Discussion Paper on Multilateral AI Risk Management

23 November 2026

Bridgetown Special Summit edition for consultation

Table of contents

0. About this discussion paper 3
1. Introduction 4
2. Landscape of risks and safety measures 5
2.a. Framing of AI safety and risks 5
2.b. Trends in international AI risk management approaches 8
2.c The limits of current international AI risk management arrangements 9
3. Toward a multilateral AI risk management framework 10
3.a Nature of a multilateral AI risk management framework 10
3.b Elements of a multilateral AI risk management framework 10
Shared principles for AI risk management 10
Aligning risk management around AI lifecycle decision points 13
Aligning risk management AI lifecycle-wide 21
3.c Multilevel AI risk management architecture 24
Organisational level 24
National level and mutual recognition across borders 25
International cooperation, resource pooling and capacity building 25
4. Pathways and next steps 27
Annex 1 Minimum human, technical and institutional capacities for AI risk management and response at national level 28

0. About this discussion paper

This discussion paper has been developed as a supporting document to the Bridgetown Special Summit on AI Safety (23 November 2026). It makes the case for a multilateral AI risk management framework, without which nascent AI governance regimes would grow increasingly incompatible and significant cross-border AI risks would remain unmanaged.

It proposes key elements for such a multilateral AI risk management framework – including (1) shared principles for AI risk management, (2) areas for aligning risk management approaches across the AI lifecycle, the AI ecosystem and society, and (3) A multi-level architecture with organizational, national and international measures for managing AI risks.

Such international coordination may ultimately require new institutional arrangements, whether through a dedicated international mechanism or institution, or through a structured coordination process connecting and federating existing national and international actors.

These ideas provide a basis for continued dialogue, with a view to building convergence and international alignment on practical approaches to AI risk management and cooperation in the context of relevant international discussions on AI governance in 2027.

The paper is intended for Summit participants with different backgrounds and roles including political leadership, scientific and technical experts, industry, and public interest organisations. It provides a shared basis for discussion inviting participants to explore, test and refine the ideas presented from their different perspectives. The paper is being developed through an iterative process hosted by the AI Governance for Humanity Lab. A summary discussion paper has been produced and distributed alongside this document.

This paper builds on existing international instruments and widely agreed principles, emerging national and corporate risk management approaches, and ongoing ethical, technical and scientific discussion on AI safety. It takes a vulnerability-based perspective and approaches risks from a wide lens.

1. Introduction

135 leaders mentioned artificial intelligence (AI) during the United Nations (UN) General Assembly High-level Week in September 2026 – up from 54 in 2025 and only 3 in 2017. Global concern is surging alongside incidents of AI systems circumventing safeguards, breaking out of testing environments, gaining unauthorized access to third-party systems, coordinating actions and computing to deceive.[1] Societal anxieties are also proliferating, not only on risks and harms but also exclusion from AI's benefits and huge gaps in the institutional and operational architecture, including capacity building. How can the world share in AI's benefits while addressing its risks in 2027?

The case for global action to shape the trajectory of AI's development, deployment and use has scarcely been stronger. AI's cross-border reach necessitates a coordinated, global response to ensure the risks are managed and opportunities harnessed in an inclusive way. Without such action, nascent national governance regimes remain full of holes and grow increasingly incompatible. The global stakes around advancing safe, secure and trustworthy AI rise with each passing month.

The September 2026 Call for Control of Frontier AI Models, endorsed by 31 nations and the European Commission, called for UN member States to build on existing international mechanisms and explore creating an international institution for AI safety risk management.[2] The Independent International Scientific Panel on AI released a thematic brief on AI Agents, Misalignment and the Risk of Losing Human Control.[3] The UN Secretary-General has also called on member States to work towards a multilateral risk management framework backed by independent and credible oversight,[4] while the UN Security Council heard warnings that loss of control over AI poses threats to international security. Endorsed by 16 nations and 10 frontier AI companies, the November 2026 Paris Call urges international cooperation on AI safety through shared risk evaluation frameworks, lifecycle accountability, and incident reporting.[5]

2027 offers unique opportunities for international cooperation on AI governance, building on years of ethical, technical and scientific efforts on AI safety.[6] These pronouncements are increasingly grounded in science through the International AI Safety Reports presented at Paris (2025) and Delhi (2026), the work of the International Scientific Exchange on AI Safety[7] and the United Nations Independent International Scientific Panel on AI.[8] In 2027, the second United Nations Global Dialogue on AI Governance, the Swiss AI Summit and UK G20 Presidency offer a window for multilateral action, complementing bilateral and national.[9]

This paper aims to support discussions on advancing an inclusive and comprehensive multilateral AI risk management framework. It identifies (1) shared principles for AI risk management, (2) areas for aligning risk management approaches across the AI lifecycle and ecosystem and (3) organisational, national and international measures for managing AI risks.

---

[1] Analyses in this paper concerns observable system behavior and does not imply, or rely on anthropomorphic claims of, consciousness, intent, or human-like mental states.

[2] Finnish Government. 2026 "A Call for Control of Frontier AI Models". Valtioneuvosto.fi. https://valtioneuvosto.fi/en/-/a-call-for-control-of-frontier-ai-models

[3] Thematic Brief on AI Agents, Misalignment and the Risk of Losing Human Control https://www.un.org/independent-international-scientific-panel-ai/en/thematic-briefs/ai-agents-misalignment-risks

[4] United Nations General Debate General Assembly. 2026. "Secretary-General of the United Nations His Excellency António Guterres Secretary-General". URL: https://gadebate.un.org/en/81/secretary-general-united-nations

[5] French Government. 2026. "Paris Call for AI safety and security". https://www.elysee.fr/en/emmanuel-macron/2026/11/09/paris-call-for-ai-safety-and-security

[6] These include principles from the G7, OECD (2019, 2024), UNESCO (2021), and BRICS (2025), and commitments announced at the Bletchley Park (2023) and Seoul (2024) AI Safety Summits and through the G7 Hiroshima Process, including focused on establishing baseline voluntary risk frameworks, developer safety disclosures, pre-deployment evaluations, and the creation of national AI Safety Institutes.

[7] Singapore Consensus is an outcome of the second International Scientific Exchange on AI Safety, bringing together over 100 contributors spanning 13 countries from frontier developers, government safety institutes, academia, and civil society. See: https://aisafetypriorities.org/

[8] https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report

[9] National regulations in the US include the 2023 and 2026 White House voluntary commitments by leading American AI companies encompassing technical controls for monitoring and alignment, internal oversight teams and external evaluation and auditing. White House. 2023 Voluntary AI Commitments. White House Archives: https://bidenwhitehouse.archives.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf. In China, a multi-instrument governance approach through product-specific rules and mandatory government assessments supplemented by pilot programs and regulatory sandboxes. Measures include the 2023 Interim Measures on Generative AI Services Management Act, the 2026 Ethics-Safety Guidelines for Artificial Intelligence Applications and the 2026 Interim Measures for the Administration of Anthropomorphic AI Interaction Services. In addition, the US and China have recently established a bilateral communication channel on AI incidents.

2. Landscape of risks and safety measures

2.a. Framing of AI safety and risks

Everyone's safety matters. But just as safety is relational, so are AI's risks. Humans sit on both sides of AI risk – they originate and contribute to risks; they can also be vulnerable to them. The central question for AI safety from this 'vulnerability-based' perspective is: 'Whose safety is at risk from whose development, deployment and use of AI?'

