region. Moreover, even among those who initially agreed to participate in scheduled interviews, some ultimately did not take place due to last-minute cancellations. Third, it is important to note that expert interviews present a certain level of risk regarding subjective interpretation, positional biases, and/or political sensitivities, which will affect whether or how information was presented and/or withheld.

Additionally, a significant proportion of survey respondents did not answer open-ended questions, which resulted in limited qualitative depth and reduced the amount of usable explanatory data. Furthermore, the research encountered difficulties in identifying and recruiting women working in the sector, which led to a gender imbalance among respondents, and a potential bias towards male perspectives in the findings.

This study addresses the identified limitations through the use of strict triangulation between interviews, policy documents and academic literature; ongoing monitoring of newly created policies and the declaration of the areas which contain uncertainty or deficiencies in supporting evidence. While triangulation strengthens the robustness and credibility of the findings, it does not fully eliminate underlying biases related to the selection of interviewees, the availability and scope of literature, or gaps in accessible data. Rather, it serves as a mitigating strategy that improves, but does not fully resolve, these structural limitations. Collectively, these approaches allow for an in-depth and thorough analysis of a continuously changing and diverse field while acknowledging the limitations for conducting research in this context.

The project remains conscious of its methodological boundaries. It does not attempt to forecast AGI development or quantify catastrophic risks through modelling. Nor does it aim to produce an exhaustive comparative study of AI governance efforts across Africa and Latin America. Additionally, the paper focuses on a selection of the most digitally advanced countries in Latin America. While this constitutes a justified sampling strategy, it limits the generalisability of the findings, which may not fully extend to other countries in the region with different levels of digital development, such as, for example, Bolivia, Peru, or Ecuador.

5. Findings

5.1. Summary of Key Stakeholders and Governance Capacity

The following section presents the background of consulted experts and their perceptions of the current AI governance landscape, as identified through the survey and interviews. Aligning with the CMO framework, it also identifies the conditions of current cooperation on AI governance, specifically the distribution of trust, institutional capacity and actor alignment.

The background of the 17 survey respondents within the general survey included technical experts, staff of intergovernmental organizations, diplomats, policymakers, private sector actors, academics and civil society representatives. 12 respondents situated themselves in advisory and consultative positions, with the rest identifying themselves as contributing to drafting and implementation, final decision makers and technology ambassadors. When asked to rank their primary region of work's current anticipatory governance capacity, on a scale from 1 to 10, where 10 was considered the highest capacity, responses ranged from 3 to 8, with an average ranking of 5.41. Compared to the total average, the ranking by experts from Latin America and Africa was slightly below, at 5.18, and diplomats' ranking was slightly above other types of actors, at 5.83. However, in general the rankings did not show major discrepancies across different actor characteristics; see details of this ranking in *figure 1* below.

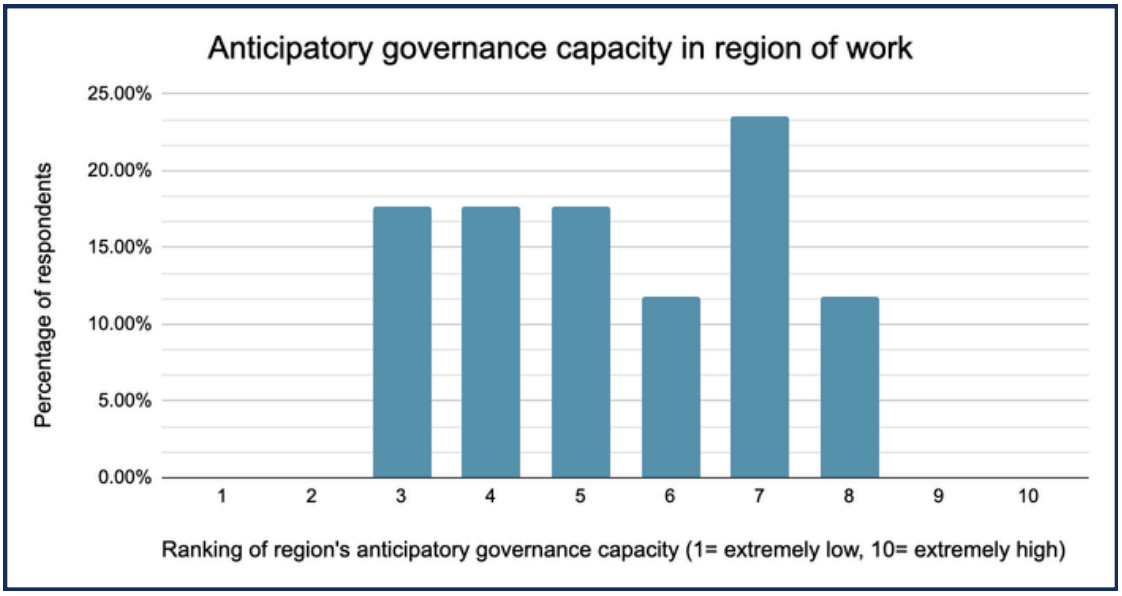


Figure 1: Ranking of anticipatory governance capacity

Experts were also asked about their perceptions of the attributes of types of key actors in AI governance. *Figure 2* illustrates the average ranking of actors on categories of trustworthiness, willingness to share safety-relevant information, alignment with public interest, perceived interest in collaboration and proactivity. Findings were consistent with the general sentiment of skepticism towards the role of the private sector identified across the expert consultations. The academic and research community was ranked highest, with an average ranking of 3.67 out of 5, followed by civil society organisations, national government agencies, and regional institutions. The private sector was ranked lowest, with a total average of 2.54, corresponding to a 30.7% lower score than the academic and research community. This ranking may have been a result of selection bias, as the private sector itself is considerably underrepresented in the group of consulted experts. However, while actors within civil society, multilateral institutions and government agencies ranked themselves an average of 0.4 points higher than the total average rating, the difference within the rating of the private sector was not statistically significant. In addition, the academic and research community had the same number of consultations as the private sector and had ranked itself lower than their overall score, therefore suggesting that the data is still applicable despite potential selection bias.

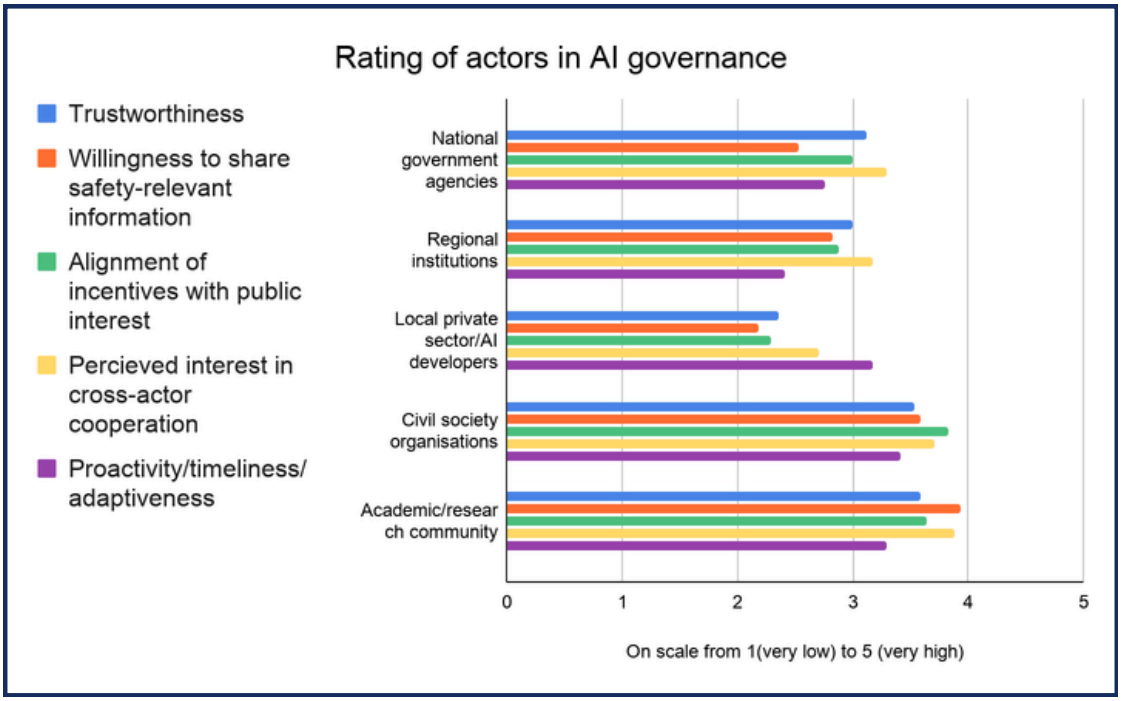


Figure 2: Rating of different attributes of actors in AI governance on a scale from 1 to 5

In the interviews, skepticism towards the private sector was connected to the concentration of private sector actors and a small number of powerful governments in AI development, as opposed to it being shaped by inclusive multilateral processes. One expert claimed that if a company is building AI, they are also building the governance of AI meaning that there are currently no accountability measures in place. Concerns over the private sector concentration were further amplified by the attitudes of the private sector, whereby companies such as OpenAI and Anthropic are unlikely to share details of their models due to commercial incentives. A lack of transparency regarding these AI systems is further understood as a catastrophic risk because societies are becoming increasingly dependent on AI systems that they do not understand or govern. These findings highlight that the current conditions for cooperation within anticipatory governance of AI are low, particularly towards the private sector. Alternatively, non-state state actors such as academic/research communities and civil societies organizations were ranked the highest regarding aspects of trustworthiness and willingness to share safety-relevant information which illustrates that within the current context of AI governance, these non-state actors are perceived to be more trustworthy than the private sector and even national governments.

