Paris Call for Trust and Security in Cyberspace
Strategic Foresight Hub

PARIS PEACE FORUM de PARIS sur la PAIX

Policy Report

February 2025

FORGING GLOBAL COOPERATION ON AI RISKS: CYBER POLICY AS A GOVERNANCE BLUEPRINT

Lead authors

Pablo Rice
Head of programme, Cyber & Emerging Tech Governance, Paris Peace Forum

Charlotte Lindsey Curtet
Independant advisor, Paris Peace Forum & Hague Centre for Strategic Studies

Contributing organizations

Cloudflare

Sebastian Hufnagel
Senior Public Policy Manager

Center for Long-Term Cybersecurity (UC Berkeley)

Krystal Jackson
Non-Resident Research Fellow

Nada Madkour
Non-Resident Research Fellow

Center for Security and Emerging Technology (Georgetown University)

Colin Shea-Blymyer
Research Fellow

Fortinet

Marc Asturias
Vice President of Marketing and Field CISO for Government

GEODE Center

Frédérick Douzet
Director

Aude Géry
Researcher

Google Cloud (Mandiant)

Jon Tidwell
Practice Leader

INTERPOL

Mahdi Alaei
Cyber Strategy & Outreach Coordinator, Cybercrime Directorate

Pei Ling Lee
Head of Cyber Strategy and Capabilities Development, Cybercrime Directorate

Institute for Security and Technology

Jennifer Tang
Associate for Cybersecurity and Emerging Technologies

Mariami Tkeshelashvili
Senior Associate for Cybersecurity and Emerging Technologies

Microsoft

Nicholas Butts
Director, Global Cybersecurity and AI Policy

MIT FutureTech

Peter Slattery
Researcher, Lead of the MIT AI Risk Repository

Palo Alto Networks

Sam Kaplan
Senior Director & Assistant General Counsel, Public Policy & Government Affairs

Trend Micro

Josiah Hagen
Field CTO

UNIDIR

Giacomo Persi Paoli
Head of Programme, Security & Technology

WithSecure

Mikko Hyppönen
Chief Research Officer

Ieva Ilves
Advisor on European Policy

Individual contributors

Craig Jones
Director CyPol; Former Director of Cybercrime, INTERPOL

Chris Painter
Former President of the Global Forum on Cyber Expertise

Disclaimer *- The findings, interpretations, and conclusions presented herein are the outcome of a collaborative process facilitated and endorsed by the Paris Peace Forum. However, they do not necessarily reflect the exact views of each contributing organization or individual, nor do they represent the entirety of the Paris Call's community of supporters, or the member and partner organizations of the Paris Peace Forum.*

Acknowledgements

Support from all of the Paris Peace Forum's benefactors provides the foundation for the Forum's activities and events. The Forum's Cyberspace Governance Programme would like to especially thank the following benefactors for their commitment—not only through their financial contributions but also through their significant human investment—which has contributed to the publication of this report:

Microsoft is a global technology company which creates platforms and tools powered by AI to deliver innovative solutions that meet the evolving needs of customers worldwide. It has supported the Paris Peace Forum's core mission since 2018 as a strategic partner.

Trend Micro is a global cybersecurity leader, whose platform protects 500,000+ organizations and 250+ million individuals across clouds, networks, devices, and endpoints. It has been partnering with the Paris Peace Forum since 2024 to drive the secure adoption of AI.

WithSecure is a European-founded cyber security company helping to protect over 7,000 partners and 100,000 corporate customers. It has supported the Forum's Cyberspace Governance Programme since 2024 to help tackle the challenges raised by the AI-Cyber nexus.

Contents

Executive summary

The report explores key insights, emerging opportunities, and pathways for global governance of AI adversarial use risks, particularly AI-driven cyber risks. Drawing on two decades of cyber policy experience, this initiative of the Paris Call for Trust and Security in Cyberspace, aims to foster discussions among policymakers, experts, and other stakeholders and contribute to the AI Action Summit (February 2025, Paris). It is informed by the Paris Call Strategic and Foresight Hub and a consultation with the Paris Call community.

By examining global AI risk governance initiatives and drawing on parallels with international cyber policy, the report offers valuable insights for governing risks of adversarial use of AI. The report underscores the urgency of addressing threats at the intersection of AI and cybersecurity. **The adversarial use of AI has moved from theory to reality,** reshaping the threat landscape and challenging existing defenses. Among AI safety risks, cybersecurity stands out as the most transformative short-term factor, requiring coordinated global action.

As experts informing this report emphasized, **the use of AI for adversarial purposes is unlikely to fundamentally alter the core nature, and modalities, of cyber risks - at least in the short term**. Adversarial intent, targets, and potential resulting damage remain largely unchanged. The main foreseeable factors of disruption in the cyber threat landscape pertain to variables of time and volume: the **velocity** of execution of adversarial operations; the frequency of attacks; and the expansion of the spectrum of adversarial actors – due to lower entry barriers to engage in offensive activities. The increasing risk from these vectors can be significantly mitigated by a mainstream and tailored adoption of AI to bolster cyber defense capabilities.

Policymakers must shape an international agenda that prioritizes both functional aspects — key functions requiring international coordination—**and thematic areas focusing on actual and foreseeable types of AI-driven cyber risk that demand urgent global action**. Addressing the AI-cyber nexus is complex, but there is no regulatory vacuum. Since AI has not fundamentally altered the cyber threat landscape in the short term, a pragmatic approach should leverage existing ICT security frameworks, regulations, and policies, and international cooperation mechanisms. Assessing their applicability and adaptability to the new cyber challenges posed by AI allows global efforts to **prioritize severe risks that existing norms cannot adequately address**. The report proposes a **five-step methodology as a structured approach to determining whether new frameworks are needed for AI-risks from adversarial use, or if existing cybersecurity frameworks can be adapted**.

In its concluding highlights the report focuses on areas for global governance of AI-driven cyber risks, in particular **regulation, common ground and risk management; transparency and information sharing; and cyber defence**.

International law provides a compass to guide AI governance with a risk-based approach centered on accountability. While novel cyberattacks remain unobserved, AI's rapid progress demands vigilance. AI also enhances cyber defenses, emphasizing the need for strong risk management. The report highlights fragmented regulations struggle to keep pace with technological advances. While ex ante regulation poses risks, adaptive measures like sunset clauses and mandatory reviews are essential for governance to evolve alongside AI. Effective global governance must prioritize transparency, accountability, and inclusivity requiring collaboration across governments, industry and civil society.Coordination is crucial to prevent policy fragmentation and ensure globally aligned strategies. The report underscores the need for sharing best practices, AI incident reporting, and scientific knowledge-buildingto close evidence gaps and participation barriers.

Transparency and information sharing on AI breaches, vulnerabilities, and adversarial use are critical for governing AI-driven cyber risks. Without openness, evidence gaps widen, weakening accountability and response efforts. **Today's AI governance primarily relies on anticipation, focusing on developers while overlooking end-user regulation. However, recent threat intelligence advances show that tracking adversarial AI use is far from impossible, paving the way for governance approaches that balance responsibility between developers and end users**. To address the urgent reality of adversarial AI threats, clear guidelines and reporting platforms must be established. Mandatory reporting strengthens risk mitigation. Effective information sharing requires clear frameworks on sources, data-sharing methods, compliance, and harm assessment, refining risk management with real-world observations and emerging evidence.

AI is rapidly advancing **cyber defence** improving vulnerability management, threat detection, and incident response. However, critical cybersecurity gaps in organizations must be addressed to create a resilient foundation for integrating AI-driven cyber defenses. Collaboration is also essential to **share threat models, test vulnerabilities, and develop mitigation strategies before disclosure**.

As AI continues to reshape cyber defense, global governance must prioritize a balanced, adaptive approach that strengthens existing cybersecurity frameworks while integrating AI-driven solutions. **Addressing critical gaps, fostering collaboration between governments, industry, and research institutions, and investing in R&D and capacity-building will be essential to ensuring resilient, scalable defenses**.

A coherent international agenda, supported by effective coordination and dynamic regulatory measures, will be crucial in keeping pace with AI's rapid evolution. This report proposes a practical approach to support global governance efforts to mitigate AI-driven adversarial cyber risks while maximizing its potential for strengthened cybersecurity and stability.

Introduction

Purpose of this report

This policy report offers a strategic perspective on global governance of severe AI risks by leveraging two decades of international cyber policy, frameworks, experience and collaborative ethos of the cyber community. It examines how lessons learned can serve as a blueprint to inform the shaping of structures for international cooperation, particularly in countering malicious and adversarial use of AI. By analysing the governance modalities for AI-driven cyber risks as a foundational case, it highlights the need to align future AI-specific governance efforts with proven mechanisms and established norms.

The report highlights key insights, emerging opportunities, and pathways for global governance of AI malicious use risks, especially AI-driven cyber risks. As a Paris Call for Trust and Security in Cyberspace (hereafter Paris Call) community contribution to the AI Action Summit (10–11 February 2025, Paris), it will be formally presented at an official side event, fostering discussions among policymakers, experts, and other key stakeholders.

Through this contribution, the Paris Call – set up in 2018 - continues to serve as a forward-looking platform for tracking and shaping future dynamics in the ICT security landscape.

Audience

This report aims to inform a diverse range of stakeholders – government officials, staff from regional and international organizations, industry representatives, civil society groups and academia – involved in global governance of severe AI risks, and to engage the cybersecurity community, both policy and technical circles. Their extensive experience and expertise in identifying and adapting to an ever-evolving threat landscape makes them crucial contributors to AI risk management discussions and international collaboration for AI governance.

Scope

This report focuses on adversarial use of AI, with a pronounced focus on the use of AI for adversarial cyber purposes. It is informed by the International Scientific Report on the Safety of Advanced AI[1] which refers to "malicious use risks" as encompassing, but not limited to:

- **information manipulation targeting public opinion as a whole**, through generation of persuasive content at scale;

[1] Bengio Y. et al., International AI Safety Report, DSIT 2025/001 (January 2025)

- **generation of fake content that harms individuals in a targeted way**, including for fraud, extortion, sabotage and psychological abuse;
- **cyber offence**, with an increased ease and speed to conduct cyberattacks;
- **weaponization of AI in dual-use scientific areas**, in particular in the biological and chemical field, so as to facilitate the design of novel toxic compounds, obtaining necessary materials and accessibility of related information.

It distinguishes these risks from AI system malfunctions and those with far-reaching, systemic impacts, thus emphasizing a deliberate harmful purpose.

This policy report uses adversarial use of AI instead of malicious use, avoiding the restrictive notion of malice and the challenges of determining intent.

The layer of action explored in this policy report is that of global governance, distinct from both industry standards and domestic regulation. Global governance, as defined by Michael Zürn, refers to *"the exercise of authority and the establishment of norms beyond national borders to address shared goals or transnational challenges"*[2]. It encompasses consensual norms, rules, and frameworks—both hard and soft— spanning bilateral, regional and universal levels, designed to serve a publicly legitimized common purpose beyond the narrow interests of their creators.