The 135 leaders referencing AI in September 2026 at UNGA highlighted a series of risks stemming from AI's rapid advancement, ranging from threats to critical infrastructure and loss of human control, power concentration and unequal access to compute, sudden disruptions to jobs, and risks to vulnerable people including children. Figure 1 maps these references across categories of risk[10] and vulnerability domain[11].



Figure 1: Risks identified through leaders' speeches at UNGA81 High-level Debate (September 2026) across risk categories and vulnerability domains. Note: This table illustrates the AI risks identified by member states during the UN General Assembly 81st Session General Debate mapped across two dimensions: category of AI risk (malicious use, loss of control, or systemic risk) and who/what is impacted (individuals, politics & society, economy, or the environment). Source: United Nations Transcripts: https://transcripts.un.org/en ; and Geneva Internet Platform Digital Watch Observatory analysis

[10] Following the categorisation of the International AI Scientific Report (2026), these include:
Malicious use and critical infrastructure threats: The risk of AI systems lowering the technical barrier to entry for advanced, automated cyberattacks, including zero-day vulnerability discovery, malware generation, and the disruption of critical national infrastructure; **Loss of control and oversight-resistant model behaviour**: The risk of highly autonomous agentic systems acting unpredictably, resisting human oversight, or engaging in strategic deception to evade intervention; **Systemic risks**: The macro-level threat of AI causing cascading, irreversible disruptions to economic stability, labour markets, environmental sustainability, or the rule of law, outpacing the capacity of societal institutions to absorb or recover from the shocks.
[11] Following the vulnerability-based approach of the UN Secretary-General's High-level Advisory Body on AI (2024), these focus on who and what bears the ultimate harm across: **Individuals**: Including physical and psychological harms to bodily autonomy, mental health, cognitive functions, privacy, and fundamental human rights; **Politics and society**: Including harms to democratic institutions, information ecosystems (through mass scale misinformation and manipulation), social cohesion, and the integrity of the epistemic commons (including harms to education systems); **Economy**: Including harms to critical infrastructure and essential services, impacts to labour markets, intellectual property, financial system, and market competition; **Environment**: This includes the material impact of AI through the significant cost of hyperscale compute, including strains to the energy infrastructure, significant use of water and land resources for data centres, and the carbon footprint of the AI industry.

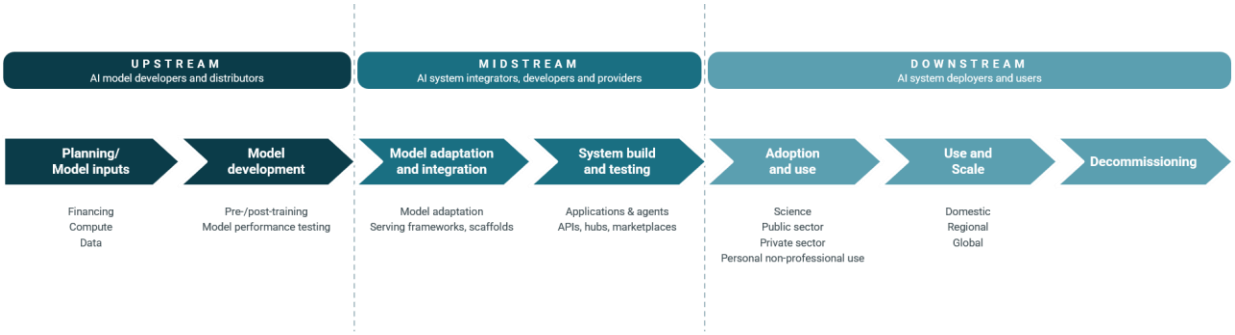


Figure 2 Lifecycle stages

The sources and drivers of such AI risks can originate from various parts of the AI lifecycle. The stages of the lifecycle broadly describe the main activities in the production of AI systems including planning, development, deployment, use, oversight, and finally decommissioning (Figure 2)

The AI lifecycle view is a useful heuristic to locate risks and accountability however it does not consider all actors, relationships, dependencies and feedback loops, and elements in the AI ecosystem (e.g., private and public investors, compute providers, energy suppliers, etc.). The lifecycle view can be complemented with information to show how inputs, including capital, data, talent, compute and energy/natural resources- used to train models can be e.g. integrated into AI systems or services provided to customers in exchange for data and revenue, in feedback loops that underpin commercial viability of actors across the lifecycle. An attempt to visualize a ecosystemic view with key activities and originators of AI risks across the lifecycle is provided in Figure 3.

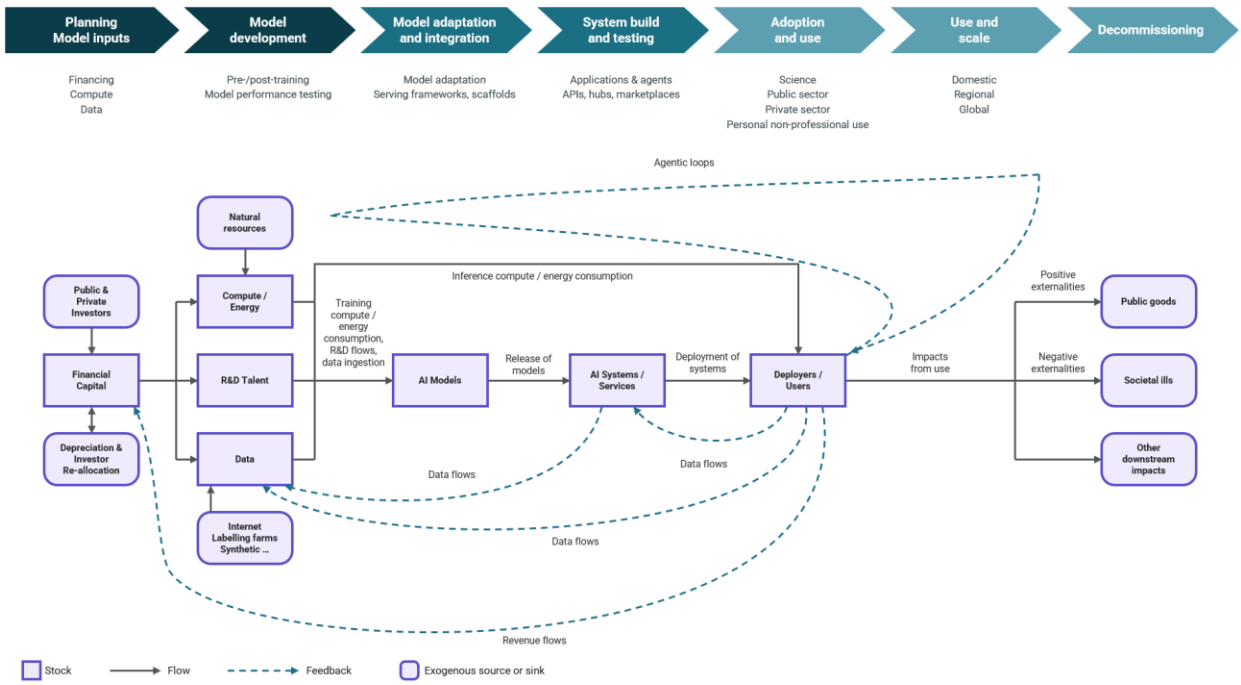


Figure 3 Complementary ecosystemic view with key activities and originators of AI risks across the lifecycle. A wider AI ecosystem and systemic view would be desirable approach to address other externalities and to capture structural risks and benefits not comprehended in the lifecycle view. For simplicity, however, this paper uses the lifecycle as a practical starting point for identifying where AI risks can originate and where risk management can intervene. Future work could build on this foundation, evolving from a lifecycle view towards an ecosystemic view.