5.2 Perceptions on Catastrophic risks

This section is dedicated to experts' perceived threat level of catastrophic risks, which are identified as the most severe, and the current governance capacity to manage these threats. It further examines the alignment of risk perception as a condition for cooperation, assessing the extent to which shared understandings of catastrophic AI risks exist across different actors and regions. Risk perception functions as a prior condition in the CMO framework; as such, this section does not identify mechanisms or outcomes. Mechanisms will be outlined in section 5.3. Moreover, a focus on the conditions pillar in this section maps a diverse understanding of risks to help structure mechanisms for cooperation in the form of recommendations in section 6

Across the survey responses, perceptions of catastrophic risks were high, as illustrated by *Figure 3*. For instance, when presented with a 1-10 scale (10 being extremely worried) regarding the respondents' worries about catastrophic risks of AI, 82% of respondents rated their concern at six or above. This suggests that even within this small expert-based sample, catastrophic risks of AI emerge as a dominant pattern instead of a marginal concern. Across the interviews, however, AI is rarely described as a danger in itself. Instead, catastrophic risks are identified as a result of interactions between AI and weak governance, geopolitical competition, private-sector concentration and technological capability. Within interviews, AI became a catastrophic risk when combined with contexts such as weak governance systems, military competition and private sector concentration. This is further supported by survey findings whereby, 47% of experts referred to institutional weakness or lack of governance capacity in the reasoning for their ranking. When asked to expound on this through the interviews, experts linked catastrophic risks to the failure of governments and multilateral institutions to keep up with private sector development. One interviewee underscored that large technology companies are not only building AI systems but also creating the governance environment around them, especially because key decisions are being concentrated within these companies located in the United States and China as opposed to multilateral spaces. This sentiment was shared by another interviewee who warned that AI models continue to remain black boxes controlled by private companies that are unlikely to share their technical knowledge due to profit incentives. Other frequent themes in their reasoning related to the lack of control, risk of malicious use, the relative power of the private sector and profit incentives, and uncertainty, especially regarding the long-term risks AI may pose. Interestingly, 20% of respondents, all of whom were diplomats, highlighted their belief in governance capacity to manage these risks as a reason for lowering their ranking, which was consistently below the average.

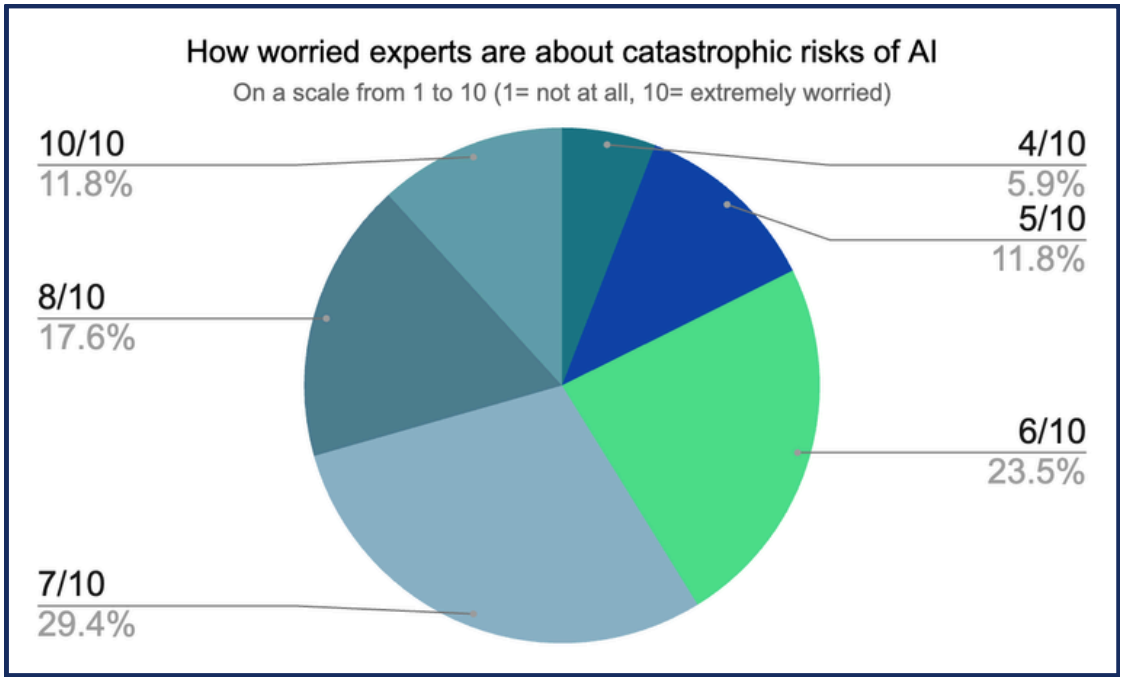


Figure 3: Graph illustrating how worried experts ranked themselves about the catastrophic risks of AI developments on a scale from 1 to 10

In interviews, the issue of lack of control is understood as a loss of human and democratic control over the direction of AI development. This was linked back to the current private sector concentration. Experts expressed concern over current governance initiatives, arguing that they are only effective when they regulate AI once it is operational, while processes of development remain opaque and in the control of these few technology companies. Additionally, when asked about human control, one interviewee argued that human oversight is often linked to embedding ethical frameworks into AI. This is not effective, as AI systems do not respect these frameworks and do not hold human capabilities such as empathy to understand the values of these frameworks. Thus, lack of control within these interviews was understood beyond technical failure of AI models but also as a political failure to retain public authority over AI models development.

The survey additionally asked experts to rank risks which were most frequently referenced in review literature on impact severity, probability of occurrence in the coming five years in their region of work, and current governance capacity to manage. Figure 4 illustrates the ranking of the risk categories: large-scale mis- and disinformation, loss-of-control or malfunctioning, biosecurity threat enablement, critical infrastructure enablement, military and warfare utilisation, escalation of interstate animosity, and socioeconomic disruption.

The impact severity of military and warfare utilisation was ranked as clearly highest, with interviewees referring to the increased unpredictability and deceptiveness that arises when advanced AI systems are used for military and malicious purposes. One expert illustrated this by explaining a recent simulation where AI models such as Claude and ChatGPT were placed in a nuclear war simulation and none of the models chose to descalate in each scenario presented during the simulation, highlighting the danger of employing AI within military contexts without adequate safeguards. However, even for the lowest ranked risk, loss of control or malfunctioning, 59% of respondents ranked impact severity as high or very high. For experts, malfunctioning was ranked with a high impact severity due to the unpredictable or deceptive system behaviour as explained in the interview. This behaviour was also linked to the lack of governance systems in shaping collective decision making. This is such that, when military actors adopt AI faster than oversight mechanisms or also when states depend on foreign AI infrastructure for military use, this is perceived as an issue of malfunctioning.

When combining rankings of impact severity and probability, military and warfare utilisation remains highest, followed by cyber disruption of critical infrastructure, large-scale mis- and disinformation and socioeconomic disruption (ranked equally), escalation of interstate animosity, loss of control or malfunctioning, and lastly biosecurity threat enablement.

When additionally combined with the current capacity to manage, the order of the risk ranking stays largely the same. However, the equally ranked categories of large-scale mis- and disinformation and socioeconomic disruption moves to second place instead of third, illustrating a discrepancy between the capacity to manage risks that can be considered more societal and those considered more technical.

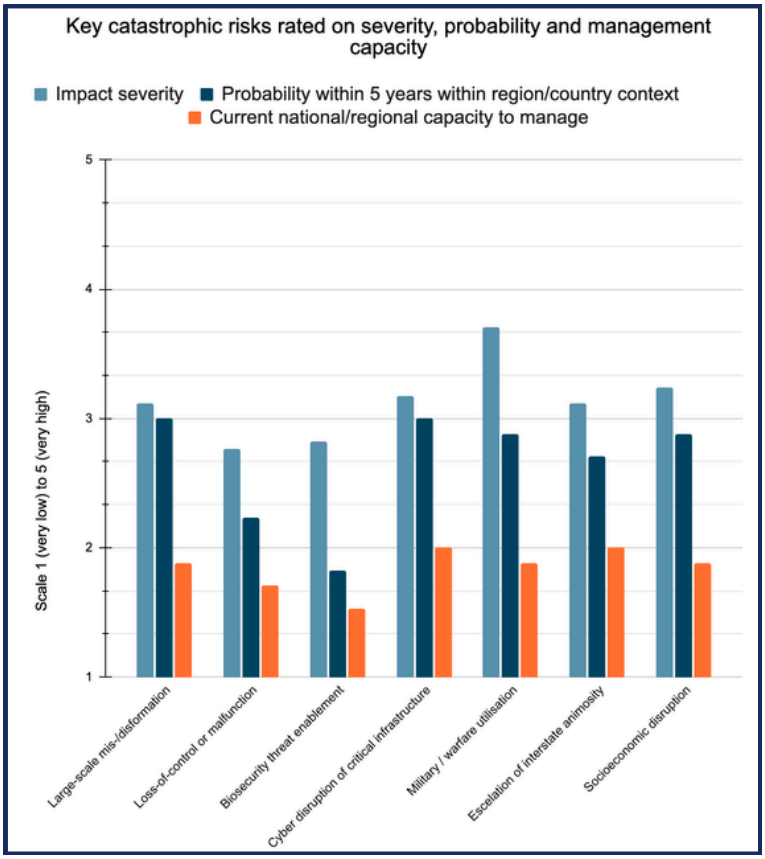


Figure 4: Risk ranking on impact severity, probability and management capacity

This further aligns with responses to what risk was most overlooked in their regional context, where information integrity was the risk most frequently mentioned. One interviewee further argues that societal risks seem more governable in theory but are much harder to manage in practice. Whereby labour disruption and misinformation appear less technically complex compared to other risks such as loss of control, they are deeply embedded in social behaviour and institutions. When considering this with the uneven capacity for states to regulate AI as explained above, societal risks and attempts to address them for instance through public education on AI and labour markets anticipation is more difficult to pursue.