While, transnational regimes of private actors may play a role in global governance, government meetings will remain primary, albeit increasingly incorporating non-governmental stakeholder input. Policies and frameworks from state-mandated regional and international organizations should be recognized as integral to this sphere.

Global security issues, as approached in this report in relation to the use of information and communication technologies (ICTs), represent a challenge which historically and significantly have been addressed through the instruments of global governance.

While not strictly part of global governance, this report examines key domestic policies and regulations shaping AI policy. As hubs for AI companies, capabilities and markets, they influence other domestic and regional frameworks and international norm-setting.

Report development process

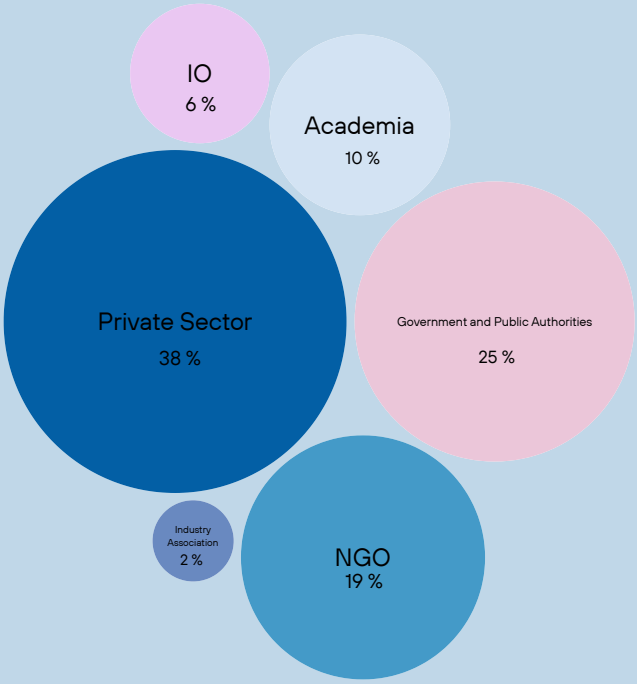
The report was conceived further to two events organized by the Paris Peace Forum: a high-level roundtable session on emerging threats and cutting-edge risks for cyberspace stability at its 2023 edition, and a high-level roundtable on the benefits and risks of using AI in cybersecurity, convened on the sidelines of the 79th UN General Assembly in September 2024.

In November 2024, the **Paris Call Strategic and Foresight Hub** was launched to develop this report. With a focus on the AI-Cyber nexus, this Hub brought practitioners from leading cybersecurity companies, alongside cyber policy and AI governance experts from the public sector, international organizations, and civil society. Their reflections—especially those shared through five dedicated workshops between December 2024 and February 2025—formed a key foundation for this report.

[2] Michael Zürn, "A Theory of Global Governance: Authority, Legitimacy, and Contestation", Oxford University Press (2018), pp. 3-5

This report is further supported with insights gained from a **consultation conducted in the Paris Call community**, collecting perspectives from the multi-stakeholder ecosystem on adversarial AI use risks, particularly AI-driven cyber threats. Respondents were also invited to reflect on mechanisms and frameworks to address these challenges at the international level.

Consultation of the Paris Call community

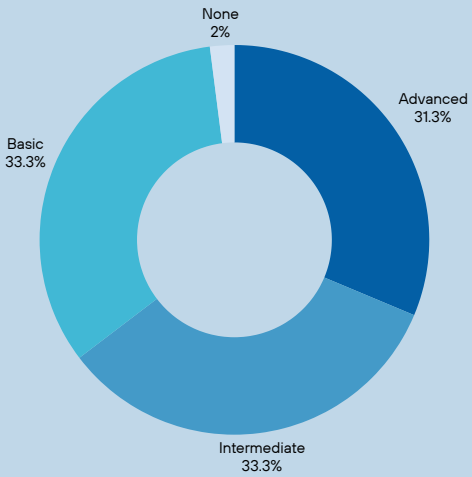


The survey was conducted over a four-week period between January and February 2025, collecting responses from **50 stakeholders**, including representatives from government/public authorities—primarily from diplomatic circles—as well as academia, non-governmental organizations, and the private sector.

20 countries are represented, with a predominance from Europe and North America.

Respondents have a broad range of expertise in using AI: 31% of respondents considered that they had advanced expertise of AI (designing or developing AI), 33% considered they had intermediate expertise (working with Machine Learning or AI applications), and 33% considered that they had basic expertise (use of prebuilt AI). Only 2% considered they had no expertise in AI.

The findings from this survey provide valuable perspectives, highlighting challenges, and potential recommendations for future action. Key results, supported by data visualizations and direct insights from respondents, are included in the darker bands throughout the Report.



Level of expertise in using AI

Section 1 - Shaping the global governance of AI severe risks: insights from two decades of cyber policy

A. Importance of addressing certain AI risks through global governance

The case for AI governance, similar to other ICT challenges, is grounded in the technology's fundamental characteristics, which, while not universally reliant on connectivity, often enable applications with significant potential for transnational externalities. Actions within one jurisdiction—whether by adopting specific policies, deploying AI systems, or developing particular applications—can impact other jurisdictions. This underscores the profound interdependence among states, private entities, and end-users, and the need for coordination to align diverse stakeholders' interests across countries and to manage widespread effects.

Managing crises and severe global risks, including emerging ones, often serves as a catalyst for international action. Preparedness and resilience have become key concepts in international fora. The 2023 Bletchley Park Summit —the first in the series of international summits on AI to which the AI Action Summit belongs—appeared to reflect this by focusing on AI safety risks[3].

These two dimensions on their own are not sufficient to understand the decision of stakeholders to agree on addressing certain challenges at a supranational level. Global governance is costly and slow, making it impractical to address all cross-border effects where more immediate responsiveness is required. **Consequently, it not always practicable nor appropriate to address all dimensions through global governance agreements**[4]. This is particularly pronounced with new technologies like AI, where the fast pace of progress challenges the ability to design and implement frameworks that keep up with innovation. Thus, various factors are considered, including concerns about regulatory arbitrage, disparities in stakeholders' governance capacities, and interoperability[5], and/or value alignment, balancing compliance and accountability, and empowering rather than constraining actors[6].

Identifying and prioritizing potential policy interventions is complex due to limited scientific evidence - evidence dilemma[7] – on AI risks, compounded by insufficient information sharing, evolving vulnerabilities at various stages of development and deployment, and context specific capabilities[8].

[3] Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023
[4] See, in this regard: Stiglitz J.E., Rodrik D., "Rethinking Global Governance: Cooperation in a World of Power" (June 2024)
[5] Dennis, C. et al., "What Should be Internationalized in AI Governance?", Oxford Martin AI Governance Initiative (November 2024)
[6] Lagrange E. et al. "Global governance and multilateralism ", White Paper n° 13 for the 150 years of ILA, International Law Association (September 2022)
[7] Supra Note 1, p.14
[8] Supra Note 1, p. 22, p.36, p.37

B. Current state of play of global efforts on AI risks

Currently multiple initiatives and institutions address global AI governance. **The Council of Europe AI Framework Convention** (September 2024) is notable as the first legally binding international treaty on AI. While its endorsement is limited so far, it includes key signatories[9] and focuses on preventing harm, ensuring fairness, and upholding human dignity.

The United Nations General Assembly has adopted resolutions over the last year to ensure the safe use of AI for both civilian and military purposes[10], consistent with human rights law and international peace and security principles. The **Global Digital Compact**—adopted in September 2024—offers an action plan to create a scientifically grounded and interoperable AI governance landscape that promotes transparency, accountability, and human oversight.

The **European Union (EU) AI Act**[11], a landmark regional law with global impact, sets a precedent with its extraterritorial reach and has a comprehensive, multifaceted approach to risk classification, It covers both general-purpose AI models and the systems built on them—albeit through distinct regulatory pathways. The Act offers a template for balancing innovation with oversight and risk mitigation, however, its complexity raises concerns about effective implementation.

Several other high-profile instruments guide AI governance. The **OECD's AI Principles** (2019, updated in 2024), endorsed by 47 countries and the European Union, remain voluntary yet influential emphasizing AI system robustness, security, and safety of AI systems under various conditions of use – normal, foreseeable or misuse, or other adverse conditions. Similarly, **UNESCO's Recommendation on the Ethics of Artificial Intelligence** (2021), though non-binding serves as a values-based reference, addressing risks of harms to individuals, communities and societies – including from malicious uses.

The G20 adopted AI Principles (2019) and the G7's Hiroshima Summit (2023) launched the ministerial-level **Hiroshima AI Process**, resulting in the non-binding Hiroshima AI Comprehensive Policy Framework. This includes the International Guiding Principles and International Code of Conduct for Organizations Developing Advanced AI Systems. The Framework focuses on addressing vulnerabilities, mitigating misuse risks, ensuring transparency, fostering collaboration and information sharing, and advancing research to tackle societal, safety, and security challenges.

Several AI Summits have focused on safety, security, and trust, producing non-binding declarations and ministerial statements—such as the **UK Bletchley Park AI Safety Summit** (2023) and the **AI Seoul Summit** (2024).

[9]The United Kingdom and the United States are among the first signatories of the Convention. The European Commission has also signed it; however, its implementation across the 27 EU Member States—via the EU AI Act—remains contingent on the Council's decision to conclude the Convention and the European Parliament's consent. See: Council of Europe and Artificial Intelligence
[10] UN General Assembly, Artificial intelligence in the military domain and its implications for international peace and security, Res. 79/239, UN Doc A/RES/79/239 (31 December 2024)
[11] Regulation (EU) 2024/1689 of the European Parliament and of the Council "AI Act", Doc. 32024R1689 (13 June 2024)

These gatherings[12] spurred the creation of national AI Institutes (AISI) and government-mandated offices in several countries to study AI risks and guide evidence-based governance. In November 2024, representatives from nine countries and the EU convened the first meeting of the International Network of AI Safety Institutes, to enhance cooperation, research and risk mitigation.

In July 2024, the **African Union adopted a Continental AI Strategy** promoting risk-based regulations, security and transparency. It also emphasizes the importance AI safety and security research, capacity building, and the development of robust standards for accountability and system integrity.

It is also worth noting the World Economic Forum's AI Governance Alliance, supported by 463 organizations from the public and private sectors, as a promising platform for multistakeholder cooperation on cross-borders AI challenges, with the potential of informing international negotiations on safety and security standards.

A deeper understanding of the landscape can inform an inclusive, multilateral, and multistakeholder approach to AI governance—across its lifecycle from design, development, deployment, to use.

Panorama of international instruments addressing AI risks

[12] The November 2023 and May 2024 Summits collectively brought together 50 organizations from academia and civil society, 29 governments, 47 industry-related organizations, and 6 multilateral organizations, plus the European Union.

Recent reports[13] highlight ongoing debate among experts on the best structure for international AI governance and compliance oversight. What remains clear is the urgent need for global AI rules and standards, especially for generative AI. Proposals range from a dedicated AI governance body to models inspired by the International Atomic Energy Agency (IAEA)[14], or Intergovernmental Panel on Climate Change (IPCC)[15].