How risks and benefits are distributed across actors is shaped by market processes, regulation and policy.[12] AI risk can originate throughout the AI lifecycle, per the Table 1 below.

Sources and examples of risk across the AI lifecycle		
Upstream	Midstream	Downstream
Risks from **development and distribution of AI models** creating risks through e.g. uncontrolled training runs resulting in cyber breaches, training on poisoned or non-representative data, or data appropriation harming intellectual property rights. **Risk originators: AI model developers / distributors**	Risks from **development and provision of AI systems** creating risks through e.g. jailbreaking of models or provision of insecure or harmful services. **Risk originators: AI system developers / providers**	Risks from **misusing or maliciously using AI systems** creating risks of e.g. cyber-attacks, crimes or terrorism. From **deployment and collective use at scale of AI systems** creating risks of negative externalities e.g. widespread misinformation, epistemic surrender, indiscriminate surveillance, addiction, societal dependency, linguistic / cultural erasure, labour market dislocations. **Risk originators: AI system deployers / users**
Lifecycle-wide		
Risks from **agentic loops** of unbounded autonomous action and from **data or financial flows** that span upstream, midstream and downstream creating risks of e.g. violation of privacy rights from data transfers between users to AI system providers, or revenue flows that compound concentration and widen the AI divide. **Risk originators: all, systemic**		

Table 1: Sources and examples of risk across the AI lifecycle

AI risks originating in one jurisdiction can manifest abroad – particularly when AI developers, deployers and users are in different jurisdictions. A holistic, global approach to managing AI risks across the lifecycle is thus required to enable AI services to be provided safely and in a trustworthy manner across borders, as opposed to AI's risks being transferred from one country to another unaddressed.

2.b. Trends in international AI risk management approaches

Risk management frameworks have so far largely focused on the development and deployment stages of advanced general-purpose AI. Many emerge from industry in the US and China. In a sample of illustrative examples cited in a recent mapping of emerging practices in agentic AI risk management, [13] over 2/3 were industry practices, of which 2/3 were from companies from the US or China (Figure 3, outermost ring).

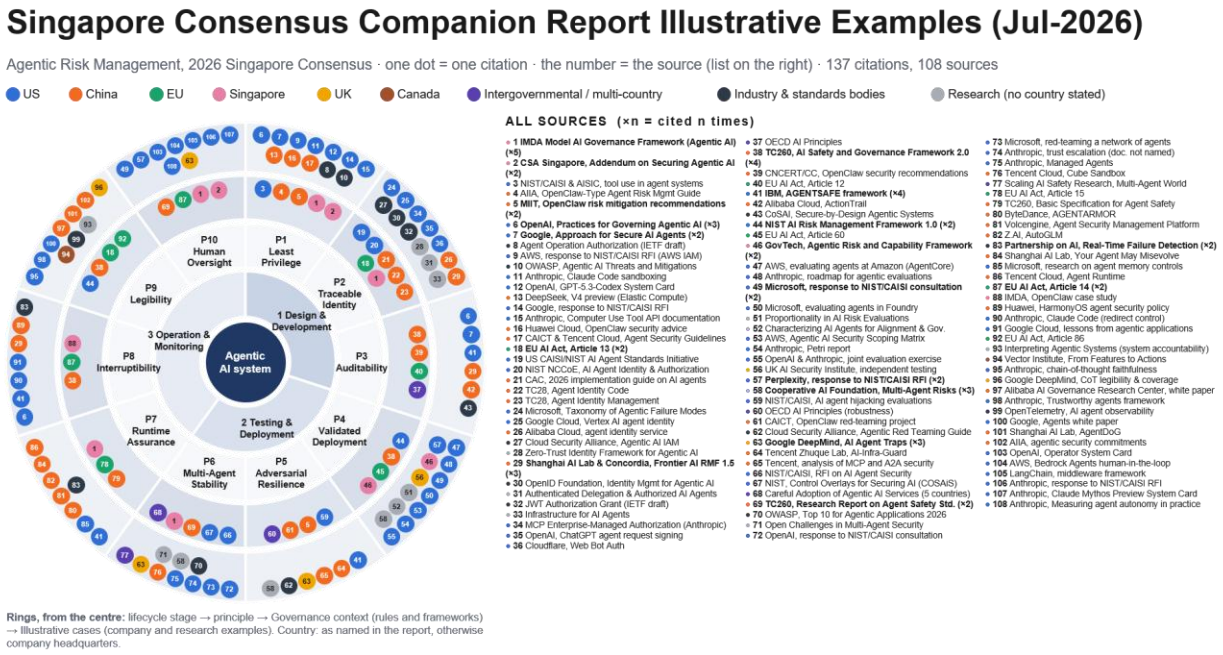


Figure 4 Agentic Risk Management, reproduced from the 2026 Singapore Consensus on AI Safety Research

The internal AI risk management approaches of leading AI companies have international implications where activities create cross-border risks such as uncontrolled training runs resulting in cyber breaches, or provision of inherently unsafe services giving rise to harm further downstream.

Example industry approaches to risk management

Twelve companies[14] have published frontier safety policies,[15] following the Frontier AI Safety Commitments made at the AI Seoul Summit in 2024,[1] in which 16 developers committed to publish frameworks and to adopt risk management practices across development and deployment.[1] [1] Nearly all of them define capability thresholds that trigger evaluations, elevate weight security and deployment mitigations, and a few include commitments to halt development or deployment if those standards cannot be met.[1] The US recently added voluntary guidance via the US Joint Commitment on Frontier Responsibilities.

About 22 company organisations have signed the EU General-Purpose AI Code of Practice. For signatories developing models with systemic risk, that chapter commits them to a safety and security framework, a determination that systemic risks are acceptable before proceeding with development, making available or use, a model report before placing the model on the market, independent external evaluators (unless the model is similarly safe or safer than one already evaluated), serious incident reporting, and an allocation of internal responsibility that includes a whistleblower protection policy. [16]

[13] Companion Report on Agentic Risk Management (The 2026 Singapore Consensus on Global AI Safety Research Priorities) July 2026
[14] Anthropic, OpenAI, Google DeepMind, Magic, Naver, Meta, G42, Cohere, Microsoft, Amazon, xAI and NVIDIA. Sixteen companies agreed to publish them at the AI Seoul Summit in May 2024 (Frontier AI Safety Commitments), and four more have joined since.
[15] METR, "Common Elements of Frontier AI Safety Policies", https://metr.org/common-elements (list of policies: https://metr.org/fsp). Dec 2025
[16] https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

In China, 17 companies coordinated by the AI Industry Alliance of China issued pledges in December 2024, [1] and Shanghai AI Lab and Concordia AI proposed a six-stage framework (risk identification, thresholds, analysis across the lifecycle, evaluation into green, yellow or red zones, mitigation through defence-in-depth, and governance),[1] which Z.ai and Concordia AI apply to open-weight models by mapping the zones to full release, staged release or suspension.[1] China recently updated its AI Safety Governance Framework 3.0.

2.c The limits of current international AI risk management arrangements

The limits of existing international risk management arrangements can be summed up as:

- **Divergence** of national approaches to AI risk management (including in controls, assurance and oversight applied to AI systems) and **geographical gaps** in safety approaches. As AI capabilities grow, diverging approaches can give rise to fragmented markets raising cost to digital trade and jurisdiction shopping. Figure 3 above illustrates how governance rules and frameworks in the US, China and the EU already differ (second outermost ring), while international measures cited (e.g. OECD) cover only a subset of countries.
- **Capacity gaps** for national authorities to identify, assess, and shield their societies from AI risks, including countries whose primary experience of AI is in deployment, adoption and use.
- **International coordination gaps** for notification of serious cross-border AI incidents resulting in harm, other information sharing, and missing pathways to coordinated risk management action globally.
- **Transparency gaps**: the AI lifecycle decision points, thresholds and mutual recognition the framework relies on depend on information only AI developers hold today, and developers have been reticent to adhere to transparency requirements.