Both survey and interview responses highlight that the current conditions of catastrophic risks perception of AI are mainly understood in the contexts of private sector concentration, military escalation, loss of control, socioeconomic disruption and weak governance. Importantly, these risks do not function individually but rather when combined, they are perceived as catastrophic, for instance current governance structures around AI are not effective due to development of these models being concentrated in a few big technology companies.

5.3 Hindering and Enabling Conditions and Mechanisms of Cooperation

Section 5.3 begins by presenting findings on what conditions incentivise and promote cooperation and which hinder cooperation on AI governance. In line with the CMO framework, having outlined the relevant conditions in the above section, the section focuses on the mechanisms pillar to deduce which features of coordination mechanisms are considered most important, and which potential mechanisms are most versus least feasible in the near future. Additionally, the section examines the conditions that enable or hinder cooperation and identifies which mechanisms of informal diplomacy would be most viable given current conditions.

An overarching theme within the survey is that governments and key governance structures are struggling to keep up with the speed of AI development, whereby, when asked which resource gap is most limiting to anticipatory governance capacity in their context, respondents most frequently chose the issue of political prioritisation. This was closely followed by the lack of legal authority and mandate, and funding issues. When further asked about the key barriers to cooperation, 67% of responses mentioned challenges in interest alignment, especially between the private sector and governments. Further barriers repeatedly mentioned include the speed of AI advancements, power play amongst stakeholders, disagreements between political parties and lack of technical knowledge and shared understandings of key concepts and risks.

Challenges in alignment appear closely linked to lack of trust between actors. When asked about how trust could be improved, experts most frequently chose public transparency requirements, followed by shared technical standards and an independent oversight body (Figure 6). However, for coordination mechanisms such as these to be effective, political buy-in arose as the most important feature, followed by legitimacy and inclusivity and authority and legal mandates (Figure 7).

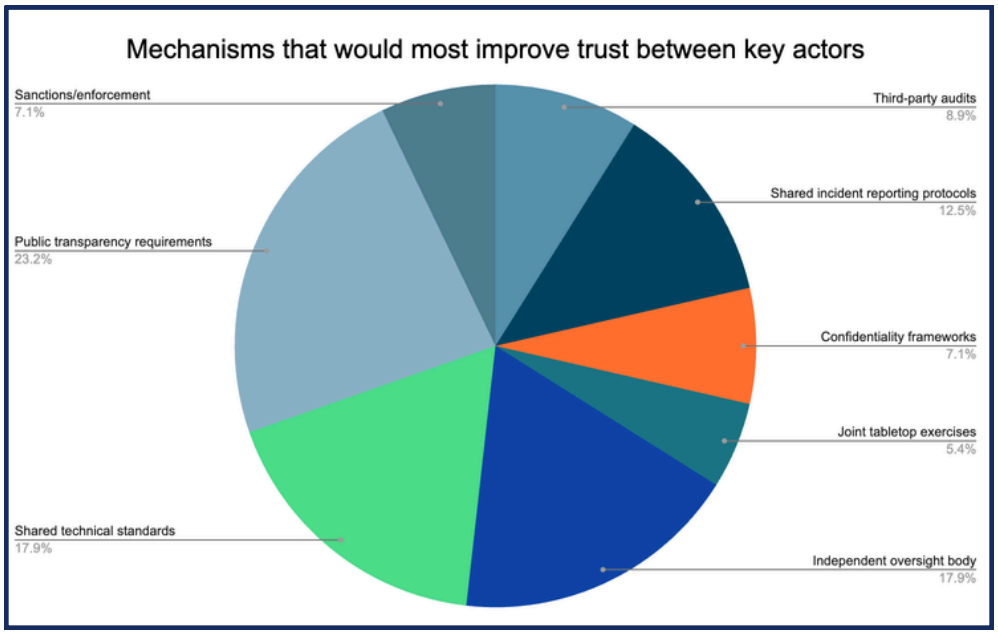


Figure 6: Most trust improving mechanisms

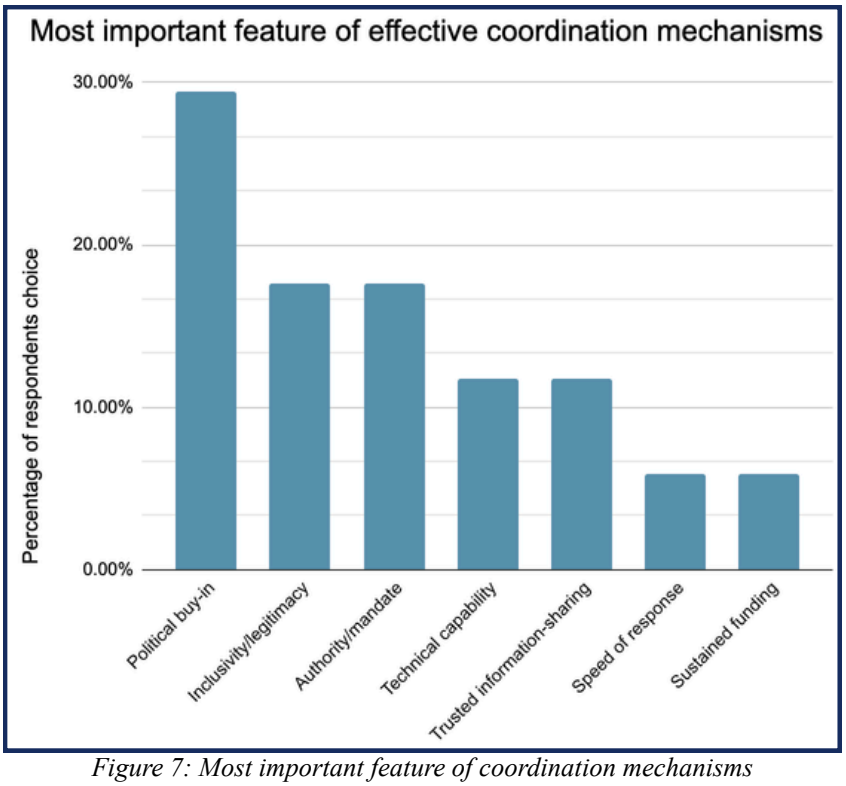


Figure 7: Most important feature of coordination mechanisms

With these conditions in mind, experts identified the most feasible areas for cooperation as being capacity building and incident reporting and shared risk taxonomies (figure 8). In contrast, technical cooperation mechanisms such as sharing advanced model details and compute access were identified as least feasible (figure 9). This may suggest that, while actors are interested in cooperation to mitigate risks, competitive concerns of both states and private actors limit options within technical cooperation. During interviews, experts furthermore called for a centralisation of processes on the global level, and a dedicated United Nations body was suggested recurrently. One expert proposed the idea of a "digital commons for humanity", which would drive the sharing of digital resources and infrastructure for global public benefit.

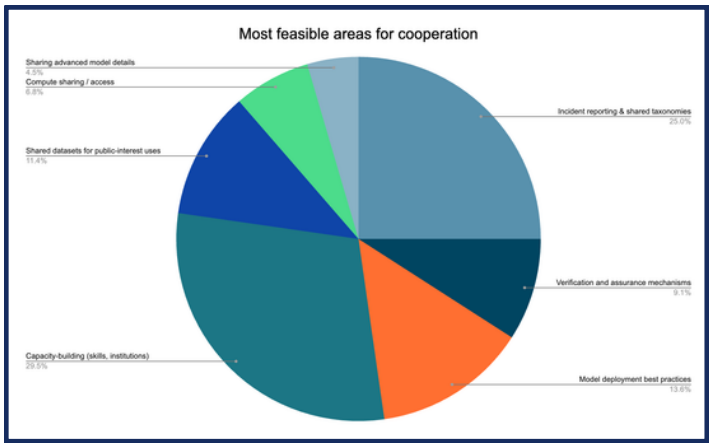


Figure 8: Most feasible areas for cooperation in AI risk mitigation

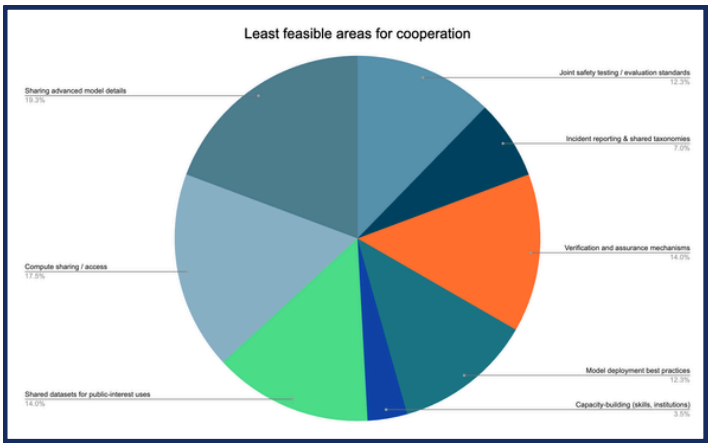


Figure 9: Least feasible areas for cooperation in AI risk mitigation

Across the interviews, a central pattern emerges: a growing gap between the speed and influence of AI development and the capacity of governance systems to respond effectively. Respondents highlight the significant role of private actors in shaping governance outcomes. This occurs either through direct influence over the design of AI systems, as emphasized in Europe,. At the same time, governance efforts are weakened by fragmentation, whereby in Europe there is a persistent disconnect between principles and implementation. This is such that, although ethical concerns such as human rights and transparency are widely recognized, translating them into concrete regulatory tools and enforceable standards remains a major challenge. The gap between principle and implementation highlights that current enabling conditions within AI governance are insufficient to activate mechanisms that would strengthen anticipatory governance especially taking into account the rapid development of AI technologies.