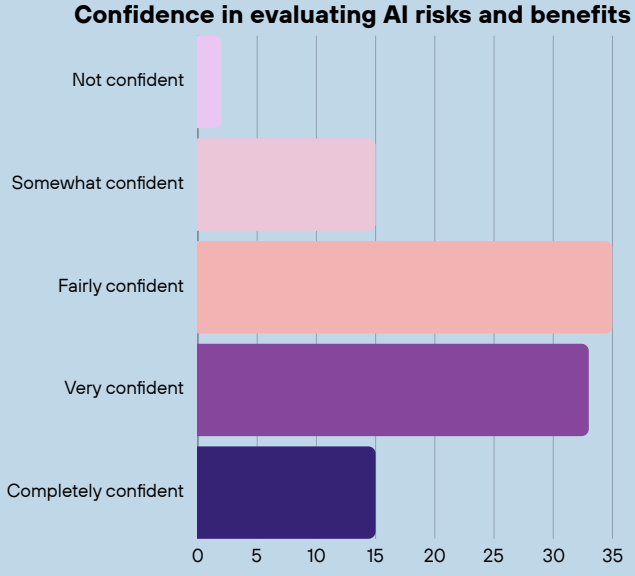
To date, global AI governance has largely relied on non-binding principles or codes of conduct without a common foundation. The UN's Governing AI for Humanity (HLAB-AI) Report highlights a serious participation gap in all major AI initiatives. Reviewing seven non-United Nations AI initiatives, it notes that seven countries are parties to all the sampled governance efforts, whereas 118 countries are parties to none.[16] This exclusion is critical, as underdeveloped countries - often most affected by AI risks – are left without a voice, mirroring climate change's disproportionate impact on vulnerable regions.

C. Parallels between global governance of AI risks and international cyber policy

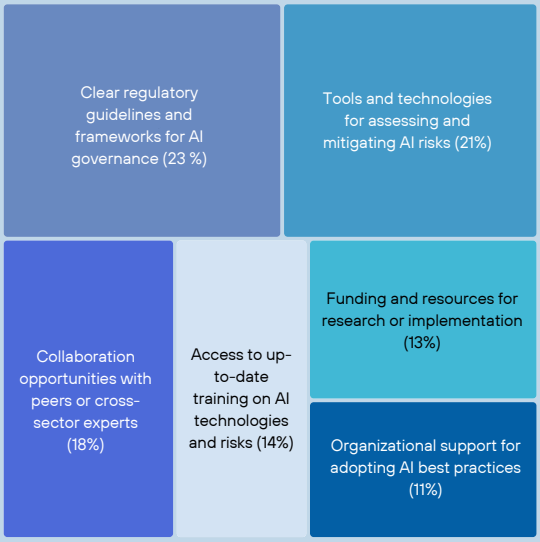
Recent debates on suitable international AI risk governance often draw on more mature global governance models. **Cyber policy stands out as a particularly relevant reference in this context**[17]. Unlike aviation, chemical or nuclear governance, with which parallels are often drawn, cyber policy shares AI's rapid innovation and foundational similarities (infrastructural, logical, or informational). This is reflected in the growing interest from cyber policy and technical experts —as seen in the consultation of the Paris Call community.

The consultation sought to identify the awareness, expertise, and confidence levels of respondents in relation to AI and AI risks, and the basis upon which this was determined.

The vast majority of consultation respondents (83%) express at least a 'fairly confident' stance in evaluating AI risks and benefits, with 33% being 'very confident' and 15% 'completely confident'. Only a small percentage (2%) report having no confidence at all.



[13]See, in particular: G'Sell F., Regulating Under Uncertainty: Governance Options for Generative AI, Stanford Cyber Policy Center, Freeman Spogli Institute, Stanford Law School, p.407
[14] Gambrell J., OpenAI CEO suggests international agency like UN's nuclear watchdog could oversee AI, Associated Press (June 2023)
[15] Suleyman M., Schmidt E., We need an AI equivalent of the IPCC, Tribune, Financial Times (October 2023)
[16] See: High-Level Advisory Body on AI (UN HLAB-AI), Governing AI for Humanity: Final Report, United Nations (September 2024), p.75
[17] See, in this regard: Morse J., "Frameworks and Outcomes for International AI Governance," Global Governance: Goals and Lessons for AI, Microsoft Publications (2024): "In some cases, participants offered cybersecurity as an area worth considering given international governance efforts and a perception of mixed results".



Resources lacking for cybersecurity specialists to adress AI-driven cyber risks

Among respondents identified as specialized ICT security practitioners, **the most commonly cited missing resource is clear regulatory guidelines for AI governance (32%)**, followed by tools for assessing and mitigating AI risks (29%). Collaboration opportunities (24%) and training on AI technologies (19%) are also key gaps, while funding for research (17%) and organizational support for AI best practices (14%) remain notable concerns.

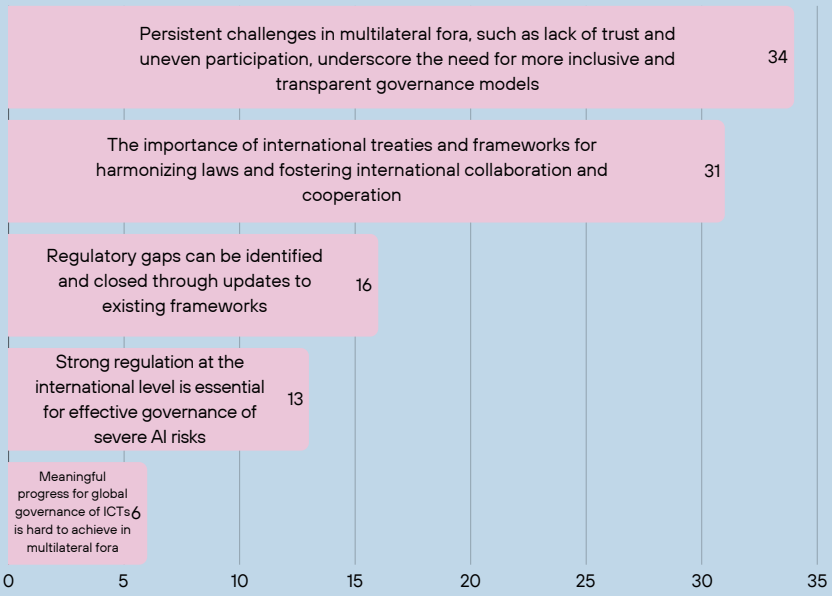
Whilst global cybersecurity policy lacks dedicated formal institutions, its governance processes offer valuable insights for governing adversarial AI use risks, particularly in the methods and processes for building a broad international consensus, impactful agreements, and ultimately paving the way for effective governance mechanisms and frameworks. The next section examines key achievements and challenges in international cybersecurity governance, laying a foundation to inform global governance of AI severe risks.

The pursuit of cybersecurity governance provides valuable lessons and frameworks that can inform the pursuit of global governance of AI severe risks. In this regard, findings from the Paris Call community consultation are insightful.

Lessons from international cyber policy can inform the global governance of AI risks

Respondents mostly emphasized that **persistent challenges in multilateral fora on cyber issues, including the lack of trust and uneven participation,** should be considered in future efforts to govern AI risks globally. These challenges underscore the need for more inclusive and transparent governance models.

Additionally, respondents recognized the **importance of international treaties and frameworks for harmonizing laws and fostering international collaborations.**



i. Achievements to build upon

a. Building global trust through consensus-driven structures

The convening power of international organizations, forums and initiatives are critical in cybersecurity governance, fostering consensus, collaboration and global norms and standards.

The development and implementation of shared principles, norms, rules, and decision-making processes that have shaped the evolution and use of the internet involve a complex, multistakeholder ecosystem that includes governments, international organizations, private sector entities, civil society, and technical communities. Unlike traditional governance models, internet governance is decentralized and relies on collaborative mechanisms to address key issues and Institutions like the Internet Governance Forum (IGF), the International Telecommunication Union (ITU), and the Internet Corporation for Assigned Names and Numbers (ICANN) play crucial roles in facilitating dialogue and coordination. The multistakeholder approach has enabled the internet to remain open and interoperable while also posing challenges in balancing regulatory oversight with innovation and rights protection.

Since the late 1990's the UN General Assembly has addressed technology's role in international security, with the First Committee and Third Committees tackling threats and misuse. Six groups UN groups of governmental experts[18] (GGE) have considered existing and potential threats in information security. The UN GGE[19] and the Open-ended Working Group[20] (OEWG) have played pivotal roles in convening Member States to discuss responsible behaviour in cyberspace, including with regard to the 11 voluntary norms for responsible behaviour in cyberspace, which identified key areas of consensus such as the protection of critical infrastructure and the explicit inclusion of health care in response to the cyberattacks during the COVID-19 pandemic[21].

The Council of Europe Convention on Countering Cybercrime (Budapest Convention) is a binding international framework against cybercrime, open to both Council of Europe and non-member States, with 76 parties. The 2024 UN Cybercrime Convention, negotiated under UN auspices, offered broader inclusiveness. Cybercrime laws at national, regional and international levels vary in maturity but continue to evolve, reflecting a clearer understanding of ICT use and misuse.

Inclusive, consensus-driven platforms serve as Confidence-Building Measures (CBM) in the cyber governance domain, and play a role as mechanisms for fostering trust and risk mitigation. The Organization for Security and Co-operation in Europe (OSCE), the UN GGE and OEWG promote transparency, collaboration, and capacity building, and for designated points of contact, dialogue, and information sharing.

[18] UN Group of Governmental Experts 2003-2004, 2009-2010, 2011-2013, 2014-2015, 2016-2017, 2019-2021.
[19] The UN Group of Governmental Experts on advancing responsible State behaviour in cyberspace in the context of international security (GGE) agreed norms in 2010, 2013, 2015, and 2021, herewith: https://www.un.org/disarmament/group-of-governmental-experts/
[20] OEWG, 2021 report, United Nations, UN Doc. A/75/816 (18 March 2021)
[21] Ibid.

The OSCE's 2016 initiative[22] highlighted regular consultations to ease ICT tensions, leveraging its forums to exchange good practices and prevent conflict escalation.

The UN GGE has similarly encouraged sustained interaction at multiple levels, supported by stakeholders from the private sector, academia, civil society, and technical communities[23]. Despite varying state capacities, self-reporting is seen as vital in this opaque domain, given the challenges in assessing and monitoring national capabilities. As the technical community has long led incident management standards, some observers view recent multilateral CBM processes as reiterations of established practices[24]. Nevertheless, there is broad agreement of the political value of uniting states under these platforms to reduce uncertainties and misunderstandings that could otherwise escalate into instability.

These platforms highlight the importance of fostering dialogue in building consensus, setting norms and mitigating risks. **Given the need for a *universal and networked approach* to international AI governance**[25]**, analyzing cyber fora dynamics —along with the drivers, incentives, and costs of achieving broad consensus—can offer key lessons**. Upcoming UN negotiations on a permanent mechanism for states to discuss the use of ICT (commonly referred to as the Cyber Programme of Action) may also provide valuable insights.