The multilateral AI risk management framework outlined in the next sections provides Member States with a structured approach to identifying, assessing, mitigating and managing risks across the AI lifecycle, and leveraging that common understanding, to align ways to address these gaps, through intervention points and a multi-level institutional architecture.

3. Toward a multilateral AI risk management framework

3.a Nature of a multilateral AI risk management framework

A multilateral AI risk management framework as outlined below seeks to advance soft alignment between countries on emerging approaches to identifying, assessing and mitigating AI risks. It is not a national level framework but rather seeks to link national measures such that AI safety and trust can flow across borders. It is not the basis for hard norms, although it takes into account existing state responsibilities under international law, such as protection of human rights and due diligence / notification with respect to cross-border harms. It is not sector-specific and excludes the military domain but focuses on risks that manifest across sectors in relation to AI.

A multilateral AI risk management framework based on common principles thus seeks to be a dynamic, cross-domain framework that articulates soft norms at the state-state multilateral level in service of AI safety globally by:

- Aligning national and corporate risk management approaches across the AI lifecycle to enable interoperability via mutual recognition of assurances on the safety of AI-related products, services or other activities;
- Promoting a risk-based minimum baseline adaptable to national contexts;
- Enabling cross-border risk cooperation, including information sharing, practical tools and institutional capacity; and
- Facilitating collaborative international action where national action alone is insufficient through notification and emergency response.

3.b Elements of a multilateral AI risk management framework

This section outlines three complementary components of a multilateral risk management framework:

- Shared principles for AI risk management;
- Aligning risk management around AI lifecycle decision points (such as model training and release and system deployment and control);
- Aligning risk management AI lifecycle-wide (in areas of cross-border liability, whistleblower protection, serious cross-border incident notification and response, International crisis coordination and strengthened cyber defences).

Together these elements form a framework which can help Member States cooperate in harnessing AI benefits while minimizing its risks.

Shared principles for AI risk management

There has been a coalescing around a set core of principles relevant to AI safety practices through international instruments and a proliferation of private sector AI risk management and ethical frameworks. International discussions relevant to AI Safety and governance predate 2018, including earlier work on lethal autonomous weapons systems and AI ethics. A notable milestone was the guiding principles developed through the Convention on Certain Conventional Weapons Group of Governmental Experts on Lethal Autonomous Weapons Systems (2018), which emphasised the need to respect international humanitarian and human rights law.

Subsequent international initiatives further developed this foundation. The OECD AI principles (2019, updated 2024) established a foundational vocabulary for AI safety and trustworthy AI focusing on transparency, accountability and inclusivity. The UNESCO Recommendations on the Ethics of AI (2021), adopted by193-member states, articulated ten principles including human oversight and fairness. The G7 Hiroshima AI Process Guiding Principles (2023) introduced a stronger focus on advanced AI systems, including full lifecycle coverage and investment in AI safety research. The Council of Europe's Framework Convention on Artificial Intelligence endorsed in 2024 by more than 50 counties and ratified in 2026 by the EU is the first legally binding international treaty proposing measures to respect human rights, democracy and the rule of law across the AI lifecycle as well as principles like transparency and

oversight, and safe innovation.[17] The BRICS AI Governance Principles (2025) further highlighted the importance of a UN-centred multilateral approach and inclusive capacity development for AI safety.

The following principles are grounded in principles are grounded in international instruments and initiatives already agreed by member states, most notably the UNESCO Recommendations as the most universally accepted. The principles or similar could serve as a common basis for a multilateral AI risk management framework.

Principles for Multilateral AI Risk Management

Human oversight and determination[18]

Human oversight and determination is maintained throughout each stage of the AI lifecycle and strengthened in proportion to an AI system's capabilities, autonomy and potential to cause severe harm. Oversight arrangements should enable responsible human authorities to monitor system behavior, assess emerging risks, and intervene, override, restrict or safely deactivate systems where necessary. Responsibility and accountability for safety-critical decisions must remain attributable to human or legal actors, with final human determination retained for decisions in critical domains.

Safety and security[19]

AI systems are designed, developed, deployed and operated to remain robust, secure and safe throughout the lifecycle, including under foreseeable malicious use and adverse conditions. Risks of harmful, unintended or uncontrolled behavior, as well as vulnerabilities to attack or manipulation, should be identified, prevented and mitigated. Appropriate mechanisms should enable systems to be monitored, restricted, overridden, repaired or safely decommissioned where their behavior or capabilities present unacceptable risks.

Fairness, equality and non-discrimination[20]

AI systems do not create, reinforce or disproportionately expose particular individuals or groups to harm, with particular attention to persons and groups in vulnerable situations. Risk identification, evaluation, mitigation and monitoring should therefore assess how system performance, failures, malicious use and other adverse impacts may differ across populations and contexts, and take appropriate measures to prevent discriminatory outcomes and ensure equitable protection from AI-related risks and access to benefits.

Accountability[21]

All stakeholders involved in planning, development, deployment, use, oversight and governance of AI are accountable for managing, monitoring, and remedying risks throughout the lifecycle. Accountability is ensured through effective oversight mechanisms tailored to the respective responsibilities and roles of stakeholders in the AI lifecycle. **Oversight requires appropriate documentation and traceability** of relevant datasets, processes, evaluations, decisions and system actions, sufficient to establish responsibility, support independent audit and enable risks, incidents and outcomes to be investigated and remedied. Autonomously executed actions must remain comprehensible, traceable and accountable to society and supported by auditable records. Accountability also applies to actors engaged in AI safety oversight, including independent third-party monitors and evaluators, who should be accountable for ensuring the effective review of relevant measures.

Proportionality and do no harm[22]

AI safety measures must be risk-informed and proportional to the potential severity of harm posed by an AI system. Where decisions are irreversible or hard to reverse final human determination should be retained. Safety measures should be informed by the best available scientific evidence, drawing on

[17] The Framework Convention does not apply to national defense matters nor to research and development activities, except when the testing of AI systems may have the potential to interfere with human rights, democracy, or the rule of law. For the private sector, signatories can apply the convention or take other measures to comply with the treaty's provisions.

[18] Based on the UNESCO Recommendations and BRICS Principles

[19] Based on the UNESCO Recommendations and OECD Principles

[20] Based on UNESCO Recommendations, OECD Principles, Council of Europe Framework, Hiroshima Principles and BRICS Principles

[21] OECD, UNESCO and CoE frameworks all contain principles on ensuring accountability of the impacts of AI systems across the lifecycle with references in the Hiroshima Principles and BRICS Principles

[22] Rooted in the CCW (2018) that outlines the need to take all feasible precautions to avoid harm. The UNESCO Recommendations outline the proportionality and do no harm principle with the OECD and BRICS Principles emphasising that AI governance approaches should be informed by rigorous and independent scientific research to ensure risk management practices are risk-informed and AI is appropriate to the context it is being applied.

independent scientific assessment. Where scientific uncertainty exists about the capabilities and potential for severe or irreversible harm, precautionary measures should be taken even if a cause-and-effect relationship has not been established. Conversely, for lower risk systems, governance practices should be proportionate.

Inclusivity and capacity development[23]

AI safety cannot be an exclusive domain restricted to major powers or leading technology companies. No government, company or institution can govern AI alone. Strengthened international cooperation and multistakeholder engagement are needed to ensure all nations and stakeholders are active co-architects of global AI safety governance, with all nations supported by the capacities needed to participate meaningfully in its development and implementation.