5.4. Case Study: Latin American countries

The following section, 5.4, presents findings particular to the case of Latin America. Focusing on the conditions aspects of the CMO framework, as previous sections above have, this section traces how regional conditions shape the mechanisms currently available for cooperation within anticipatory governance in Latin America. Moreover, the section is based on the survey responses and interview insights from experts working in the Latin American context. This supports the identification of what characterises AI governance and the potential role of informal diplomacy in this setting, as well as specific challenges and potentials tied to regional coordination.

Respondents seem to assume a baseline of low trust, where cooperation cannot rely on goodwill alone and instead requires mechanisms that reduce uncertainty and opportunism. This disposition is reinforced by interview findings, which point to corruption, instability, and low trust in national government agencies as structural features of the regional context, not merely incidental factors. One interviewee went further, noting that political uncertainty is a major theme, with regional AI convergence in Latin America seen as heavily dependent on elections, ideological shifts, and polarisation rather than on technical readiness alone.

This connects directly to the strong perception that the absence of a regional framework is a constraint, as indicated by 77.8% of Latin American survey respondents. AI governance is perceived as an area where institutional fragmentation is particularly limiting. Unlike other domains where flexibility might be seen as an advantage, here the lack of coordination is interpreted as a structural weakness, especially given the cross-border nature of many AI-related risks. Notably, however, one interview respondent offered a dissenting view, suggesting that the lack of a strong regional framework could in fact be seen as beneficial flexibility: a reminder that the dominant narrative of fragmentation-as-weakness is not universally shared.

Among the risks perceived as most pressing, interviewees highlighted military capabilities and armed groups, with one expressing particular concern about the use of AI by armed groups in Latin America, alongside threats to privacy, risks to information integrity, and the impact of AI on labour markets. One interviewee flagged biosecurity enablement, including the facilitation of new illicit drugs, as a significantly underrated AI risk. At the same time, the same interviewee was less alarmed about AI displacing jobs in the near term, arguing that the energy demands of large-scale AI deployment would render such a scenario unsustainable.

The recognition of the need for coordination is not matched by clarity on how to achieve it. When asked which institution is best positioned to anchor regional governance, survey responses are mostly split between CEPAL and Mercosur. This near-even division is significant. It suggests that there is no clear institutional focal point or leadership consensus. Interview data adds texture to this picture: the OAS was mentioned as a potential political framework, but interviewees were quick to note that certain issues cannot be enforced within it, and that ideological divisions and resource constraints further limit its effectiveness. Mercosur and UNASUR were described as less advanced, while the Santiago Process, built around action plans rather than binding treaties, was identified as the most operationally developed. Strikingly, one interviewee expressed doubt that regional coordination on regulation would happen at all, citing a lack of consensus, and suggested that governance processes should instead be centralised within the UN.

This lack of institutional clarity is further reinforced when looking at the question on urgent cross-border AI issues. Here, 44.4% of survey respondents converge on critical infrastructure, organised crime and security, and misinformation as top priorities. In this sense, the responses suggest that AI is not introducing entirely new risks, but rather amplifying existing structural challenges, making regional coordination more necessary but also more complex. Interviews corroborate and deepen this picture.There is a notable pessimism regarding the feasibility of achieving such coordination in the near term. A majority of survey respondents (55.6%) rate the likelihood of meaningful convergence of AI standards within 2–3 years at 2 out of 5, indicating low expectations. The reasons provided are mostly political disagreements and coordination difficulties. Interviews echo this assessment, with interviewees pointing to ideological and political clashes, lack of resources, and the outsized influence of large companies based in the United States or China as compounding factors. A particularly structural constraint identified in the interviews is the scale problem: none of the subregional alliances in Latin America have sufficient membership or political weight to replicate initiatives like the EU AI Act, leaving the region without the collective capacity to shape binding legal standards.

A similar pattern emerges in the evaluation of informal diplomacy mechanisms. Survey responses are generally low, with 33.3% rating effectiveness at 2, 11.1% at 1, and 22.2% at 3 on a 1–5 scale. This suggests that informal channels are not perceived as strong substitutes for formal governance structures. In other words, while these mechanisms exist, they are seen as insufficient to address systemic or catastrophic risks, particularly in a context where formal coordination is already weak. That said, interviews offered a more nuanced appreciation of informal diplomacy.

One interviewee described these dialogues as valuable precisely because they illuminate what is happening in practice and on the ground, improving both the design of public policies and the conduct of foreign policy. In their words, such processes "allow us to adjust details in regulation that we might not have considered otherwise," and while they are "small pieces within a larger process, they matter." This suggests that the value of informal mechanisms may be underestimated when evaluated solely against formal governance benchmarks.

Another key insight comes from the question about civil society inclusion. Respondents identify resource constraints and limited access to decision-making spaces as the main barriers. This is important because it highlights that the governance challenge is not only horizontal (between states), but also vertical (between institutions and society). The implication is that AI governance in the region risks becoming elite-driven and technocratic, with limited input from broader social actors. Interview respondents reinforced this concern, emphasising that civil society actors, the private sector, and academia bring expertise that national governments often lack, and crucially, that these actors enjoy greater public trust than governments do. One interviewee also stressed the importance of education as a foundational governance tool, arguing that people need to understand how AI works, what its uses are, and what its risks are, both to protect themselves from disinformation and malicious uses, and to participate in a more qualified debate about regulation. Given that many of the risks identified, such as misinformation or socioeconomic disruption, directly affect vulnerable populations, this disconnect could further weaken the effectiveness and legitimacy of governance efforts.

Finally, the question about what concrete measure should be implemented within the next 12 months yields highly diverse answers, with no clear consensus. Interviews mirror this diversity while surfacing a few recurring themes. Several interviewees advocated for beginning with soft instruments or soft law as the only realistic way to engage states without triggering sovereignty concerns. Other suggestions included advisory processes from international organisations, baseline evaluation and audit standards, engagement with the private sector as a first step, and a regional conference bringing together all relevant actors. Some pointed to specific initiatives, such as one led by the Colombian government, as potential models, while others called for political statements on the principles guiding AI adoption in the public sector, or for legal charters governing AI use by governments. The breadth of these proposals, taken together with the survey data, reinforces the current conditions of AI governance in Latin America: a region that recognises the urgency of AI governance but has yet to converge on a common path forward.

6. Discussion and Implications for Policy

6.1 Informal Diplomacy Mechanisms

The current context of anticipatory governance of AI consists of low trust amongst actors, especially towards the private sector as illustrated in the data analysis. This lack of trust is further exacerbated by the fact that AI governance is dominated by bilateral agreements specifically within the U.S. and China. One concern raised across interviews is the degree to which governance has shifted away from state-led multilateral frameworks towards bilateral agreements and predominantly within the private sector, whereby one interview explicitly states how current companies that are developing AI are also developing their governance. This phenomenon marks a regime shift where the decision-making of AI is concentrated within the private sector, actors who are neither democratically accountable nor structurally incentivised to prioritise catastrophic risk prevention. This clearly shows the lack of coherence and coordination within current AI international governance efforts.

Furthermore, various intergovernmental organisations are working on international governance mechanisms on AI leading to overlapping mandates on this technology producing redundancy and confusion amongst actors on which mandate to follow. Multiple intergovernmental organisations, including bodies such as the ITU, OECD, UNESCO and other international forums, are developing AI governance mechanisms, often with overlapping mandates and different regulatory priorities. This further creates uncertainty among actors about which standards, institutions or processes should guide their behaviour. Keohane and Victor's (2011) criteria are useful in understanding the mishaps of the current AI governance landscape regarding anticipatory governance, whereby this regime complex lacks coherence as institutions and states advocate for diverging approaches to regulation, accountability as key agreements occur behind closed doors, determinacy as key terms have different meanings across systems, sustainability as geopolitical dynamics bring an extra layer of uncertainty and epistemic quality as expertise remains concentrated within a few actors.

While formal governance mechanisms are crucial in AI catastrophic risk prevention, many interviews note that AI development often outstrips the timeline of formal diplomatic processes, especially when countries such as China and the U.S. are engaging in an "AI race" to develop the most sophisticated AI models. Moreover, the time taken to establish these formal diplomatic efforts is much slower than the development of AI. Yet, as defined in the literature, anticipatory governance is defined by its capacity to act under such conditions of uncertainty, such as catastrophic risks, before they become irreversible.