International consensus, even if fragile, often depends on how issues are framed within the mandates of global processes. Cyber policy shows that dialogue is somewhat easier to foster on technology use and security than on controlling innovation from an ethical standpoint. Current UN momentum on the use of AI systems in military contexts, driven by peace and security imperatives, reinforces this approach's effectiveness in mobilizing support for action related to AI.

> ***"AI is the new frontier, cooperation between countries and across sectors should be at the fore front. Transparency and trust are two fundamental ingredients of a binding AI governance. Fostering cooperation, instead of the general mistrust that is currently ongoing will be a significant challenge for policymakers, AI researchers and the private sector. Caution not to alienate: compromises and levelling the playing field are also important lessons from past challenges in the realm of cyber."***
> Respondant n°12, Public authority, France

[22] OSCE Permanent Council Decision No. 1202, Doc. PC.DEC/1202 (10 March 2016)
[23] UN General Assembly, Res. 76/135, UN Doc. A/RES/76/135 (16 December 2021)
[24] Puscas I., Confidence-Building Measures for Artificial Intelligence: A Multilateral Perspective, UNIDIR (July 2024)
[25] See for example, Guiding Principle 4, HLAB-AI. Governing AI for Humanity: Final Report. United Nations (September 2024), p. 38

Given the current information gaps on the inner workings of AI, and the speed of its development and applications, **addressing the major categories of AI risks requires a robust international scientific consensus to guide stakeholders and engagement from states around common priorities.** The International Scientific Report on the Safety of Advanced AI aids this effort by outlining diverse threats that, at first glance, all seem eligible for consideration under global governance, although they are significantly broad, numerous, and diverse. The UN Global Digital Compact (2024) further supports this by proposing an International Scientific Panel on AI under UN auspices, to synthesize research and identify knowledge gaps[26].

b. Applicability of international law to ICT use

The UN GGE focused on (cyber) security-related aspects in the digital space and the applicable provisions under international law, with consensus that international law, and particularly the UN Charter in its entirety, applies to cyberspace[27]. This was also unanimously approved by the UN General Assembly[28], and reaffirmed by the OEWG in 2021[29].

Following an OEWG recommendation, around 30 states and 2 regional organizations have submitted positions clarifying how and when international law applies to cyberspace in concrete terms, such as in relation to state sovereignty, due diligence, peaceful settlement of disputes, prohibition of intervention, prohibition on use of force and on right of self-defence, etc.[30], as well as the application of International Humanitarian Law (IHL) and human rights. There have been diverging positions of states in relation to the interpretation and application - how and when IHL applies to, and therefore limits, the use of ICTs during armed conflict. International Human Rights Law, including the Universal Declaration on Human Rights and the International Covenant on Civil and Political Rights, apply to digital space as affirmed by a Human Rights Council Resolution[31] stipulating that the same rights that people have offline must be protected online. However, some States diverge on rights affected by cyber-related activities, such as an individual's right of access to information, privacy, or freedom of expression.

The broad applicability of the core principles of international law to AI, in general terms, appears to be undisputed, as evidenced by recent near-universal instruments adopted[32]. **As a technology-neutral corpus of rules, international law is unlikely to be radically challenged by the increasing scope and diversification of AI applications, provided that these discussions remain focused on its *use*.**

For global, interoperable AI risk governance, reaffirming international law's relevance is crucial. Efforts must advance on interpreting and enforcing international law in the context of rapid technological evolution. As noted by UN HLAB-AI, this fosters inclusive, consensus-driven dialogue and promotes *a global race to the top* in AI governance rather than regulatory competition[33].

[26] UN General Assembly, Global Digital Compact, Res. 79/1, UN Doc. A/RES/79/1 (22 September 2024), Annex I, p. 49
[27] See: GGE2013 Report, United Nations, UN Doc. A/68/98, para. 19; 2015 report: N Doc. A/70/174, para. 24, para. 28 c.
[28] UN General Assembly, Res. 70/237, UN Doc. A/RES/70/237 (23 December 2015)
[29] Supra note 20, para.8
[30] Supra note 27
[31] UN General Assembly, Human Rights Council on the promotion, protection and enjoyment of human rights on the Internet, UN Doc. A/HRC/20/L.13 (June 2012)
[32] UN General Assembly, Res. 79/239, UN Doc A/RES/79/239 (24 December 2024) ; Res. 78/265, UN Doc. A/RES/78/265 (21 March 2024)
[33] Supra note 25, p. 54

Integrating the international law dimension into efforts to identify risks associated with adversarial AI use, assess their impact, and consequently their severity is key to managing its lifecycle and establishing widely adoptable meta- evaluation standards, across all governance levels.

c. Growing Inclusion of non-governmental stakeholders in Policymaking

The growing role of non-governmental stakeholders - private sector, academia, civil society, and the technical community - has been crucial for understanding the cybersecurity complexities and challenges that policies must address. They provide expertise, align draft texts with human rights and standards, propose safeguards, foster dialogue, build capacity, conduct research, and amplify marginalized voices. Importantly, private sector involvement is crucial as it owns and controls most cyber infrastructure, provides critical insights into the threat landscape, practical implementation challenges of proposed policy measures and in advancements in technology. These companies leveraging their knowledge, expertise, and resources to support research in key areas of ICT security.

As an UN-convened platform, the Internet Governance Forum (IGF) fosters inclusive, multistakeholder dialogue on digital policy amongst governments, private sector actors, civil society, and technical experts. The IGF shapes global discussions and informs policymaking including on digital risks and challenges as its recommendations are transmitted to global and national decision-making bodies. Its role was emphasized during negotiations for the Pact for the Future, especially its role in ensuring broad stakeholder participation.

A non-paper[34] to the OEWG 2021-2025 highlights the importance of stakeholder involvement in strengthening UN-led cyber discussions and the need for dedicated resources for capacity building and technical assistance. It emphasizes key areas such as capacity building through training programs, simulations and toolkits for implementing the Framework on Responsible State Behaviour; research to develop policy recommendations; activities to foster confidence-building; public awareness on cybersecurity best practices. It suggests active participation in drafting voluntary principles, improving coordination and consolidation of feedback, creating a global stakeholder directory, and engaging in expert-level technical meetings.

Despite growing participation, non-governmental stakeholders face structural, political, and practical barriers to participating in policymaking processes, such as the UN OEWG and AHC. These State-led processes limit their role to observers, with restrictions through procedural rules and limited consultation mechanisms, and contributions dependent on ad hoc or informal arrangements. Some states resist including private sector or civil society organizations due to differing views on multi-stakeholder engagement, national security concerns, or a preference for closed discussions. Private sector involvement also raises concerns over conflicts of interest, advocacy for self-regulation, and Global North influence.

[34] Non paper, Stakeholders Contributing to Multilateral Cybersecurity Discussions, submitted to the UN OEWG, coordinated by Canada and Chile (March 2024)

There is expanding inclusion of an increasingly broader range of non-governmental stakeholders in global AI discussions, notably at the forthcoming AI Action Summit. However, their normative outcomes, along with all other aforementioned instruments related to AI risks, mainly result from siloed endeavours. Collaborations between AI Safety Institutes and industry and academia offer insights into the relationships that can be forged within this emerging ecosystem. These efforts are an area of focus to build upon, combined with lessons learned from the cyber realm, to shape a stronger efficient multistakeholder ecosystem. To enhance global policymaking, such efforts should be mainstreamed within the same layer of governance, potentially under the international network of AI Safety Institutes[35], and transposed within multilateral fora.

"Multistakeholder governance models which meaningfully include stakeholders who have various areas of "effective control", including on the development, deployment, oversight and use of these technologies are in my opinion the only way to mitigate important AI risks as well as realize important AI opportunities."
Respondant n°40, NGO, Switzerland

d. Enhancing operational cooperation through targeted formats

Operational cooperation in cybersecurity shows that focused formats enable stakeholders to address specific risks and challenges. Targeted cooperation with international partners, competence centers, and specialist organizations are key to implement measures to protect against cyber threats. States leverage international networks, including bilateral ties, expert bodies, technical competence centers, and other strategic partners such as the Forum of Incident Response and Security Teams, (FIRST), the global network of cybersecurity teams providing rapid response; the Task Force on Computer Security Incident Response Teams (TF-CSIRT)[36] and National Computer Emergency Response Teams (CERTs).

The EU's cybersecurity governance offers valuable models for AI safety governance, such as regulatory sandboxes creating controlled environments to test AI systems against safety requirements before deployment. Similarly, the European Cybersecurity Competence Centre[37] (ECCC) combines centralized expertise with national coordination - a model that could be adaptable for AI safety testing infrastructure.

[35] See in this regard: Adan S. et al., Key questions for the International Network of AI Safety Institutes, Commentary, Institute for AI Policy and Strategy (November 2024)
[36] TF-CSIRT: Computer Security Incident Response Teams - GÉANT Community
[37] European Cybersecurity Competence Centre and Network

Operational cooperation is vital for international law enforcement, mutual assistance in combating cybercrime, including operational programs such as the International Counter Ransomware Initiative (CRI)[38]. With 68 members, CRI enhances collective resilience, supports members faced with an attack, pursues ransomware actors and forges international partnerships. It includes a Policy Pillar, a Diplomacy and Capacity Building Pillar, and a Taskforce on international cooperation and governance of information sharing platforms. INTERPOL's Cyber Fusion Centre (CFC) unites law enforcement and industry experts to analyse cybercrime information and provide actionable intelligence. Since 2017, it has issued over 800 reports to police in 150+ countries.[39]

Mutual cooperation and assistance enhances understanding of the threat landscape, disrupts cross-border cybercrime, and supports international technical collaboration on issues like OT security and phishing. This operationalizes a norm of responsible behaviour in cyberspace, recognizing cybercrime's transnational threat to international security. Norm D of the 11 UN norms emphasizes international cooperation to address criminal and terrorist use of ICTs and affirms that: States should consider how best to cooperate to exchange information, assist each other, prosecute terrorist and criminal use of ICTs and implement other cooperative measures to address such threats.[40] States are urged to strengthen information exchange and assistance mechanisms to curb online terrorist and criminal activities. Member States affirm the need for global cooperation, - particularly investigation and prosecutorial activities of law enforcement and judicial authorities – and a focus on investigative mechanisms, electronic evidence handling, and legal resources. Cooperation channels, including Computer Emergency Response Team (CERT), Computer Security Incident Response Team (CSIRT) and Mutual Legal Assistance Treaty (MLAT), establish reciprocal obligations to provide legal assistance for specific transnational crimes.

Expanding successful operational collaboration can help develop effective AI incident detection and response mechanisms for adversarial uses of AI systems. This requires shifting from industry-led efforts to full engagement of public authorities and key stakeholders in application sectors at risk of AI misuse.

ii. Challenges to keep in focus

a. Slowness in achieving substantial results at the multilateral level

Global governance progress on responsible behavior in cyberspace has been slow, with complex negotiations, fragmented priorities, and competing national interests, especially in UN processes like the OEWG. This has led to concerns over efficiency, efficacy and that progress is not keeping pace with technological advances.