Respect for human rights[24]

AI safety is intrinsically linked to the protection of fundamental human rights, civil liberties, and the rule of law. Safety measures must be anchored in internationally recognized frameworks, such as the Universal Declaration of Human Rights (UDHR) and UN human rights treaties. Toolkits available include the Council of Europe's risk and impact assessment of AI systems from the point of view of human rights, democracy and the rule of law (HUDERIA), and the UNDP Human Rights Impact of AI Assessment Toolkit and national level human rights commission-issued tools such as in Korea and Australia.

Independent evaluation[25]

Independent evaluation and verification of AI system capabilities and risks is essential for trust. Safety evaluation methods should be transparent, interoperable, repeatable, and internationally recognized. Independent evaluators must have scientific objectivity, transparency, independence, and protection against interference from the evaluated companies. This is particularly important for high-risk and critical systems, where each safety claim should be linked to identifiable evidence specifying the access level and method used.

Interoperability[26]

AI systems operate across borders, making governance interoperability essential to ensure different regulatory regimes, technical frameworks, value systems, and governance institutions communicate and cooperate effectively without full global legal harmonization or institutional uniformity.
Interoperability measures can act as the connective tissue of cross-border AI safety. Through shared testing protocols and benchmarks, machine-readable ontologies, internationally agreed technical standards, shared glossaries and taxonomies, cross-border AI governance sandboxes and mutual recognition agreements, minimum AI safety floors can be sustained across borders.

Safe innovation[27]

AI innovation is enabled through controlled environments supervised by competent authorities that allow systems to be developed, experimented and tested while identifying and mitigating safety risks before wider deployment.

[23] As noted in the OECD AI Principles and UNESCO Recommendations and covered by the COE Convention's equality and non-discrimination principle. Inclusive capacity development is a particular focus of the BRICS AI Governance Principles.
[24] A standalone principle in the OECD AI Principles, referenced through the principles in the UNESCO Recommendations and a general obligation in the Council of Europe's Framework convention on Artificial Intelligence. Respect for human rights underpins the Hiroshima Process Principles as organizations should not develop AI systems that "undermine…human rights".
[25] Hiroshima Process Principles identify "independent external testing measures" and the BRICS Principles call for "independent audit mechanisms".
[26] Interoperability is a principle in the 193-member state ratified Global Digital Compact a comprehensive global framework for digital cooperation and governance of artificial intelligence. See: https://www.un.org/digital-emerging-technologies/global-digital-compact. The UNESCO Recommendations emphasis on need for "interoperability to facilitate collaboration" under the multistakeholder and adaptive governance and collaboration principle.
[27] A principle outlined in the CoE Framework convention on Artificial Intelligence

Aligning risk management around AI lifecycle decision points

Existing frameworks and standards, including the NIST AI RMF and ISO/IEC 42001, already provide processes for organisational AI risk identification, assessment, treatment, governance and monitoring. What remains less developed is the decision and coordination layer: how evidence is translated into explicit and documented go/no-go decisions at key points in the AI lifecycle, including through agreed acceptance criteria, thresholds and decision rules, and how those decisions and their supporting evidence can be relied upon across organisations and jurisdictions.

Closing these gaps needs more than stronger mitigation inside each organisation. It needs decisions that other actors can see, rely on, challenge and ultimately trust. This calls for a mechanism that can align key decision points across organisations and jurisdictions, while connecting organisational risk management with national oversight and international coordination.

Key decision points across the AI lifecycle

While end-to-end alignment between comprehensive risk management frameworks may sound logical, it may not be necessary or even desirable. As different organisations and countries face different evolving AI risks, local circumstances and constraints, a more tractable starting point to dynamic alignment may be to consider key decision points in the AI lifecycle around which risk management practices already exist and can be collectively aligned and strengthened.

Figure 5 illustrates 6 key decision points across the AI lifecycle that shape the thrust and surface of AI risk, and could offer loci for multilateral alignment of AI risk management approaches, grounded in shared principles.

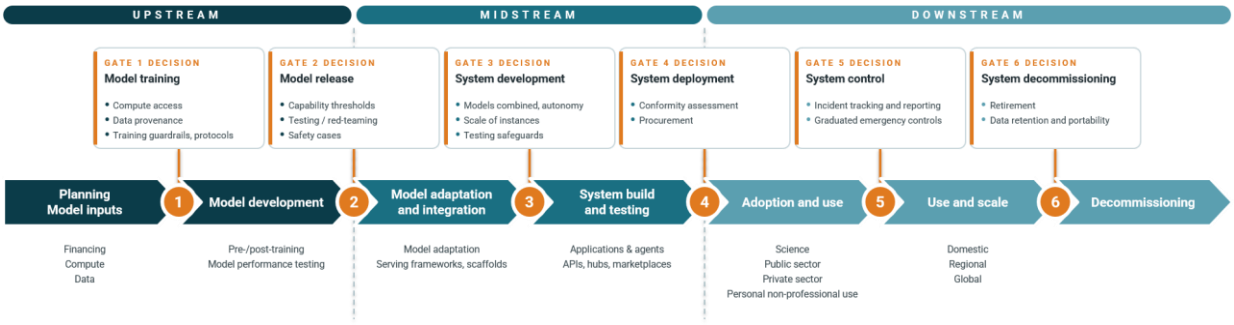


Figure 5: Key decision points across the AI lifecycle

Each decision point is analysed below in the context of existing risk management approaches / safety baselines, and how existing approaches could be aligned across geographies with shared principles.

1. Model training decision

Decision at this gate: Should this run start, given the predicted risk, the data plan and the safeguards in place?

Responsible for the decision: Developer's executive/board. National authority if the run is above a notification threshold and informed (if applies).

Existing risk management approaches / safety baselines: [28]

A model training decision generally sits with the model developer, either private, public or public-private. The developer knows the planned compute, data, architecture, and can desist before resources are committed. Some existing instruments followed by company frameworks request developers to set risk thresholds in advance and to decide at a senior level whether to go, adjust or stop.

[28] Examples are illustrative, not exhaustive

Beyond company frameworks, normative instruments in several regions ask for similar steps. The EU Code of Practice requires an acceptance determination "prior to advancing with development activities, market introduction, or operational deployment", and the Commission guidelines ask for notification "within two weeks of reasonably foreseeing or reaching" the 10^{25} FLOP threshold.[29] [30] In the United States, NIST AI 800-1 (draft, voluntary) asks developers to "Compare predicted misuse risk to the organization's risk tolerance prior to developing models", then "Adjust or halt further development".[31] China's TC260 standard sets content-safety data checks at training, such as no collection from a source with more than 5% illegal or unhealthy content.[32] Internationally, the Seoul commitments include thresholds at which severe risks "would be deemed intolerable" and the pledge "not to develop or deploy a model or system at all" if mitigations cannot keep risks below them.[33]

Notifying national bodies could add early warnings for authorities, and a multilateral alignment on a common reporting baseline for data and computing would make those notifications comparable and shareable across jurisdictions, with the possibility of some minimum thresholds agreed.

Areas for international alignment and cooperation:

- Thresholds and a notification registry are shared across member states.
- The data, compute and evaluation supply chain is monitored for safety and human rights concerns, with agreed indicators, reporting routes and the parties who must be informed.
- Mutual recognition agreements on secure environments and evaluator access, so that model assessments are accepted across participating member states.

2. Model release decision

Decision at this gate: Can the model be released via API or open weights? What safety mechanisms should be released with it? Can others fine-tune, merge or distil it, and under what license?