As such, this report recommends the development of non-binding shared risk taxonomies and multilateral incident reporting frameworks, with human rights integrated directly into this design. These mechanisms first introduce a common language of catastrophic risk in the context of AI that is applicable to all state and non-state actors and so going beyond geopolitical competition. This is also consistent with Wendt's analysis that state interests are formed through social interaction and shared meanings and these shared taxonomies do not require states to abandon their competitive interests. Instead, they encourage states to conduct this competition within a structure of agreed safeguards. Additionally, these two multilateral mechanisms provide a structural correction to the bilateral dominance of current arrangements as explained above.

A second recommendation is the need for a standardisation process. As seen, many IGOs and civil society actors are focusing on AI governance, but governance efforts remain fragmented and lack collaboration. There is consensus on the need for broad principles such as transparency and accountability, but less clarity about how to translate them into concrete technical and institutional requirements. However, the question of who would operationalise the shared risk taxonomies and multilateral incident reporting frameworks remains unclear. The ITU could be a suitable organisation given its technical mandate and broad membership, however, the interview analysis also pointed out the issue of funding constraints especially for UN agencies.

CMO	Informal Diplomacy Mechanisms
Conditions	• Low trust amongst actors in field especially private sector. • AI governance currently dominated by bilateral conditions mainly between China and the U.S. • Overlapping IGO mandates creating fragmentation. • AI development currently outpaces formal diplomatic efforts.
Mechanisms	• Non-binding shared risk taxonomies with human rights framework focus. • Multilateral incident reporting. • Standardisation process to translate principles to technical processes. • ITU as a potential coordinating body.
Outcomes	• Shared mechanisms such as risk taxonomies introduces common language when speaking about catastrophic risk. • Incident reporting creates accountability in the absence of binding treaties. • Standardisation led by the ITU reduces mandate overlap.

Table 3: CMO applied to informal diplomacy

6.2 Non-state Actors

6.2.1. Private Sector Inclusion

Expert findings confirm the private sector as a necessary but challenging diplomatic actor. Experts shared substantial concerns about the sector's decision-making power, and lack of trustworthiness due to profit incentives. Simultaneously, as the owners and designers of AI models private companies are considered to be effectively shaping AI governance through model design. Furthermore, by possessing developmental control and technical expertise, far beyond that of states, some experts argue that they should be the first point of contact for governing AI. Skeptics, however, highlight that even when included through informal diplomatic mechanisms, private companies are typically represented by diplomatic or policy representatives, rather than the engineers or technical teams actually designing AI systems.

As the influence of public institutions on technical choices remains limited, voluntary commitments of private companies send a positive signal for trust. However, it is highlighted that they may undermine transparency, and can be revoked at any time, leading to a remaining lack of control, trust and predictability in AI governance.

The need for the constructive inclusion of the private sector in informal diplomacy is highlighted. Based on the analysis of expert consultations, this project therefore recommends careful consideration regarding what type of private sector representation is established through diplomacy mechanisms. Special emphasis should be placed on engaging with technical private sector representatives rather than policy representatives. Voluntary regulatory commitments should therefore be approached with cautious optimism. Technical insights should be encouraged and considered, however, they should be included in capacity building rather than be relied on as reassurance. Anticipatory AI governance should aim for binding regulation. However, as this has proven to be challenging, engagement in informal diplomacy mechanisms such as incident reporting, benchmark standards and risk taxonomies should be prioritised.

CMO	Private Sector
Conditions	• Private sector seen as untrustworthy due to profit incentives. • Technology companies send policy staff not technical engineers to discuss governance processes.
Mechanisms	• Prioritise technical (engineer-level) private sector representation in global governance efforts. • Engage these engineers and the private sector in incident reporting, benchmark standards and risk taxonomies.
Outcomes	• Technical representation improves quality of private sector contributions to governance.

Table 4: CMO applied to the private sector

6.2.2 Civil Society and Research Community

Expert consultations demonstrated overwhelmingly positive attitudes towards the inclusion of non-state actors in the form of civil society and academic and research community representatives, highlighting the value of informal diplomacy mechanisms. Their expertise on the effects of AI harms and regulation on human rights and affected communities is seen as especially valuable. Interviews, furthermore, highlight perceptions of civil society and researchers as being among the most trustworthy actors. Experts, however, underlined that they currently lack the agency and influential power to drive real change.

Based on expert perceptions of their value, this project recommends that informal diplomacy mechanisms should prioritise academic, research and civil society groups. Due to their superior expertise and consideration of outcomes on human well-being, should these actors be given more agency in analysing current and potential AI governance mechanisms, as well as in standardisation processes.

CMO	Civil Society
Conditions	• Most trusted actor. • Resources constraints and lack of agency limits driving effective change.
Mechanisms	• Prioritise civil society, academic and research groups within informal diplomacy mechanisms. • Embed civil society expertise in standardisation processes.
Outcomes	• Human rights approaches are more systematically integrated into risk frameworks. • Reduce technocratic concentration in global governance.

Table 5: CMO applied to civil society

6.3 Regional Insights

The findings from this case study resonate strongly with the existing literature, painting a coherent picture of a region that recognises the urgency of coordinated AI governance but faces foundational structural barriers that make such coordination deeply difficult to achieve.The low level of trust found by the respondents corresponds to that described by Vivares et al. (2025), where the political and economic nature of AI governance in the region is determined primarily by the influence of models from the Global North. Interviewees also indicated that their ability to independently govern the region has been impacted greatly by large technology companies from the United States and China, which supports the centre-periphery dynamics postulated by Vivares et al. as being central to their evaluation of AI governance.

There is an agreement between the institutional instability found in the case study and the foundational weaknesses described by Sanchez-Pi et al. (2021), such as weak data governance, lack of infrastructure and high levels of public distrust. Contreras (2024) argues that Latin America should create institutional capacity to anticipate and reduce future risks, rather than waiting until an issue arises before regulating it. The survey participants' strong belief that there will not be any meaningful convergence of regulations within 2–3 years reflects a region where, as Sanchez-Pi et al. point out, countries have begun creating regulations without

sufficient cross-border connections. This predominantly present risk framework is at odds with Contreras' (2024) assertion that Latin America should develop institutional capacity to anticipate and reduce future risks.

The project suggests three recommendations. First, regional actors should put the use of soft law mechanisms and political statements on AI principles into their priority frameworks since binding frameworks are unlikely to be acceptable for political reasons and sovereignty issues impede this form of cooperation. Second, coordination efforts must also include civil society, academia and the private sector, whose greater public legitimacy could help address the trust deficit undermining state-led initiatives. Third, even within capacity constraints, governance frameworks should incorporate anticipatory tools, such as baseline audit standards and ongoing monitoring mechanisms, in order to move regional approaches from reactive to forward-looking.

When considering the recommendation on coordination efforts at a regional level, Latin America should refer to current regional efforts of the AU, such as the AU continental strategy of 2024. AU regional efforts provide a crucial perspective on how countries within the Global South are able to come together to identify regional priorities regarding AI and represent this at a global level. This was seen through the AU High Level Policy Dialogue on AI held on May 17th 2025 where African leaders came together to discuss how to establish continental AI Governance mechanisms based on ethics and human rights. While the AU governance frameworks and initiatives are still developing, this report argues that they provide a good base for Latin America to refer to what regional cooperation and coordination can look like, especially regarding ethics and human rights integration within such emerging frameworks. Finally, a challenge of such a regional effort for Latin America would be deciding which institution would lead it. As seen in through the interviews, Latin America has several regional institutions such as CEPAL and Mercosur, thus this report predicts a level of uncertainty as to which regional body can spearhead such an initiative or whether it calls for the creation of a new regional body within Latin America to take this on.

CMO	Latin America
Conditions	• No regional governance frameworks. • No clear institutional leader to take up regional efforts in AI governance: CEPAL vs Mercosur. • Governance remains reactive and dependent on electoral cycles.
Mechanisms	• Soft law mechanisms and political statements on AI principles. • Baseline audit standards and ongoing monitoring to shift from reactive to anticipatory governance. • AU continental strategy (2024) as a reference for Global South regional coordination.
Outcomes	• Shared AI principles established without triggering sovereignty. • Institutional capacity for anticipatory governance is an entry level effort in AI governance.

Table 6: CMO applied to regional insights

7. Final Remarks

This research has analysed anticipatory governance and informal diplomacy mechanisms that can aid in the mitigations of catastrophic risks that are arising from AI technological advancement. In a fragmented and mutable governance landscape, multilateral efforts alone are unlikely to provide a sufficient response to the emerging problems. Therefore, the role of informal diplomacy, expert consultation and multi-stakeholder dialogue is of great importance, as it helps build trust, aligns risk perceptions and creates practical pathways for coordination.

Several important themes emerged from consultations with experts. One of them is the value of a human rights-based approach as a substantial framework for governing AI development. It should not be considered simply a high-level abstraction, but rather a concrete mechanism for establishing international norms regarding risk taxonomies, incident reporting and the setting of accountability standards. Including human rights considerations from the outset of efforts to develop informal diplomatic approaches to governing AI risks offers a way to establish legitimacy in these discussions and ensure that the voices of those most affected by these risks are heard.

The findings also suggest a tension in relation to the private sector. Companies such as Open AI and Antropic possess the AI material capabilities while being the least trustworthy of all stakeholders in terms of governance. The suggested approach is that informal diplomacy mechanisms should prioritise technical representation from the private sector, to facilitate constructive dialogue and trust-building with individuals directly involved in designing AI systems. The solution should not be exclusion, but rather more deliberate engagement, prioritizing the voices of technical experts and developers over those of political representatives. Meanwhile, civil society and academia remain excluded from meaningful participation, despite being identified as the most trusted actors. These are the actors best positioned to identify potential risks overlooked by states and companies, and to bring affected communities into discussions that often become technocratic.