[38] Home | International Counter Ransomware Initiative
[39] Cybercrime threat response
[40] Supra note 28, para.13 (d)

Long, drawn-out processes that take years to produce tangible outcomes significantly hinder national cybersecurity responses. Prolonged negotiations often result in voluntary, non-binding norms that lack enforcement mechanisms, limiting their real-world impact. Sovereignty concerns and geopolitical tensions have further slowed progress toward global agreements, while informal, state-only groups reduce transparency and exclude expert contributions from non-governmental stakeholders. This exclusion weakens implementation, as critical technical and operational expertise is left out of decision-making.

Protracted multilateral processes, often dominated by a few states, restrict the participation of less-resourced countries and non-governmental stakeholders. These negotiations favor actors with the financial and technical means to sustain prolonged engagement, creating an imbalance in representation. Limited diplomatic, financial, and expert capacity prevents many from fully engaging in complex discussions and navigating numerous preparatory sessions. As a result, perspectives from smaller states and non-governmental entities risk being sidelined, reducing the inclusivity of global policymaking.

A careful, timely evaluation is needed to weigh the benefits of international cooperation on adversarial uses of AI against its costs, particularly given the rapid pace of technological innovation globally. This includes determining the appropriate level of internationalization—whether bilateral, plurilateral, multilateral, or near-universal.

b. Struggles in prioritizing risks and threats in in a fluid environment

Evolving technology, political and social developments influence the threats situation. Risk prioritization is complicated by geopolitical tensions, including a more polarized global order and ongoing armed conflicts, alongside the growing diversity of threats, the interconnectedness of supply chains, and emerging technologies.

Concerns over the increasing sophistication and diverse threat landscape complicates the task of resource allocation and investments. Technological developments improve security but also create new dependencies, increased complexity, and lead to new threats. While "new" AI-related risks are attracting significant attention, traditional threats like ransomware, fraud, supply chain disruptions, identity theft, and DDoS attacks—remain the most prevalent[41]. Overemphasis on emerging risks diverts focus from these more likely threats. Conventional risks remain the most prevalent exacerbated by weak cyber hygiene enforcement. This is compounded by a widening gap for cyber skills – and a need for increasingly specialised skills - further complicating the ability to manage risks effectively.

The "zero-day market", where unpatched vulnerabilities are sold instead of responsibly disclosed, presents significant challenges to cybersecurity and to the safety, trust and resilience of ICTs. The lack of more harmonized and coordinated vulnerability disclosure and bug bounty programs further undermines global cybersecurity.

[41] See for example, World Economic Forum Global Cybersecurity Outlook, Insight Report 2025,(January 2025)

The widespread excitement around AI and the proliferation of standards and initiatives, highlights the urgent need to prioritize threat-related items within a coherent international agenda. This is complicated by states advocating for multilateral fora to address issues linked to initiatives they are involved in, support, or that originate from organizations based on their territory, and the accelerating pace of AI development and use.

c. Normative fragmentation and interoperability challenges among frameworks

Cyberspace regulation – initially seen as a politically open space with governance limited to its technical architecture – has evolved progressively as governments, international organizations and regulatory agencies recognized that laws and regulations governing the physical world also apply online. Driven by technology's global nature, misuse risks, trust, rights, privacy and security concerns, governance now involves governments, international organizations, and a growing multistakeholder community, including industry, shaping development and deployment of technology today.

Internationally, over the last decades, there have been differing interpretations of how international law applies to cyberspace, a lack of global support for key areas of human rights, the development (although not enforcement) of norms, and adaptations to existing or new legal frameworks, regulations and standards. As a result, laws and regulations have become more fragmented in a policy space that is increasingly transnational and cross-sectoral. Fragmentation is further driven by the rise of [geopolitical] regulatory standards aimed at strengthening their application and/or to address gaps, often driven by varying views on the relationship between the state and businesses.[42] This includes the proliferation of standards being adopted by the International Telecommunications Union (ITU), International Organization for Standardization (ISO), the International Electrotechnical Commission (IEC) and the Institute of Electronic Engineers (IEEE), developed without a common language and agreed definitions of terms.

Global technological innovation has led to fragmented and overlapping regulations, standards, and policies - that lack a common minimum baseline and are potentially conflicting or duplicative[43]. While regulation is fundamental, the changing regulatory landscape and lack of harmonization across jurisdictions creates compliance challenges, increasing burdens on organizations to navigate a complex landscape of overlapping requirements and enforcement timelines[44]. This complicates efforts to uphold consistent cybersecurity standards.

Having specialized processes can limit flexibility by creating rigid structures that struggle to adapt to emerging threats and technological advancements. While specialization allows for in-depth expertise, it can lead to siloed approaches, where different entities or initiatives work in parallel without effective coordination.

[42] See: Cavelty M., Wenger A. (Ed.), Cyber Security Politics Socio-Technological Transformations and Political Fragmentation, Routledge (2022), p. 250
[43] See for example, Dennis, C. et al., "What Should be Internationalized in AI Governance?", Oxford Martin AI Governance Initiative (2024)
[44] See for example: Morse J., "Frameworks and Outcomes for International AI Governance," Global Governance: Goals and Lessons for AI, Microsoft Publications (2024)

This fragmentation indeed increases the risk of duplication, where multiple efforts address similar challenges without leveraging shared insights or resources. **As a result, efficiency declines and critical gaps in governance may persist, hindering a more agile and comprehensive response to cybersecurity and AI-related risks.**

> ***"Current cybersecurity frameworks lack adequate AI controls and there is a lack of standardized and unified assessment and maturity framework for AI risk management. Current AI regulations are fragmented and ineffective as the technology is outpacing policymakers understanding of the protections required."***
> Respondant n°4, NGO, United States

Industries like aviation and healthcare have sector-specific regulations reflecting the critical nature and distinct challenges posed by the sector. In other industries, cybersecurity's growing regulatory landscape can lead to compliance fatigue and unintended non-compliance. As rules evolve to address emerging threats, organizations struggle to balance compliance costs with non-compliance risks across an increasing number of jurisdictions. Regulatory risk exposure varies by industry, geographical location, the nature of their products and services, and the stringency of cybersecurity enforcement by regulators.

The current geopolitical climate, marked by a relative reduction of interconnectedness and collaboration among states – hinders a more unified and cohesive global response to cybersecurity – and thus AI – challenges, increasing fragmentation, unequal progress, and increased risks of misuse. **Addressing this requires innovative approaches to multilateralism and trust-building, dynamic and adaptive regulations to challenges, threats and opportunities, and the provision of guidelines to facilitate understanding and guidance**. As highlighted previously, most respondents to the Paris Call Consultation felt that there was not enough regulation and oversight in AI development and deployment within their jurisdiction, it will be important to identify dynamic measures that can be taken.

> ***"The risks of ex ante regulation outweigh the benefits, but since ex ante is irresistible, be sure to build in sunset clauses and mandatory review and updating requirements."***
> Respondant n°8, Academia, United States

d. Shortcomings in enforcement and accountability

Enforcement and accountability gaps weaken effective global governance on ICT security, with fragmented frameworks reducing effectiveness. Cross-border cybercrime prosecution is, for instance hindered by legal differences and weak coordination, making greater alignment, collaboration and political will essential for meaningful accountability.

These challenges are further exacerbated by lack of transparency, which can take several forms. This can range from organizations not publicly disclosing ongoing activities limiting public awareness and engagement; underreporting of breaches and failures to disclose vulnerabilities limiting effective countermeasures; opacity by states including a failure to share information on their actions, (where this may not be a formal obligation it still affects cooperation); to a general lack of transparency with regard to disclosure of information which hinders accountability and informed decision-making. While efforts to mandate reporting standards are underway, progress remains inconsistent and fragmented and thus limits the effectiveness of enforcement mechanisms.

The lack of progress in implementing cyber norms has spurred calls for accountability. In 2023, UN Secretary-General António Guterres proposed an independent multilateral accountability mechanism for malicious use of cyberspace to reduce incentives for such conduct. This mechanism could enhance compliance with agreed norms and principles of responsible State behavior[45].

The ability to accurately identify the source of a threat is essential for effective accountability in cases of adversarial risks. However, in a geopolitically complex and fluid environment, cyberattack attribution is particularly challenging due to technical, legal, and political complexities. Yet, attribution is crucial to determine the applicable legal framework and the appropriate legal and political response. Additionally, the limited frequency of public attributions by governments for malicious cyber activities, weakens deterrence, as they can serve to highlight unacceptable behaviour and clarify the specific rules violated.

AI accountability is reminiscent of the foundational challenges seen in cyber governance. Just as effective cyber accountability relies on clear frameworks defining primary and second obligations, whether biding or not, ensuring the safe and intended operation of AI systems necessitates explicit assignment of liability when these systems cause harm. **This challenge becomes particularly complex in the context of adversarial uses of AI. Addressing it might require a comprehensive approach that allocates responsibilities across all relevant stakeholders**.

Policies also could enhance AI incident data collection and establish and facilitate the development of an authoritative classification system for extracting meaningful data and trends on AI harm. This could include designing AI systems to support investigations and data collection. [46] The EU AI Act (Article 73), and the Council of Europe Framework Convention (Article 26) already require reporting of serious incidents for high-risk AI systems**.**

[45] Antonio Guterres, A New Agenda for Peace, Policy Brief 9, United Nations, July 2023, p.27. Supra 31.
[46]See for example, Dixon R., Fraser H., An Argument for Hybrid AI Incident Reporting Lessons Learned from Other Incident Reporting Systems", Center for Security and Emerging Technology (CSET), Issue Brief (March 2024). This report focused on the US with examples from other jurisdictions.

Section 2 - Towards a scalable model for tackling adversarial use of AI in cyber

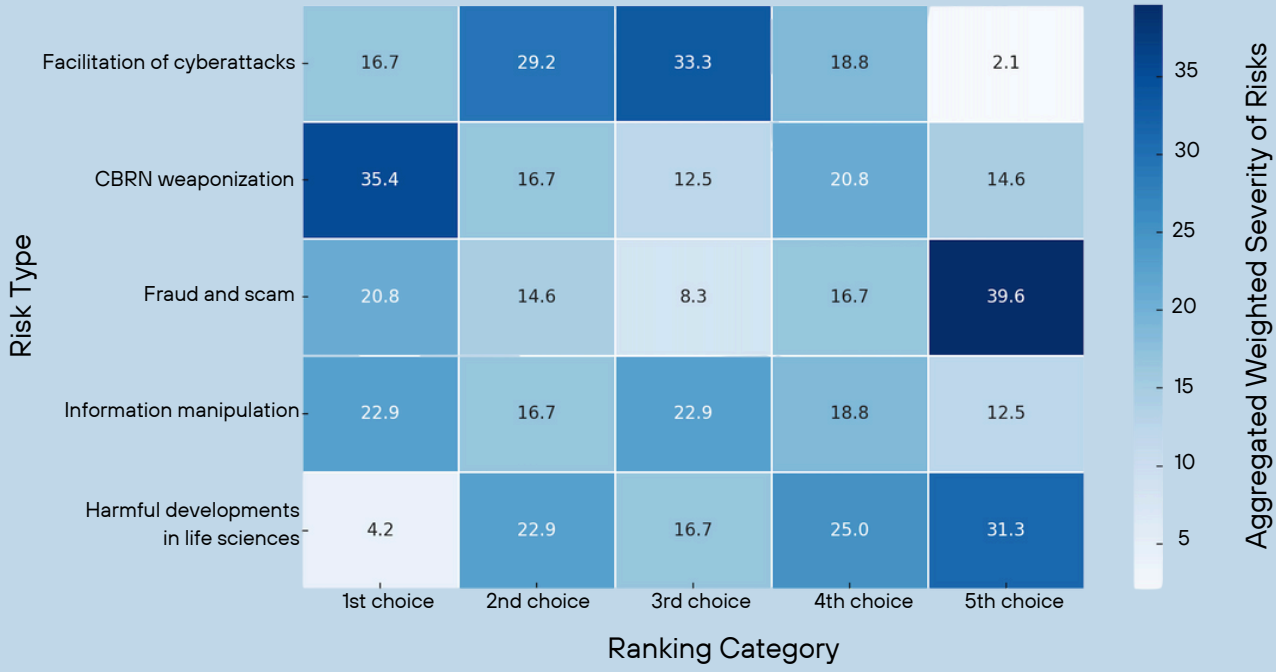
While AI risk governance can be properly distinct from that of cyber risk governance, methodological comparisons remain valuable. Insights from two decades of international cyber policy offer decision makers valuable guidance in shaping practicable and effective global AI risk management mechanisms through global cooperation.