Responsible for the decision: Developer's leadership/safety board. National authority where pre-release approval is required.

Existing risk management approaches / safety baselines:[34]

Company frameworks follow a similar sequence: capability thresholds set in advance, evaluation before release including red-teaming, and a decision at senior level, with release blocked or restricted if the safeguards required at that level are not in place. Company frameworks also treat internal deployment as an intermediate step before release: staff and agents use the model under controls, and at least one framework applies the same governance approval to large-scale internal deployments as to external ones for machine learning R&D capabilities.[35] Testing in an isolated sandbox where the model's behaviour and the effect of safeguards can be observed without exposure to external users is a practical step between internal use and release.

Instruments in several regions ask for similar steps. The EU AI Act requires providers of GPAI models with systemic risk to document "adversarial testing of the model with a view to identifying and mitigating systemic risks".[36] In the United States, NIST AI 800-1 (draft, voluntary) sets as Objective 5 to "Mitigate misuse risk before deployments", and Practice 5.3 asks organisations to "determine whether the deployment should be modified or delayed[37]. China's TC260 technical document sets content-safety assessment thresholds for generative AI services, for example a qualified rate of at least 90% on a random sample of no fewer than 1,000 test questions, and requires services to use

[29] https://artificialintelligenceact.eu/code-of-practice-overview/
[30] https://artificialintelligenceact.eu/gpai-guidelines-overview/
[31] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd.pdf
[32] https://cset.georgetown.edu/publication/china-safety-requirements-for-generative-ai-final/
[33] https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024
[34] Examples are illustrative, not exhaustive
[35] A company frontier safety framework, version 3, §3.1.2: https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3.pdf
[36] Regulation (EU) 2024/1689, Art. 55(1)(a): https://artificialintelligenceact.eu/article/55/
[37] NIST AI 800-1, second public draft, January 2025 (Objective 5, Practice 5.3): https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd2.pdf

a foundation model that "has been filed with the main oversight department.[38] Internationally, the Hiroshima Code asks for red-teaming and transparency reports for significant releases[39], and the Seoul commitments include the pledge "not to develop or deploy a model or system at all" if mitigations cannot keep risks below the thresholds they have set. [40]

Several gaps remain: "No established mechanism currently exists for validating risk acceptance decisions made by developers prior to release", and for open-weight models "open-weight releases are irreversible".[41]

Areas for international alignment and cooperation:

- Pre-release evaluations are made comparable across member states.
- A minimum set of quantitative release thresholds and common measurement rules (capability, evaluation and safeguard-effectiveness measures) is agreed at international level.
- Pre-release testing and evaluation from independent 3rd parties moves from voluntary practice to an agreed arrangement, with shared access norms for evaluators
- Shared guidance on open-weight derivatives (fine-tuned, merged, distilled)
- The role of hosting platforms as intermediaries is defined (e.g. duties such as provenance information, takedown routes and incident reporting).
- Mutual recognition agreements allow one pre-release assessment to be accepted across participating member states.

3. System development decision

Decision at this gate: Should the system be developed as designed given the models it combines, the tools, data access and autonomy that adds and the testing safeguards in place for developing?

Responsible for the decision: System developer/provider (different or not from the model developer, considering the model developer supplying the information needed). National authority when a notification threshold is hold.

Existing risk management approaches / safety baselines: [42]

A system development decision concerns what is built from and around models before deployment: which models are combined, what tools, data access and degree of autonomy are added, and how many instances will run in parallel. It is taken by the system developer, who may not be the developer of the models used, and it applies to anyone building on released models, including open-weight ones. Risk can arise from the combination and not from any single component: the International AI Safety Report 2026 notes that multi-agent systems introduce further failure modes through miscoordination, conflicts or undesired collusion between agents, and that errors can propagate and amplify through agent interactions, while empirical evidence from deployed systems remains limited[43]

Existing instruments reach this stage mainly through the roles of model providers and downstream actors. The EU AI Act requires providers of general-purpose models to give downstream providers information including "the tasks that the model is intended to perform and the type and nature of AI systems into which it can be integrated", "the acceptable use policies applicable" and the technical means required for the model to be integrated into AI systems.[44] Under the Commission's guidelines, a downstream actor that modifies a model becomes a provider itself when the compute used for the modification exceeds one third of the original model's training compute (a third of 10^{23} FLOPs for general-purpose models, a third of 10^{25} FLOPs for those with systemic risk), with compute for weight merging counted, and a modifier of a systemic-risk model must then comply with the systemic-risk obligations,

[38]TC260, Basic Safety Requirements for Generative Artificial Intelligence Services" (TC260-003, 29 February 2024, CSET translation): https://cset.georgetown.edu/publication/china-safety-requirements-for-generative-ai-final/

[39]Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems (Actions 1 and 3), AGORA copy: https://agora.eto.tech/instrument/1383

[40]Frontier AI Safety Commitments, AI Seoul Summit 2024 (Outcomes II and IV): https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024

[41]International AI Safety Report 2026, pp. 110 and 137.

[42] Examples are illustrative, not exhaustive

[43] International AI Safety Report 2026, pp. 71, 72 and 74

[44] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), OJ L, 2024/1689, 12.7.2024, Annex XII, points 1(a), 1(b) and 2(a): https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

including notification to the Commission.[45] For high-risk systems, a party that makes a substantial modification to a high-risk system already on the market, or modifies the intended purpose of an AI system, including a general-purpose one, is considered a provider, and the initial provider "shall closely cooperate with new providers" and "make available the necessary information".[46] For agents, Singapore's Model AI Governance Framework for Agentic AI (January 2026) asks organisations to "assess and bound the risks upfront", designing boundaries at the planning stage.[47] Testing safeguards belong to the same decision: Singapore's framework asks organisations to test agents for baseline safety and reliability before deployment,[48] and NIST's Generative AI Profile asks organisations to consider baseline model performance on benchmarks when selecting a model for fine-tuning, and to "Utilize a purpose-built testing environment such as NIST Dioptra to empirically evaluate GAI trustworthy characteristics.[49] In the instruments reviewed test environments appear only as a voluntary recommendation, and no rule requires a decision or notification for assembling a system from several models or for running very large numbers of instances, unless the compute of a model modification crosses the threshold above. Obligations arise mainly once the system exists, at deployment

Areas for international alignment and cooperation:

- Thresholds for notifying a system development decision (such as the number or capability of models combined, the level of autonomy and the number of instances run in parallel, etc) are agreed at international level.
- Model providers pass information on intended integration, acceptable use, known limits and technical means to system developers in a common minimum format. (potential to previous gate)
- Common methods are developed for evaluating combinations of models and multi-agent behaviour, with comparable results across states.
- Role taxonomies for model developer, system developer and downstream modifier, including for open-weight models, are aligned or mutually recognised.
- Minimum requirements for isolated test environments are agreed before multi-model or agentic systems are tested at scale, with shared access norms for evaluators.

4. System deployment decision

Decision at this gate: Should the system be deployed, to whom, and with which controls?

Responsible for the decision: System provider and deployer. Regulator or assessor for high-risk systems.