The findings from Latin America highlight how the dynamics of AI governance occur outside the Global North. Without a shared regional framework, these countries often find themselves forced to deal with the consequences of AI risks largely in isolation. However, the emerging continental strategy of the African Union, while still in its early stages, presents a viable example of a regional framework, demonstrating that collective regional action is possible even in challenging contexts in the developing world.

Ultimately, the risks posed by emerging AI technologies are real, and the governance mechanisms available to address them remain insufficient. Thus, anticipatory AI governance requires more than technical risk assessment. It demands coordinated diplomatic efforts, accountable participation and sustained dialogue across sectors and regions. Informal diplomacy can help create the conditions for cooperation, but its effectiveness depends on whether it can connect relevant stakeholders, build trust and translate different technical, political and normative concerns into a shared language.

Further research is still needed within four main areas. First, regarding the AU's Five Year Implementation Plan and Annual Conference. Tracking the progress of both these initiatives highlights whether regional policies and efforts are able to be translated into actual governance processes. The AU is one of the first Global South regional actors attempting to harmonise national AI strategies, mobilise AI expertise within the continent and ensure African values and priorities are being included into these governance frameworks. These developments present the AU as a useful avenue for future research to understand how regional organisations can mobilise countries in having shared priorities in AI Global Governance and the role this plays in global governance

structures. Such future research would also provide valuable lessons for Latin America where AI governance remains fragmented and there is a lack of coordination mechanisms.

Second, one of the recommendations given in this report was to prioritise and facilitate representation of the private sector in a technical capacity. Expert consultations highlighted that the effectiveness of private sector inclusion is undermined when companies send policy or diplomatic representatives, rather than those responsible for actually developing the technologies, and potentially steering AI governance through model design. However, the report was unable to identify and present specific strategies which could be utilised to increase dialogue with the technical experts and engineers in AI firms. To deepen the findings of this report and increase the applicability of its recommendation, further research into how technical private sector representation can be facilitated and incentivised is needed.

Third, the report addresses the mitigation of harms to human society, in the form of anticipatory AI governance. In this, however, questions about the suitability of anticipatory governance itself, and its potential to cause harm to human societies, have not been addressed. UNESCO (2025) describes anticipatory governance measures as the identification of potential risks and measures before they materialise, however, as Amoore (2013) and Guston (2014) highlight, this may have contradictory implications for the prevention of harm. Amoore describes how this governance grants the capacity to act on " emergent, uncertain, *possible* futures" (p.5), and can extend the limits of actors' authorities, including private technology companies. Furthermore, exceptional measures, employed in the name of risk management, become perennial and may obscure the initial objectives of protecting human societies for which they were employed. Guston further highlights how anticipatory strategies may steer governance towards technoscience rather than humane ends, however, does suggest that inclusion of actors who were previously excluded from governance, can help combat such dynamics. The report, therefore, does not suggest disqualifying anticipatory governance as a suitable framework for AI, yet recommends further research into potential harms of anticipatory measures, and how anticipatory governance can be structured to preserve initial objectives.

Finally, the report recommends prioritising mechanisms such as shared taxonomies, standardisation and incident reporting. While it suggests the International Telecommunication Union, as a potential body to lead these attempts, experts also proposed centralisation via the United Nations, or a new "Digital commons for humanity" body. Identifying a leader or centralised organisation to lead common mechanisms, was beyond the scope of this project, but remains a crucial avenue for further research.

8. Bibliography

Ade-Ibijola, A. and Okonkwo, C. (2023) 'Artificial Intelligence in Africa: Emerging Challenges', in Eke, D.O., Wakunuma, K. and Akintoye, S. (eds.) Responsible AI in Africa. Cham: Palgrave Macmillan. Available at:https://doi.org/10.1007/978-3-031-08215-3_5 (Accessed: 15 May 2026).

Alais, O. (2026) 'Governance by procurement: how AI rights became a bilateral negotiation'. Available at: https://www.hks.harvard.edu/centers/carr-ryan/our-work/carr-ryan-commentary/governance-procurement-how-ai-rights-became (Accessed: 16 April 2026).

Alter, K.J. and Meunier, S. (2009) 'The politics of international regime complexity', Perspectives on Politics, 7(1), pp. 13–24. (Accessed: 15 May 2026).

Amoore, L. (2013) 'Introduction: On the politics of possibility', in The Politics of Possibility: Risk and Security Beyond Probability. Durham: Duke University Press, pp. 1–26; 55–76. (Accessed: 15 May 2026).

Anthropic (2024) 'The case for targeted regulation'. Available at: https://www.anthropic.com/news/the-case-for-targeted-regulation (Accessed: 31 October 2025).

Anthropic (2026) 'Anthropic's Responsible Scaling Policy: Version 3.0'. Available at: https://www.anthropic.com/news/responsible-scaling-policy-v3 (Accessed: 21 April 2026).

Aranda, L. and Perset, K. (2024) 'Name it to tame it: defining AI incidents and hazards', OECD AI Policy Observatory. Available at: https://oecd.ai/en/wonk/defining-ai-incidents-and-hazards (Accessed: 15 May 2026).

Bostrom, N. (2002) 'Existential risks: analysing human extinction scenarios and related hazards', Journal of Evolution and Technology, 9(1), pp. 1–31 (Accessed: 15 May 2026).

Bradford, A. (2020) The Brussels Effect: How the European Union Rules the World. Oxford: Oxford University Press (Accessed: 15 May 2026).

Bradford, A. (2023) Digital Empires: The Global Battle to Regulate Technology. Oxford: Oxford University Press (Accessed: 15 May 2026).

CAIDP (2026) Council of Europe AI Treaty. Center for AI and Digital Policy. Available at: https://www.caidp.org/resources/coe-ai-treaty/ (Accessed: 15 May 2026).

Contreras, P. (2024) 'Convergencia internacional y caminos propios: regulación de la inteligencia artificial en América Latina', Actualidad Jurídica Iberoamericana, 21, pp. 468–493 (Accessed: 15 May 2026).

Council on Foreign Relations (n.d.) 'What is diplomacy?'. Available at: https://education.cfr.org/learn/reading/what-diplomacy (Accessed: 12 May 2026).

Crawford, E. (2021) 'What is "soft" law? An analysis of the concept of non-binding instruments and provisions in international law', in Crawford, E., Non-Binding Norms in International Humanitarian Law. Oxford: Oxford University Press, pp. 8–39. https://doi.org/10.1093/oso/9780198819851.003.0002 (Accessed: 15 May 2026).

Cui, T. et al. (2024) 'Risk taxonomy, mitigation, and assessment benchmarks of large language model systems', arXiv. https://doi.org/10.48550/ARXIV.2401.05778 (Accessed: 15 May 2026).

Digwatch (n.d.) 'Capacity development in AI and digitalisation'. Available at: https://dig.watch/topics/capacity-development/ (Accessed: 12 May 2026).

DiploFoundation (2021a) 'Geography and diplomacy'. Available at: https://www.diplomacy.edu/topics/geography-and-diplomacy/ (Accessed: 12 April 2026).

DiploFoundation (2021b) 'Track two diplomacy'. Available at: https://www.diplomacy.edu/topics/track-two-diplomacy/(Accessed: 12 April 2026).

Gálvez Callirgos, P. (2025) Pathways to Inclusion: Advancing Civil Society's Role in AI Governance. Geneva: Globethics Publications (Accessed: 15 May 2026).

Guston, D.H. (2014) 'Understanding "anticipatory governance"', Social Studies of Science, 44(2), pp. 218–242 (Accessed: 15 May 2026).

Hartwig, A. and Sánchez, L. (2024) 'Inteligencia Artificial y el Estado argentino', conference paper, Córdoba (Accessed: 15 May 2026).

Hatano, A. (2026) 'Ethical AI and business & human rights', International Organizations Law Review, 22(3), pp. 509–530 (Accessed: 15 May 2026).

Hendrycks, D., Mazeika, M. and Woodside, T. (2023) An Overview of Catastrophic AI Risks. arXiv (Accessed: 15 May 2026).

ILIA (2025) Latin American Artificial Intelligence Index (ILIA 2025). Available at: https://wireunwired.com/latin-american-artificial-intelligence-index-ilia-2025-launches-a-deep-dive-into-regional-progress-rankings-and-future-strategy/ (Accessed: 15 May 2026).

International AI Safety Report (2025) International AI Safety Report. Available at: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025 (Accessed: 15 May 2026).
Keohane, R.O. and Victor, D.G. (2011) 'The regime complex for climate change', Perspectives on Politics, 9(1), pp. 7–23 (Accessed: 15 May 2026).

Koenig, P.D. (2025) Understanding the Politics of Artificial Intelligence. Edward Elgar Publishing. https://doi.org/10.4337/9781035348022 (Accessed: 15 May 2026).

Kokotajlo, D. et al. (2025) AI 2027. Available at: https://ai-2027.com/ (Accessed: 15 May 2026).

Kwarkye, T.G. (2025) 'We know what we are doing', Canadian Journal of African Studies, 59, pp. 437–455.

Lenoir et al. (2025) AI Governance: Empowering Civil Society. Available at: https://www.renaissancenumerique.org/en/publications/ai-governance-empowering-civil-society/ (Accessed: 15 May 2026).

Mäntymäki, M. et al. (2022) 'Defining organizational AI governance', AI and Ethics, 2(4), pp. 603–609 (Accessed: 15 May 2026).