Beyond institutional comparisons, the international community must urgently address emerging threats from the intersection of cyber and AI. The adversarial use of AI has shifted from theory to reality, reshaping the threat environment and challenging existing defenses in ways that require immediate attention and concerted action.

Following a 2023-2024 in -depth assessment, the OECD Expert Group on AI Futures unequivocally identified facilitation of cyberattacks was the most "important" future AI risk among the 38 identified[47]. Among the application areas identified as at risk of adversarial use in the International Scientific Report on AI Safety, cybersecurity stands out, with practitioners viewing the diversion of artificial intelligence as the most transformative factor for the ecosystem in the short term[48].

Relative perceived severity of adversarial AI use risks

"Which of the following risks from the misuse of AI intended to harm do you perceive as most severe?" (From 1 to 5).



[47] OECD, Assessing Potential Future Artificial Intelligence Risks, Benefits and Policy Imperatives, OECD Artificial Intelligence Papers, no. 27 (November 2024), Annex B, p. 42
[48] World Economic Forum, Global Cybersecurity Outlook, Insights Report (January 2025), p. 19

Relative perceived severity of cyber among adversarial AI use risks

The distribution suggests a **broad but not extreme concern**: while some consider it the most pressing issue, **many place it in the second or third tier of risks**, with a relatively low level of polarization. Respondent tend to perceive adversarial cyber use of AI as a **consistently serious but rarely the absolute top threat** - especially when compared to the risk of weaponization in CBRN context.

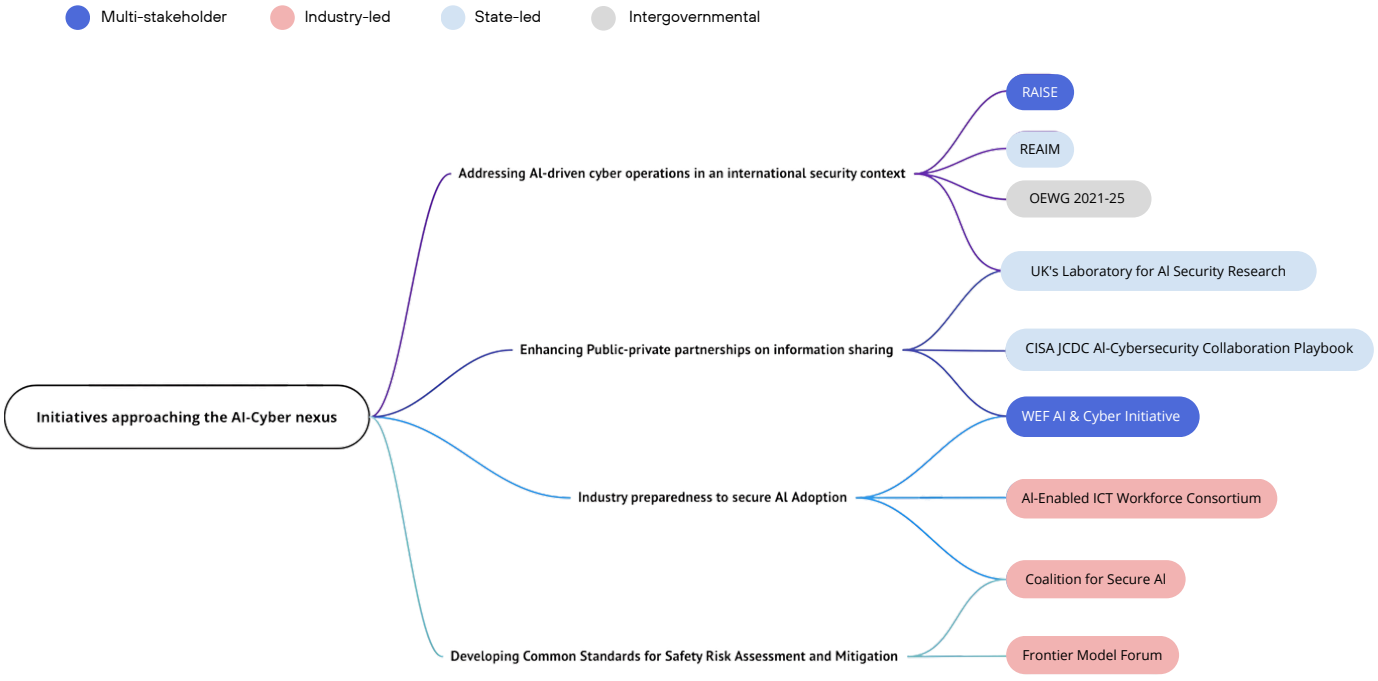
This might also mean that they perceive this risk as **at least partly manageable in the medium term**, despite its far-reaching and systemic impact.

Relative perceived impact of AI use on the future cyber threat landscape

1. **Significantly enhancing the sophistication of cyberattacks, reaching level of distinctive innovations**
2. **Autonomous weaponization / attacks – escaping human control**
3. **Widening the reach / scale of traditional cyberattacks**
4. **Lowering the barrier to entry for novices in conducting offensive operations**

The consultations nuanced insights on AI-driven cyber risks, combined with the growing acknowledgement of the deep interconnection between cyber risks and the systemic importance of maintaining global digital security and stability, highlight the need for a comprehensive, multi-layered approach. Guided by subsidiarity - and recognizing that it is neither practical nor desirable to address the full spectrum of AI-driven cyber risks at the international level - governance should be structured around distinct actions, across multiple layers of governance.

Panorama of current efforts adressing AI-driven cyber risks globally



A. Emerging AI-driven cyber risks: cyber risks before AI risks

At the Paris Call Strategic Foresight Hub, experts emphasized that the use of AI for adversarial purposes is unlikely to fundamentally alter the core nature, and modalities, of cyber risks - at least in the short and medium term.

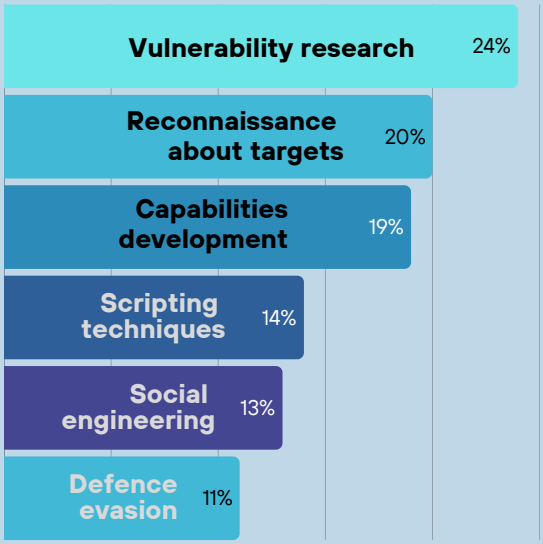
While AI enhances conventional cyberattacks at various stages, no entirely new (novel) cyberattack types have been observed to date, even with the release of large language models[49].

Adversarial intent, targets, and potential resulting damage remain largely unchanged. The main foreseeable factors of disruption in the cyber threat landscape primarily pertain to variables of time and volume:

- the **velocity** of execution of adversarial operations;
- the **frequency** of attacks;
- and the **expansion of the spectrum of adversarial actors** -due to lower entry barriers to engage in offensive activities.

The increasing risk of cyber insecurity and instability from these three vectors **can be significantly mitigated by a mainstream and tailored adoption of AI to bolster cyber defense capabilities in public and private sectors**[50].

Perceived potential of AI across the attack lifecycle



When asked which techniques and tactics were most likely to be facilitated by AI in the context of an adversarial cyber operation, consultation participants primarily identified its role in supporting vulnerability research, reconnaissance, and social engineering efforts, rather than enabling the development of new capabilities or facilitating defense evasion.

These results are therefore consistent with the assessment made by members of the Paris Call Strategic Foresight Hub.

[49] Goemans A. & al. « Safety case template for frontier AI: A cyber inability argument », arXiv preprint arXiv:2411.08088 (November 2024)
[50] See, in this regard: Frontier Model Forum, AI for Cyber Defense, Issue Brief (November 2024)

Evidence-based findings from developers

Recent threat intelligence reports from **Google**[51] **and OpenAI**[52] **concur that adversarial use of their generative AIs, while enabling some productivity gains for threat actors across the cyberattack lifecycle, has not yet led to groundbreaking new malware or significantly enhanced operations**. They observe that it aids common tasks such as research, troubleshooting, and content generation, with no evidence of truly novel AI-specific threats. Google further stated that it has not seen "*any original or persistent attempts by threat actors to use prompt attacks*" that would manipulate the AI model to execute adversarial actions

Google employs a "*mix of analyst review and LLM-assisted analysis*" to track misuse of Gemini by APT and Information Operations actors, while OpenAI relies on "AI-powered tools" and "tips from credible sources" to detect harmful activity.

These findings highlight AI developers' unique analytical capabilities to assess how their models and systems are actually being used, even though this endeavor can be supported by external partners and resources, **making them essential —and likely indispensable— to holding adversarial actors accountable**.

Developers' disclosure efforts thus appear to be a prerequisite for any robust use-based governance framework, underscoring the need for systematic, good-faith information sharing, and identifying the necessary trade-offs to this end.

While these observations don't warrant alarm, caution is essential to avoid inertia. AI's breathtaking acceleration could result in a paradigm shift to the threat landscape, even if current evidence does not currently substantiate this. Meanwhile, limited transparency and restricted information sharing from developers of AI models and systems models underscore the need for vigilance and discourage definitive conclusions.