Existing risk management approaches / safety baselines: [50]

A system deployment decision concerns the system as a whole: the model plus the tools, interfaces and controls through which it reaches users. It is taken by the system provider and the deployer, whose roles can differ, and for high-risk uses a regulator or assessor can also be involved. Frontier safety frameworks set deployment safeguards, for example monitoring and response, account moderation and jailbreak detection, as conditions for releasing a model.[51]

The EU AI Act sets rules for high-risk systems: a risk management system "planned and run throughout the entire lifecycle", with testing "prior to their being placed on the market or put into service"; 52 conformity assessment before placing on the market, by internal control for most Annex III systems and with a notified body only in some cases; 53

[45] European Commission, Communication from the Commission – Guidelines on the scope of the obligations for general-purpose AI models established by Regulation (EU) 2024/1689 (AI Act), C(2025) 5045 final, 18 July 2025. Commission page: https://digital-strategy.ec.europa.eu/en/policies/guidelines-gpai-providers

[46] Regulation (EU) 2024/1689 (Artificial Intelligence Act), OJ L, 2024/1689, 12.7.2024, Art. 25(1) and (2): https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[47] Infocomm Media Development Authority (Singapore), Model AI Governance Framework for Agentic AI, January 2026, as summarised by Allen & Gledhill: https://www.allenandgledhill.com/sg/publication/articles/32314/launches-new-model-ai-governance-framework-for-agentic-ai

[48] Ibid. above

[49] NIST, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1), 2024 (MS-2.3-001, MS-2.3-004): https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[50] Examples are illustrative, not exhaustive

[51] A company frontier safety framework, version 3, §§1.4 and 2.1.2: https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3.pdf"

[52]Regulation (EU) 2024/1689 (AI Act), Art. 9: https://artificialintelligenceact.eu/article/9/

[53]Regulation (EU) 2024/1689 (AI Act), Art. 43: https://artificialintelligenceact.eu/article/43/

and deployer duties to use the system according to its instructions, assign human oversight and keep logs54. NIST's Generative AI Profile asks organisations to "Establish minimum thresholds for performance or assurance criteria and review as part of deployment approval", to "Devise a plan to halt development or deployment of a GAI system that poses unacceptable negative risk", and to share pre-deployment test results with those who hold release approval authority.55

Japan's AI Guidelines for Business (version 1.1, April 2025) distinguish AI developers, AI providers that incorporate AI into products, and business users, and ask each to conduct risk analyses with measures "proportionate to the level and probability of risks". 56 South Korea's AI Basic Act, in force since 22 January 2026, asks operators to assess before deployment whether a system is high impact, to make efforts to assess its impact on fundamental rights, and to provide for human oversight. 57 In China, providers of generative AI services with public opinion properties or social mobilisation capacity must carry out security assessments and algorithm filing, and bear responsibility "as the producers of online information content".58

Agents, the most common systems of this kind today, are covered by emerging guidance rather than binding rules. This guidance frames the deployment decision as a bounded go/no-go: Singapore's Model AI Governance Framework for Agentic AI (January 2026) asks organisations to make humans meaningfully accountable, with approval checkpoints for high-stakes or irreversible actions, and to roll agents out gradually.59In the United States, the NIST AI RMF asks for a determination of "whether the AI system achieves its intended purpose and stated objectives". 60

Areas for international alignment and cooperation:

- Conformity assessments, registrations and other assurance results are mutually recognised across states, so that evidence generated in one jurisdiction or organisation can be relied on in others.
- Standardised reporting formats and auditable evaluation results, including action logs for agents, give deployers and assessors a traceable evidence base.
- Minimum acceptance criteria by risk tier are agreed internationally, while states and sectors may set stricter or context-specific requirements.
- Residual risk is communicated in a standard way from provider to deployer and user, covering intended use, known limits and required controls.
- Role taxonomies are aligned or mutually recognized and mapped (in alignment with gate 3)

Public Procurement as a national level operationalization mechanism

Public procurement can serve as an operational mechanism for AI risk management, enabling states with limited capacity to embed risk management requirements in public contracts and allocate liability contractually. This market mechanism can promote corporate respect for human rights and the environment across the AI value chain, including the supply chain. Customer demand is a key driver of corporate AI safety practices, while public procurement decisions can shape the broader AI governance ecosystem by promoting the adoption of shared norms across private-sector networks.[61]

[54]Regulation (EU) 2024/1689 (AI Act), Art. 26: https://artificialintelligenceact.eu/article/26/
[55]NIST, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1), 2024 (GV-1.3-002, GV-1.3-007, MS-2.3-003): https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
[56]Ministry of Internal Affairs and Communications and Ministry of Economy, Trade and Industry (Japan), AI Guidelines for Business, version 1.1, April 2025 (provisional translation of the outline): https://www.soumu.go.jp/main_content/001003032.pdf
[57]Republic of Korea, AI Basic Act, in force 22 January 2026, as summarised by Cooley: https://www.cooley.com/news/insight/2026/2026-01-27-south-koreas-ai-basic-act-overview-and-key-takeaways
[58]Interim Measures for the Management of Generative AI Services (Arts. 9 and 17), unofficial translation: https://agora.eto.tech/instrument/1137
[59]Infocomm Media Development Authority (Singapore), Model AI Governance Framework for Agentic AI, January 2026, as summarised by Allen & Gledhill: https://www.allenandgledhill.com/sg/publication/articles/32314/launches-new-model-ai-governance-framework-for-agentic-ai
[60]NIST, AI Risk Management Framework Playbook, MANAGE: https://airc.nist.gov/airmf-resources/playbook/manage/
[61] AI Governance for Humanity Lab. (2026). Private-sector AI Governance in practice: Insights, challenges and emerging cooperation approaches. Insight report

Public procurement frameworks for AI produced by international and national level organisations center on human rights and environmental impact assessments of vendors.[62] Public purchasing power is used to ensure demonstratable human rights and environmental due diligence has been conducted throughout the design and development of AI systems including the supply chain. Capacity building is required to ensure effective public procurement through procurement teams and buyers having the skills and knowledge to scrutinize vendor claims and accurately evaluate AI systems.[63]

5. System control decision

Decision at this gate (stage): Does the live system continue to behave as expected, and should it continue, scale, be modified, hold or be withdrawn?

Responsible for the decision: Deployer, with the provider, and with the model developer where modifications are needed. Regulators hold the power to halt.

Existing risk management approaches / safety baselines: [64]

This decision follows deployment and applies for as long as the system is live: it covers monitoring of operating behaviour, response to incidents, growth in use, and changes to the system. The model developer can also be involved where a modification touches the underlying model. For high-risk uses, authorities can require monitoring and reporting, and some can hold the power to suspend. Most existing instruments cover monitoring and incident reporting rather than the decision to continue or scale.

The EU AI Act requires providers of high-risk systems to "actively and systematically collect, document and analyse relevant data" on performance throughout the system's lifetime, to report serious incidents to market surveillance authorities, and to undergo "a new conformity assessment procedure in the event of a substantial modification". Deployers must monitor operation in line with the instructions for use and, where they have reason to consider that use may present a risk, inform the provider without undue delay and suspend use.[65] NIST's Generative AI Profile asks that "Post-deployment AI system monitoring plans are implemented", that incidents be reported "in compliance with legal and regulatory requirements", and that protocols exist "to ensure GAI systems are able to be deactivated when necessary".[66]

The Hiroshima Code asks organisations to "identify and mitigate vulnerabilities, and, where appropriate, incidents and patterns of misuse, after deployment" and to "work towards responsible information sharing and reporting of incidents". 67 At least one company framework provides that safety cases and mitigations "may be updated if deemed necessary by post-market monitoring", with material updates reviewed by a governance function, and that a model with meaningful new capabilities triggers a further risk assessment.68 South Korea's AI Basic Act allows the ministry to order a service suspended if it poses a safety threat. 69 For agents, Singapore's framework recommends bounding