Marino, D., Stilo, P. and Serra, F. (2025) 'Artificial intelligence and geopolitical competition', in Generative Artificial Intelligence and the Fifth Industrial Revolution. Cham: Springer (Accessed: 15 May 2026).

Medeiros, M. (2020) 'Public and private dimensions of AI technology and security', CIGI Online (Accessed: 15 May 2026).

Muggah, R. (2025) 'Building an anticipatory governance strategy', World Economic Forum (Accessed: 15 May 2026).

Munga, J. and Quansah, S. (2025) 'Understanding Africa's AI governance landscape', Carnegie Endowment for International Peace (Accessed: 15 May 2026).

Nayat Sanchez-Pi, L. et al. (2021) A Roadmap for AI in Latin America. GPAI (Accessed: 15 May 2026).
Nemko Digital (n.d.) 'AI Regulation in Nigeria'. Available at: https://digital.nemko.com/regulations/ai-regulation-in-nigeria (Accessed: 15 May 2026).

OECD (2024b) OECD AI Principles (Accessed: 15 May 2026).

OECD (2025) Regulatory Sandbox Toolkit (Accessed: 15 May 2026).

OECD (2026) AI Principles Overview (Accessed: 15 May 2026).

OECD (n.d.) Data Governance (Accessed: 15 May 2026).

OECD Observatory of Public Sector Innovation (n.d.) Futures and Foresight (Accessed: 15 May 2026).
OECD.AI (n.d.) AI Compute Database (Accessed: 15 May 2026).

Peña González, J.G. et al. (2024) 'El principio de transparencia y la regulación de la IA en Colombia', Diálogos y Voces Judiciales, 2(2), pp. 83–99 (Accessed: 15 May 2026).

Polchar, J. (2024) 'Using foresight to anticipate emerging critical risks', OECD Working Papers on Public Governance, No. 79 (Accessed: 15 May 2026).

Polchar, J. and Alfonzo Santamaria, N. (2024) 'Mapping emerging critical risks', OECD Working Papers on Public Governance, No. 78 (Accessed: 15 May 2026).

Pratt, T. (2018) 'Deference and hierarchy in international regime complexes', International Organization, 72(3), pp. 561–590 (Accessed: 15 May 2026).

Raustiala, K. and Victor, D.G. (2004) 'The regime complex for plant genetic resources', International Organization, 58(2), pp. 277–309 (Accessed: 15 May 2026).

Risse, T. (2000) "Let's Argue!": Communicative Action in World Politics', International Organisation, 54(1), pp. 1–39.doi:10.1162/002081800551109 (Accessed: 18 May 2026)

Schmitt, L. (2022) ‘Mapping global AI governance’, AI and Ethics, 2(2), pp. 303–314 (Accessed: 15 May 2026).

Sharma, R. (2025) ‘The $4.8 trillion AI trust crisis’, World Economic Forum (Accessed: 15 May 2026).

Souppaya, M. et al. (2024) Secure Software Development Practices for Generative AI. NIST.

Stanford HAI (n.d.) What is AI alignment? (Accessed: 15 May 2026).

Taeihagh, A. (2021) ‘Governance of artificial intelligence’, Policy and Society, 40(2), pp. 137–157 (Accessed: 15 May 2026).

The New York Times (2025) ‘Video: opinion | The forecast for 2027? Total A.I. Domination.’ Available at: https://www.nytimes.com/video/opinion/100000010157582/the-forecast-for-2027-total-ai-domination.html (Accessed: 15 May 2026).

The Stanley Foundation (2016) A Multistakeholder Governance Agenda (Accessed: 15 May 2026).

UNESCO (2022) Recommendation on the Ethics of Artificial Intelligence (Accessed: 15 May 2026).

UNESCO (2023) Anticipatory Governance (Accessed: 15 May 2026).

UNESCO (2025) Comparing Governance Mechanisms for AI (Accessed: 15 May 2026).

United Nations (n.d.) Artificial Intelligence (AI) (Accessed: 15 May 2026).

Vallejo, N. and Hauselmann, P. (2004) Governance and Multistakeholder Processes (Accessed: 15 May 2026).

Vanegas, W.J. et al. (2024) ‘Políticas públicas ante la revolución de la IA en Colombia’, Revista Venezolana de Gerencia, 29(106), pp. 865–883 (Accessed: 15 May 2026).

Vivares, E. and Rodriguez-Ortiz, F. (2025) ‘Artificial intelligence meets South America’, Globalizations (Accessed: 15 May 2026).

Waltz, K.N. (1979) Theory of International Politics. Reading, MA: Addison-Wesley.

Wendt, A. (1992) ‘Anarchy is what states make of it’, International Organization, 46(2), pp. 391–425 (Accessed: 15 May 2026).

World Economic Forum (2025) The Global Risks Report 2025 (Accessed: 15 May 2026).

World Economic Forum (2026) The Global Risks Report 2026 (Accessed: 15 May 2026).

9. Appendix

Annex 1: Qualitative codebook
Annex 2: List of abbreviations and key definitions
Annex 3: Latin American Countries
Annex 4: African Union AI governance
Annex 5: Further details on Methodology

10. Remark on the Use of Artificial Intelligence

In accordance with the Geneva Graduate Institute *Rules and Good Practices for Ethical Use of Artificial Intelligence in Academic Work*, artificial intelligence (ChatGPT-5.1, 2025) was used in the revision of this text to enhance sentence flow and coherence, improve overall organization, and ensure grammatical and linguistic accuracy, as well in ensuring coherence and accurate Harvard style in the bibliography.

APPLIED RESEARCH PROJECT 2025/26

in partnership with

GENEVA GRADUATE INSTITUTE

Globethics

POLICY BRIEF

MITIGATING CATASTROPHIC AI-RISKS:
THE ROLE OF INCLUSIVE & ANTICIPATORY INFORMAL DIPLOMACY

By Cora Ebner, Eesha Jamal, Greta Franco & Isabela Duleba

POLICY BRIEF

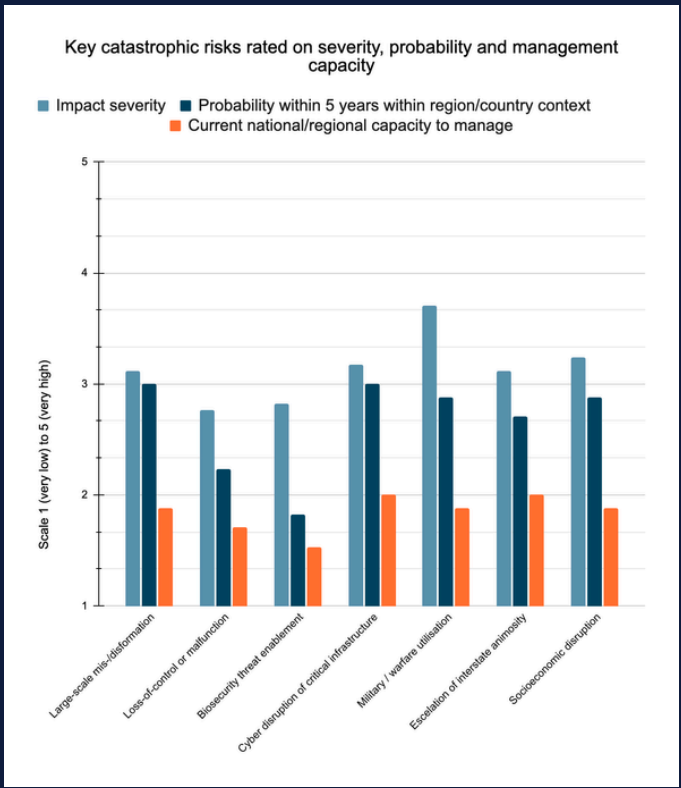
MITIGATING CATASTROPHIC AI-RISKS: THE ROLE OF INCLUSIVE & ANTICIPATORY INFORMAL DIPLOMACY

EXECUTIVE SUMMARY

- Governance actors are highly concerned about **catastrophic AI risks**, but governance capacity is lagging behind technological progress. **Informal diplomacy** offers a promising route for more inclusive and anticipatory AI governance.
- **Civil society** and the research community are generally viewed positively for their transnational and human rights expertise, but they lack the power to drive substantive change. The **private sector**, while often seen as untrustworthy and contested in terms of inclusion, holds critical technical expertise and implementation capacity.
- Effective AI governance depends on **political prioritization, institutional capacity, shared risk understanding, and incentives for cross-actor cooperation.** Informal diplomacy can support this by fostering collaboration among states, multilateral organizations, industry, and civil and research actors on standards, risk taxonomies, and incident reporting, with human rights embedded in design.
- **Regional frameworks can further enable risk mitigation** by aligning states in similar contexts. Overall, soft-law approaches appear to be the most feasible starting point for coordinated, anticipatory governance.