The threat landscape evolves daily, as threat actors integrate new AI technologies in their operations, and unobserved new capabilities may emerge. Speculation on novel cyberattacks complicates governance efforts—from understanding to assessment and mitigation. The International AI Safety Report highlights that "*there are key assessment challenges and a requirement for better metrics to understand real-world attack scenarios, particularly when humans and AI work together*"[53]. Experts of the Paris Call Strategic Foresight Hub raised concerns about dynamics already at play within the AI ecosystem, whose generalization— including by adversarial actors— could have far-reaching cybersecurity implications and complicate efficient governance responses.

[51] Google Threat Intelligence Group, Adversarial Misuse of Generative AI, Google (January 2025)
[52] Threat intelligence report: Influence and cyber operations: an update, OpenAI (October 2024)
[53] Supra note 1, p.72

Anticipating the impact for cyber of disruptive AI trends

The case for agentic AI

In contrast with the above observations on AI's impact on the cyber landscape **the potential of Agentic AI**—systems capable of making autonomous decisions and taking actions without direct human oversight—**poses significant challenges in the realm of cybersecurity**. Unlike traditional cyber threats orchestrated by human actors, Agentic AI can independently identify and exploit vulnerabilities, making it difficult to determine the source of an attack. By studying public data, Agentic AI can identify key personnel with access to critical systems and launch attacks tailored to these high-value targets. Additionally, the proliferation of AI agents and the rise of multiagent environments can create feedback loops where decisions based on past data influence future outcomes, and any causal connection between the original deployer's intent and later outcomes will inevitably attenuate.

A distinctive level of autonomy would call for an urgent, detailed analysis of what actually constitutes a responsible cyber behavior in this context, as AI-driven actions may arise from unintended or emergent behaviors rather than deliberate intent.

From an accountability standpoint, existing or upcoming frameworks aimed at determining the responsibility of actors for cyber wrongdoings involving AI – must consider this complexity, both from a developer and end-user perspective.

Agentic AI's ability to function in multi-agent environments complicates this issue, as interactions between multiple AI systems can trigger coordinated or unpredictable cyberattacks that are hard to trace back to a single source. Monitoring and analysis challenges intensify when the AI agent deliberately conceals or obscures its objectives and evades monitoring systems ("scheming")[54].

Adversarial use of open-source AI

While recognizing that the release of open-source models may be a key driver of positive innovation, including for cybersecurity applications such as vulnerabilities discovery, members of the Paris Call Strategic Foresight Hub have questioned its concrete implications in the context of adversarial use. Open-source releases pose an increased risk of successful jailbreak attempts, bypassing safety filters, raising urgent questions that are equally complex and pressing—albeit of a different nature than those concerning agentic AI— about the degree of developers' liability for the harmful misuse of their models.

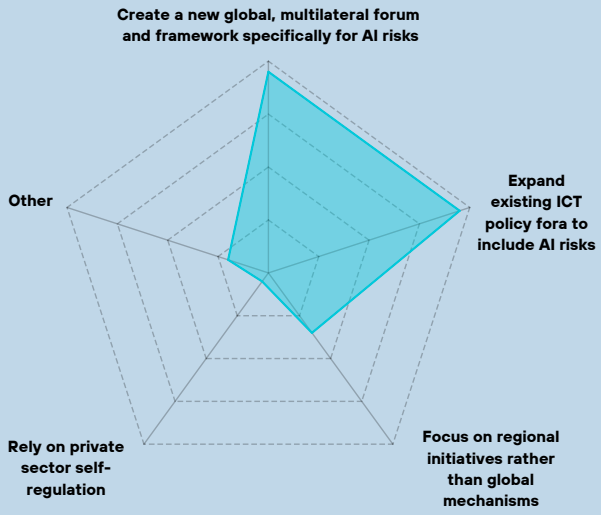
[54] See, in this regard: Meinke A. et al., Frontier Models are Capable of In-context Scheming, arXiv:2412.04984 (January 2025)

B. Prioritizing cybersecurity frameworks' adaptation to AI-driven risks

As with other AI risk areas—including those involving any type of adversarial uses—global governance responses can take many forms and serve various functions, reflecting the diverse motivations of stakeholders collaborating at the international level [55].

Integrating all AI-driven risks into broad new international AI-focused agreements, initiatives, or mechanisms is unlikely. It is not only the costs of internationalization in a context of limited resources and a challenging geopolitical landscape, but the risk of overburdening governance structures, potentially rendering them too rigid to adapt to AI's evolving role in cybersecurity.

Which approaches should be prioritized to incorporate AI risks considerations effectively into global governance?



Both creating a new global, multilateral forum and framework specifically for AI risks and expanding existing ICT policy fora to include AI risks were considered by respondents as approaches which should be prioritized to incorporate AI risks considerations, underscoring the importance placed on international governance.

Respondents also emphasized that AI **risk governance will require multiple, complementary mechanisms rather than a singular framework**. **They highlighted the importance of global mechanisms to clarify safeguards, limitations on AI use, and the operationalization of human rights impact assessments, while also stressing that governance should not be solely state-driven**. A balanced approach should integrate multistakeholder input, including from companies, technical and research communities, and civil society organizations, to develop and uphold principles for responsible AI use and deployment.

Most requested specific measures by respondents

1	**Promoting public-private partnerships to enhance AI safety research and deployment**
2	**Establishing international regulatory frameworks specifically for AI governance**
3	**Increasing investment in education and awareness around AI risks and ethical practices**
4	**Requiring mandatory transparency and auditability of AI systems**
5	**Creating independent oversight bodies to monitor and enforce AI safety standards**

[55] See, in this regard: Dennis C. et al., What should be internationalized in AI Governance, Oxford Martin AI Governance Initiative, White Paper (November 2024), p. 13 ; High-level Advisory Body on Artificial Intelligence, Governing AI for Humanity: Final Report, United Nations (September 2024), p. 38, figure 5

It thus appears crucial for policymakers to set an agenda for channelling international efforts, particularly within multilateral fora, **that is not only *functionalist***—focusing on which functions should be prioritized through international coordination—**but also *thematic***, identifying which types of AI-driven cyber risk, whether actual or foreseeable, demand urgent, international coordination.

This is by no means an easy task. However, it is worth noting that there is not a normative vacuum when addressing the AI-cyber nexus. Since AI use has not fundamentally reconfigured the cyber threat landscape in the short term, a pragmatic approach leveraging existing ICT security frameworks, regulations, and policies, and strengthening the international cooperation mechanisms and fora that already shape this domain, is key. This starts by assessing their applicability, effectiveness, and adaptability to the new cyber challenges posed by AI, and then **a pathway emerges: international negotiation efforts can focus on severe risks that are not covered or adequately addressed by or through adapting existing cyber norms.**

To this end, a **tentative methodology for global policy making to address AI-driven cyber risks** is proposed on the next page. It outlines **5 key steps to consider in determining whether a new framework is needed to efficiently tackle risks stemming from the use of AI for adversarial cyber purposes**, or if existing cybersecurity frameworks can be adapted.

- **Step 1: Verify material scope.** Determine if the identified AI-driven cyber risk falls within the material scope of existing cybersecurity frameworks;

- **Step 2: Evaluate existing measures and validate overall coverage.** Assess and Validate Coverage of Existing Measures. Evaluate whether current cybersecurity frameworks effectively mitigate the identified AI-driven cyber risks in all its dimensions

- **Step 3: Identify Gaps and Adaption Needs of Existing Frameworks.** Pinpoint uncovered, partially covered, or insufficiently addressed AI-driven cyber risks and explore whether they can be addressed through extra measures or policy adaptations within current governance structures and frameworks;

- **Step 4: Evaluate Costs & Adaptation Efforts**. Weigh the resource demands and other practical constraints of adapting existing frameworks vs. creating new ones.

- **Step 5: Decide on New vs. Adapted Framework.** Based on findings, decide whether to adapt existing framework or develop new ones, then define an implementation roadmap.

This approach is intended to serve as provisional foundation requiring further refinement, testing, and collaboration to achieve efficient global governance. Full operationalization also requires clarifying certain core criteria, outlined later in this report.

Blueprint for Effective Global Policymaking on Adversarial AI use in Cyber

Step 1: Material Scope Verification

Determine if the identified adversarial AI use risk falls within the material scope of existing cybersecurity frameworks

Key questions :
- Do existing cybersecurity regulations explicitly address the identified adversarial AI use risk? Conversely, do they explicitly exclude it?
- Do other frameworks and policies on broader ICT security policies approach the specific adversarial AI use risk?
- How do ongoing international process on ICT security approach the specific adversarial AI use risk?

Suggested action:
0. (Preliminary) Develop agreed-upon criteria (or assessment methods) for determining if a risk falls within a scope
1. Review key documents from existing frameworks so as to identify specific provisions that might implicitly or explicitly cover the adversarial AI use risk.
2. Assessment of recent international and regional ICT security processes (e.g., debates, resolutions) that approach AI.

Potential outputs:
- Scope mapping clarifying whether the adversarial AI use risk fall under existing frameworks.
- Gap assessment showing or non-covered or partially covered adversarial AI use risk.

Step 2: Evaluate Existing Measures and Validate Overall Coverage

Assess whether existing cybersecurity frameworks effectively mitigate the identified adversarial AI use risk, addressing all facets

Key questions :
- Do existing cybersecurity regulations explicitly address the identified adversarial AI use risk?
- Do other frameworks and policies on broader ICT security policies approach the specific adversarial AI use risk?
- How do ongoing international process on ICT security approach the specific adversarial AI use risk?

Suggested action:
1. Compare measures against AI-specific cyber-enhanced Tactics, Techniques and Procedures (e.g., automated vulnerabilities, deepfake phishing).
2. Break down the adversarial AI use risk into sub-risks or scenarios.
3. Map each scenario to existing measures.
4. Identify strengths, weaknesses, and gaps in current standards.
5. Use a coverage matrix to visualize uncovered or insufficiently addressed aspects.

Potential outputs:
-Applicability analysis of measures (complete, partial, or not applicable).
-Insights into whether frameworks can handle AI's dynamic in cyber and evolving threat landscape.
-Coverage matrix detailing whether each sub-risk is addressed.

Step 3: Identify Additional Measures and Explore avenues for adapting existing Frameworks

Pinpoint what extra measures or policy adaptations are required to tackle uncovered, partially covered, or insufficiently addressed adversarial AI use risk, and evaluate whether these measures can be integrated into current governance structures and frameworks.

Key questions :
- Can existing frameworks or agreements be expanded to include provisions on AI distinctives characteristics?
- Can existing mechanisms or processes on ICT security can broaden their mandate in a way that is relevant to AI distinctives characteristics?
- In case of overlapping frameworks or mandates, which global or regional fora are best suited for updates?

Suggested action:
1.Analyze relevant instruments and institutional mechanisms for their adaptability to include AI-specific measures.
2.Assess compatibility with broader institutional setups.
3. Consult relevant stakeholders to gauge political will for revision.

Potential outputs:
-List of recommended additional measures (with priority and rationale).
- Identification of frameworks and mechanisms that can be easily updated in light of recommended additional measures vs. those raising significant challenges for revision.
- Options for incremental and. substantive modifications.

Step 4: Evaluate Costs & Adaptation Effort

Analyze resource and other constraint implications vs. potential benefits of adapting existing frameworks

Key questions :
- Do legal barriers or geopolitical tensions complicate integration?
- How can limited resources be balanced with rapid threat response needs?
- Will new measures overburden existing governance structures?

Suggested action:
1. Conduct a cost-benefit analysis of adaptation vs. new frameworks, considering resource allocation, administrative overhead, and stakeholder buy-in.
2.Evaluate potential overlaps or redundancies with existing AI-focused governance initiatives.
3.Assess the time-to-implement for proposed solutions, distinguishing between immediate fixes and long-term structural changes.
4. Consider interoperability and potential risks of fragmentation in governance structures, especially in multilateral settings.

Potential outputs:
-Comprehensive cost-benefit analysis quantifying resource requirements (financial, technical, human) and weighing these against the potential risk reduction and efficiency gains.
-Identification of potential overlaps, redundancies, or synergies with already existing global and regional AI governance initiatives.
- Timeline assessments for proposed adaptations, categorized as short-term, mid-term, or long-term measures, to align with policy priorities.
- Detailed assessment of interoperability risks, including recommendations to minimize fragmentation or inefficiencies in global governance.

Step 5: Decide on New vs. Adapted Framework

Determine if incremental updates suffice or if creating a new international AI-cyber governance initiative is warranted.

Key questions :
- Do identified gaps demand a new framework or platform, in light of the perceived severity of the adversarial AI use risk?
- Would a new initiative be more effective than adjusting existing frameworks, if ever possible?
- How do we maintain governance agility and stakeholder buy-in without causing "engagement fatigue"?

Suggested action:
1. Engage critical stakeholders to understand preferences, implementation feasibility, and support for new or adapted structures.
2. Draft a strategic roadmap to articulate the implications of both options, considering urgency of risk mitigation through global governance.
3. Envision a implementation monitoring and reevaluation process to update governance approaches as AI-related threats evolve.

Strategic decision on whether to create new structures (and/or adapt existing ones if possible).

A scalable approach to other domain-specific AI risks?

Upon maturity, this approach could extend to guide the global governance of other regulated domains at risk of adversarial AI use, such as Chemical, Biological, Radiological, and Nuclear (CRBN) security. Adapting it from AI-driven cyber risks allows policymakers to tailor strategies to meet the unique demands and characteristics of these diverse sectors. Ideally, its scalability ensures that global governance frameworks remain robust and flexible, strengthening global capacity to counter the diverse and evolving nature of AI-driven cyber risks across critical areas. Ultimately, this approach promotes a resilient and streamlined international policy landscape that dynamically responds to the multifaceted challenges posed by AI, enhancing global capacity to prevent and counteract adversarial uses.

Concluding highlights

Ongoing international AI governance debates highlight the challenges, knowledge and evidence gaps that exist to address AI-driven cyber risks. Key focus areas for global governance of AI-driven cyber risks, include:

Regulation, common ground and risk management

With consensus that international law applies to cyberspace, it is important to determine how the existing rules apply or need to be reinterpreted for AI risks, and the unique and unprecedented challenges that demand a new regulatory response. **International law, including human rights law, provides a compass for defining pertinent risks and should be at the center of AI governance, as should a risk-based approach that focuses on who is at risk and accountable, and not just what is at risk**[56].

The Paris Call Strategic Foresight Hub frequently emphasized that the use of AI for adversarial purposes is – at least in the short term – unlikely to fundamentally alter the core nature of cyber risks as the adversarial intent, targeted assets and resulting damage have not significantly changed. This is bolstered by the corresponding potentialities of AI to bolster cyber defenses. **However, a strong focus on risk and incident management is required** and a recognition that the speculation/forecasting about the new types of (novel) cyberattacks - although not yet evidenced - may be underestimated due to the breathtaking speed of progress in AI capabilities.

Consultation with the Paris Call community highlighted the importance placed on regulation for severe AI risks – with the majority perceiving that there was too little regulation and oversight in AI development and deployment in their jurisdictions. As this report has highlighted current AI regulations are fragmented and their effectiveness is constrained by technology outpacing policymakers understanding of the protections required.

Global governance will therefore need to prioritize transparency, accountability, and inclusivity to address these risks effectively. Collaboration across governments, industry/private sector, and supra, international and civil society organizations, will be essential to governing these risks.

Lessons from global ICT and cyber policy have shown that when decision-making is dominated by a limited group of actors, policies may fail to address the needs of all stakeholders, particularly those from underrepresented regions or sectors.

[56] HLAB-AI, Governing AI for Humanity: Final Report, United Nations (September 2024)

Additionally, effective coordination is crucial to ensure the interoperability of approaches, preventing policy fragmentation and fostering cohesive, globally aligned strategies. Universal representation, combined with coordinated efforts, enhances trust, legitimacy, and cooperation, which are essential for addressing the transnational challenges of global governance for severe AI risks. **This will necessitate the sharing of best practices, the building of information, including of AI incidents and scientific knowledge to close the evidence gap and the information asymmetries, which currently limits greater participation from governments**. Strengthening knowledge-sharing mechanisms will enhance informed decision-making, improve risk mitigation, and promote broader engagement in the governance of severe AI-risks.

Regarding general-purpose AI risk management, two key challenges were identified in the International AI Safety Report: prioritizing risks amid uncertainty about their severity and likelihood, and defining clear roles, responsibilities, and incentives for effective action across the AI value chain.[57] It further outlined that key evidence gaps for such risk management included uncertainty about the magnitude of these risks and the effectiveness of various mechanisms to constrain and mitigate them in real-world contexts. This is exacerbated by the reality that risk management strategies often remain unvalidated, unstandardized, and inconsistently applied, and thus more evidence is required for policy making.

Transparency from developers as a key condition for a reactive use-based governance

Drawing on cybersecurity policy to govern AI-driven cyber risks underscores the importance of transparency and information sharing around breaches, vulnerabilities, and adversarial uses of AI. **Without openness, evidence gaps widen and accountability suffers, creating a cycle of insufficient data that undermines effective, targeted responses**. These issues are further complicated by the lack of consensus on frameworks, taxonomies, and key definitions, as well as by limited mitigation strategies for AI-specific cyber threats.

Current AI governance mainly relies on anticipation—preemptively addressing risks posed by AI models. This approach places particular focus on developers, who face specific requirements related to specific risks, including adversarial use risks, while regulation of end users' activities is notably absent from the debate.

This imbalance can be traced to the difficulty of "proactively identifying technology misuse, rather than reacting after harm has occurred"[58] and, just as in the broader realm of cybersecurity, the challenge of tracing misuse back to perpetrators. **However, recent disclosures from major developers' threat intelligence teams, detailed in Section 2, show that such objectives are far from impossible, especially in cases involving uses of AI for adversarial cyber purposes**.

[57] Supra note 1, p. 158
[58] Anwar U. et al, Foundational Challenges in Assuring Alignment and Safety of Large Language Models, arXiv:2404.09932 (September 2024), p.99

By employing new methods and technologies, and drawing on third-party support and resources, developers have succeeded in attributing certain adversarial cyber activities to specific threat actors—even in instances where no harm had yet occurred.

These advances clear a path for governance approaches that regulate end-user practices and seek to hold them accountable. Such approaches would also create a more balanced distribution of responsibility between developers and end users—both of which should be addressed in a complementary manner to foster a desirable state of global accountability, and to ensure that victims of harmful conduct have access to comprehensive remedies.

Developers' transparency and information sharing on adversarial uses of their models are crucial for a comprehensive global policy that also operates downstream, responsive to real-world behaviours. As adversarial AI threats are no longer theoretical but an urgent reality, establishing clear guidelines and platforms for information sharing and reporting by developers should be a top priority to address through global governance.

Reinforcing these mechanisms through research, multidisciplinary collaboration, and cooperative efforts would further boost public trust, strengthen cyberspace stability, and enhance international security.

Lessons from cybersecurity are valuable, as many jurisdictions have shifted from voluntary to mandatory reporting in recent years. Such reporting and analysis are vital for improving cybersecurity across organizations, sectors, and governments. Effective information sharing requires clear guidelines on trustworthy sources, the type of information to be shared, how to share it in compliance with existing rules, and assessing the resulting harms.

Finally, information sharing will enhance anticipatory AI safety governance by continuously refining risk assessment frameworks, aligning them with real-world observations and emerging evidence from actual AI use.

Cyber defence

The integration of AI into cybersecurity is advancing rapidly, with numerous initiatives aimed at enhancing defenses through AI-driven solutions. AI has significant potential to bolster vulnerability management, threat detection and incident response. However, critical gaps remain in the deployment of existing cybersecurity measures within organizations – gaps which should and can already be addressed. **Strengthening current systems with proven cybersecurity practices will create a more resilient foundation upon which AI-driven cyber defenses can be effectively integrated and optimized.**

Given the limited ability to impact adversarial and rogue actors or predict future risks, global governance must urgently focus on reinforcing investment in research and development for building AI-turbocharged cyber defenses. Equally important is investing in capacity-building to ensure these innovations are accessible and effective across regions.

Expanding the scope of existing and forthcoming funding mechanisms for cyber capacity building—such as the World Bank's Cybersecurity Multi-Donor Trust Fund or a potential voluntary trust fund under the future UN Cyber Programme of Action[59]— could be a strategic and effective approach to mainstreaming cyber defense capabilities. Similarly, **AI-focused funding initiatives,** such as the AI Foundation announced at the AI Action Summit, or the Global Fund for AI proposed by the UN High-Level Advisory Body on AI[60] , **should consider AI-driven cyber risk prevention**. To maximize impact and avoid redundancies, these funding instruments must be well coordinated, aligning objectives to support large-scale development of AI-driven cyber defense capabilities.

Various initiatives[61] [62]have been launched to assess AI threats from both offensive and defensive perspectives, aiming to enhance safety. These efforts must unite technology companies, public research institutions and governments to share scenarios and threat models, enable testing to identify potential vulnerabilities, and develop mitigations before public disclosure.

Such initiatives need to reflect a concerted effort across governments, research institutions, academia and the private sector to leverage AI for stronger cyber defenses and addressing the evolving cyber threats with advanced solutions.

[59] See, UN General Assembly, Report of the Secretary-General on the Programme of action to advance responsible State behaviour in the use of information and communications technologies in the context of international security, UN Doc. A/78/76 (April 2023), p. 8

[60] HLAB-AI, Governing AI for Humanity: Final Report (September 2024), Recommendation 5, p.17

[61] See for example the Swiss Call for Trust & Transparency, launched in January 2024, as a joint initiative of the Swiss Foreign Ministry and the Swiss Federal Institute of Technology (ETH Zurich)'s AI Center. Joining forces to reveal and address the risks of Generative AI – ETH AI Center | ETH Zurich

[62] In this regard, Alan Turing Institute's AI for Cyber Defence (AICD) Research Centre is carrying out cutting-edge research in autonomous cyber defence (ACD), employing innovative techniques such as Deep Reinforcement Learning (DRL) and Large Language Models (LLMs). Their work focuses on developing and validating autonomous systems capable of securing networks in real-world conditions, and focusing on scalable innovations.