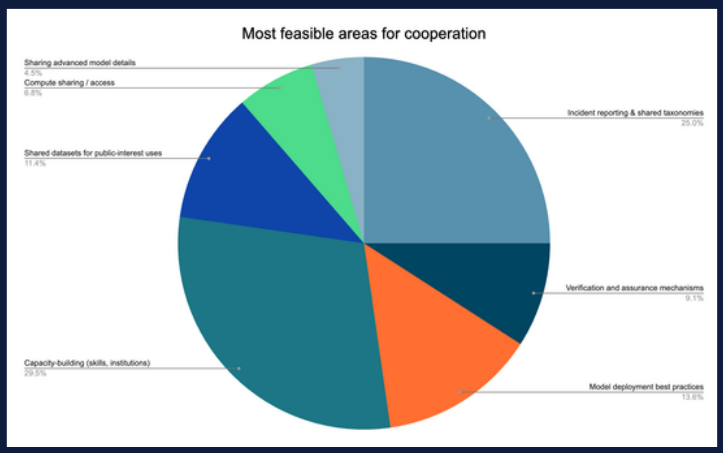
AI RISK & GOVERNANCE LANDSCAPE

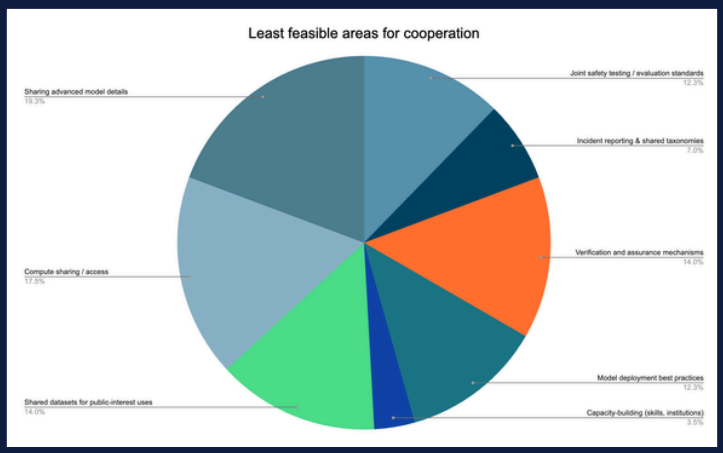
- AI became a catastrophic risk when combined with contexts such as **weak governance systems, military competition and private sector concentration.**
- Embedding ethical frameworks within AI models is not seen as **effective risk prevention** as AI systems do not respect these frameworks and do not hold human capabilities such as empathy to understand the values of these frameworks.
- The governance landscape is marked by **uncertainty, power asymmetries,** and **competitive dynamics**. As such, AI advancements are occurring in an environment that generally lacks shared global norms, trust-building channels or communication frameworks.
- While AI risks are a key concern of governance agendas globally, **key decisions are primarily made by agreements between few,** dominating states and private companies, leaving the remaining actors to adapt accordingly.



INFORMAL DIPLOMACY MECHANISMS

- Challenges in alignment on governing AI appear closely linked to a lack of trust between actors.
- Mechanisms, such as shared incident reporting and technical standards, are seen to be effective in place of formal diplomatic mechanisms such as sanctions/enforcement
- Too many organisations are developing AI governance frameworks, leading to a lack of consensus on which mechanisms are effective

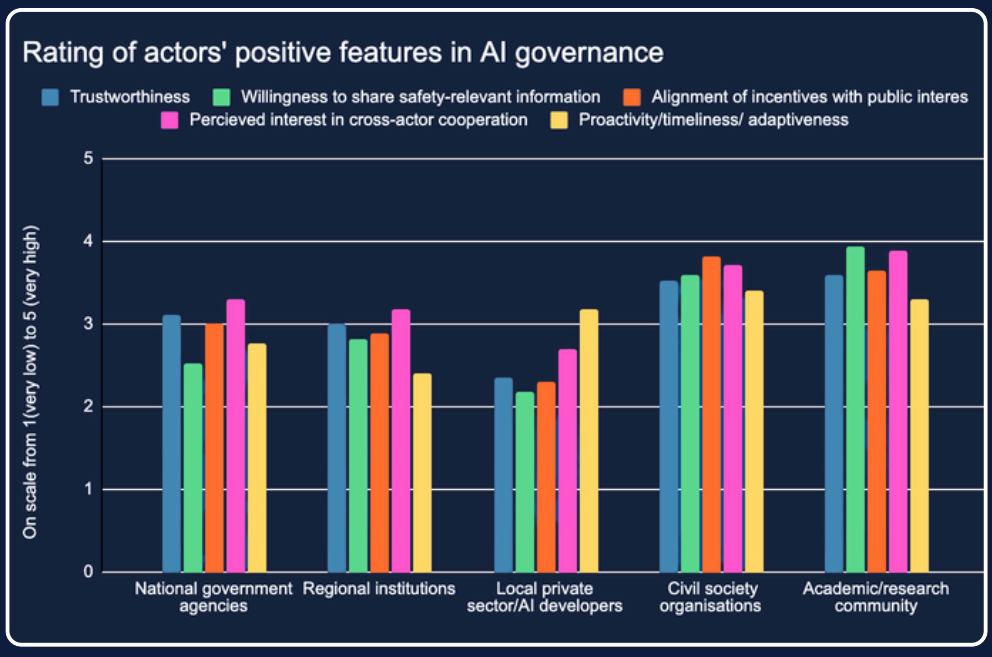




NON-STATE ACTOR INCLUSION

Civil Society & Research Community

- Highest ranked actors across governance characteristics.
-
- Their inclusion, highlighted as especially important due to their transnational and human rights expertise.



Private sector

- Experts are concerned about current large decision-making power, seen as essentially shaping governance through model design.
- Seen as untrustworthy due profit-incentives, but recognised as an indispensable technical partner due to capacity to enhance efficiency, applicability and speed of governance.

HINDERING FACTORS FOR COOPERATION

Biggest Resource Gaps

1. **Political polarization,** which undermines policy continuity and makes cooperation dependent on shifting political alignments.
2. **Lack of Legal Authority and Mandate** which weakens enforcement, accountability and the translation of cooperation into concrete outcomes.
3. **Funding issues** which restrict implementation of AI governance mechanisms.

Key Barriers to Cooperation

1. **Interest misalignment,** especially between governments and private-sector actors.
2. **Speed of AI advancements** which governance frameworks struggle to match.
3. **Power play amongst stakeholders,** including states, firms and international institutions.

ENABLING FACTORS FOR COOPERATION

Mechanisms for Improving Trust

1. **Political transparency requirements** which improve trust, accountability and clarity around AI governance decision-making.
2. **Shared technical standards** which can reduce uncertainty and support coordination.
3. **Independent oversight body** helping to improve accountability and reduce concerns about capture.

Features of Effective Coordination

1. **Political buy-in** from governments and relevant institutions, identified as the most important feature for coordination mechanisms to work.
2. **Legitimacy and inclusivity,** especially the involvement of actors beyond governments, including civil society, academia, technical experts and the private sector.
3. **Clear legal authority and mandates** which can give coordination mechanisms the institutional weight needed to move beyond voluntary dialogue.

LATIN AMERICA

Survey and interview data from across Latin America reveal a consistent pattern of institutional fragility, low trust and coordination failure.

1. **Distrust** is the defining feature of the regional landscape. 55.6% of survey respondents rate the likelihood of meaningful convergence of AI standards within two to three years at just 2 out of 5, citing political disagreements and coordination difficulties. The **absence of a regional framework is identified as a constraint** by 77.8% of respondents, yet there is no consensus on how to fill this gap, with responses almost evenly split between CEPAL and Mercosur as the preferred anchor institution.
2. **Informal diplomacy** mechanisms are not seen as adequate substitutes for formal structures, with 44.4% of respondents rating their effectiveness at 2 or below on a 1–5 scale. Civil society academia, and the private sector remain largely excluded from decision-making despite enjoying greater public legitimacy than governments.
3. The most urgent cross-border AI priorities according to 44.4% of respondents are **critical infrastructure, organised crime and misinformation**, suggesting that AI is amplifying existing structural challenges rather than introducing entirely new ones.

The region is aware of the governance challenge but has yet to develop the institutional foundations or political consensus needed to respond to it effectively.

RECOMMENDATIONS

Informal Diplomacy Mechanisms

1. Creation of **non-binding shared risk taxonomies** and **multilateral incident reporting frameworks**, with human rights integrated directly into this design.
2. Development of a **standardisation process** that brings together **current efforts by civil society and inter-governmental actors**, thus moving away from the current fragmented landscape of AI Governance
3. Standardisation process to be **spearheaded by one organisation**, such as the ITU given its technical mandate and broad membership

Private sector

1. Informal diplomacy mechanisms should prioritise **technical representation of the private sector**, to facilitate constructive dialogue and trust-building with individuals directly involved in designing AI systems.
2. Technical insights from the private sector should be leveraged for **capacity building**, rather than be relied on.
3. AI governance should currently prioritise **public-private cooperation** on incident reporting, benchmark standards and risk taxonomies.

Civil Society

1. Informal diplomacy mechanisms should have the **inclusion of civil society** and the **academic community** as a key priority.
2. The **human rights expertise** of the civil society **should be leveraged**, by giving civil society more agency in evaluating current and future governance frameworks.
3. Civil society's **transnational perspective and expertise,** should be utilised in suggested standardisation.

Regional insights

1. Prioritise **soft law** and political statements on AI principles, as binding frameworks remain politically unviable.
2. **Include civil society, academia,** and **the private sector** in coordination efforts, given their greater public legitimacy.
3. Incorporate **anticipatory tools**, such as audit standards and monitoring mechanisms, to shift from reactive to forward-looking governance.

CONCLUSION

Overall, anticipatory AI governance requires more than technical risk assessment; it demands coordinated diplomatic efforts, accountable participation and sustained dialogue across sectors and regions. Informal diplomacy can help create the conditions for cooperation, but its effectiveness depends on whether it can connect relevant stakeholders, build trust and translate diverse technical, political and normative concerns into a shared language.