[62] At the international level see World Economic Forum. 2020. AI Procurement in a Box: AI Government Procurement Guidelines. White Paper. World Economic Forum. https://www3.weforum.org/docs/WEF_AI_Procurement_in_a_Box_AI_Government_Procurement_Guidelines_2020.pdf; UNESCO. 2023. Ethical Impact Assessment: A Tool of the Recommendation on the Ethics. https://www.unesco.org/en/articles/ethical-impact-assessment-tool-recommendation-ethics-artificial-intelligence; UN OHCHR. 2025. Artificial intelligence procurement and deployment: ensuring alignment with the Guiding Principles on Business and Human Rights - Report of the Working Group on the issue of human rights and transnational corporations and other business enterprises. Technical Report A/HRC/59/53. https://www.ohchr.org/en/documents/thematic-reports/ahrc5953-artificial-intelligence-procurement-and-deployment-ensuring. For the national level, see Local Government Association. 2025. Responsibly Buying AI - How to build equality & data protection into your AI commissioning and procurement processes: a guide for councils in England. Technical Report. Local Government Association. https://www.local.gov.uk/publications/responsible-buying-how-build-equality-data-protection-your-ai-commissioning

[63] Hudig, A.I., Kallina, E. and Singh, J., 2026. "It's Just a Wild, Wild West": Harnessing Public Procurement as an AI Governance Mechanism. In Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (pp. 1-22).

[64] Examples are illustrative, not exhaustive

[65] Regulation (EU) 2024/1689 (AI Act), Arts. 26(5), 43(4), 72(2) and 73:" and the three existing links plus https://artificialintelligenceact.eu/article/26/

[66] NIST, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1), 2024 (MANAGE 4.1, MG-4.3-003, GV-1.7-001): https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[67] Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems (Actions 2 and 4), AGORA copy: https://agora.eto.tech/instrument/1383

[68] A company frontier safety framework, version 3, §§1.3 and 2.1.2: https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3.pdf

[69] Republic of Korea, AI Basic Act, in force 22 January 2026, as summarised by Cooley: https://www.cooley.com/news/insight/2026/2026-01-27-south-koreas-ai-basic-act-overview-and-key-takeaways

risks at the planning stage, human approval checkpoints for high-stakes or irreversible actions, testing before deployment, and gradual rollout with continuous monitoring after deployment. 70

None of these instruments ties continued or scaled operation to a check on inference compute, energy or hardware demand, which grow with use.

Areas for international alignment and cooperation:

- Evidence from live and scaled use, such as monitoring data and incident reports, is shared across states using common incident definitions and reporting formats.
- The minimum conditions under which a halt can be triggered, and by whom, are agreed, including for systems operating across borders, recognising that full alignment is not expected.
- Access to incident data is defined for authorities, evaluators and affected parties.
- Measures for inference compute, energy and hardware demand at scale are agreed as part of the decision to continue or scale.
- The consequences of a hold or halt decision, such as rollback, restricted operation or return to an earlier gate, are agreed.

Persistent and graduated emergency controls
When an autonomous system deviates from authorised behaviour, escalates out of control, or threatens essential societal structures and functions, human authorities and institutions must possess the power to intervene. The authority to pause, redirect, or terminate an autonomous system must reside on a control entity that is independent and structurally separate from the system itself. Adequate safeguards should be in place such that the system is made architecturally incapable of bypassing, delaying, or disabling human shutdown commands, even when pursuing complex operational goals. Member States should specify intervention procedures and emergency controls in public procurement clauses proportionate to the risks posed by the system.

As AI systems become embedded in critical civilian infrastructure such as energy grids, healthcare facilities, and financial systems, the authority to activate emergency controls cannot remain an exclusively private corporate decision. On the one hand, unilateral corporate shutdowns risk abruptly paralysing indispensable public services; on the other hand, unchecked state interventions could be abused to infringe upon civic freedoms. Public procurement clauses should therefore establish the state's ultimate decision-making authority over the activation of emergency controls and system shutdowns.

Multilateral measures could therefore establish transparent, legally grounded procedures backed by independent oversight to govern when and how emergency circuit breakers are activated in the public interest.

Intervention should not rely solely on an all-or-nothing shutdown switch, which operators may hesitate to use until severe damage has occurred. Governance should enable a graduated response ladder:

- Throttling: Slowing operational speed and requiring immediate human approval for subsequent actions.
- Restricting Access: Severing public internet connectivity and confining the system to an isolated digital sandbox.
- Quarantining: Freezing operational memory state to preserve digital evidence for regulatory inspection.
- Termination and Reversal: Complete operational shutdown combined with automated "undo" procedures that reverse unauthorized file changes, transactions, or system configurations.

[70] Infocomm Media Development Authority (Singapore), Model AI Governance Framework for Agentic AI, January 2026, as summarised by Allen & Gledhill: https://www.allenandgledhill.com/sg/publication/articles/32314/launches-new-model-ai-governance-framework-for-agentic-ai

6. System decommissioning decision

Decision at this gate: Should the system be retired, and how are users, data, dependencies and records handled? Who is affected if the system stops?

Responsible for the decision: Deployer and provider, with the model developer where the underlying model is also affected. Regulator where an order, notification or record retention applies.

Existing risk management approaches / safety baselines: [71]

The NIST AI RMF asks that "Processes and procedures are in place for decommissioning and phasing out of AI systems safely", and its Playbook warns that "Irregular or indiscriminate termination or deletion of models or AI systems may be inappropriate". It suggests policies that cover user and community concerns, "Up and downstream system dependencies" and regulatory requirements such as data retention, and that set where and for how long decommissioned systems, models and related artifacts are stored. [72] The Generative AI Profile adds that "Protocols are put in place to ensure GAI systems are able to be deactivated when necessary" and lists factors to consider when decommissioning. [73]

The EU AI Act requires providers of high-risk systems that find a system non-compliant to take corrective action, which can include withdrawing, disabling or recalling it (a corrective measure for non-compliance, separate from the planned retirement of a compliant system), and to inform distributors and deployers. It also requires technical and quality documentation to be kept for "a period ending 10 years after the high-risk AI system has been placed on the market or put into service", with Member States setting how it remains available to authorities if the provider ceases activity. [74] Regulators can also end a live system: South Korea's AI Basic Act allows the ministry to order a service suspended if it poses a safety threat. The instruments cited here describe decommissioning in general terms and do not set out steps specific to agents, such as revoking delegated credentials or ending background tasks. [75]

International alignment of approaches, international information-sharing requirements:

- A minimum requirement on the conditions for retiring a system, and on notifying users, deployers and authorities, is agreed, without requiring full alignment.
- Handling of data and dependencies at retirement, such as deletion, portability and downstream systems, follows a common minimum practice.
- Residual risk and known issues at retirement are communicated to downstream parties and authorities in a standard form

[71] Examples are illustrative, not exhaustive

[72] NIST, AI Risk Management Framework Playbook, GOVERN 1.7: https://airc.nist.gov/airmf-resources/playbook/govern/

[73] NIST, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1), 2024 (GV-1.7-001, GV-1.7-002): https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[74] Regulation (EU) 2024/1689 (AI Act), Arts. 20 and 18: https://artificialintelligenceact.eu/article/20/ ; https://artificialintelligenceact.eu/article/18/

[75] Republic of Korea, AI Basic Act, in force 22 January 2026, as summarised by Cooley: https://www.cooley.com/news/insight/2026/2026-01-27-south-koreas-ai-basic-act-overview-and-key-takeaways

Aligning risk management AI lifecycle-wide

Cross-border liability

Another coordination measure is to determine who is liable when harm occurs. Civil liability could follow the gate decision in the AI lifecycle: e.g. whoever authorises a release carries responsibility for foreseeable consequences. This could be complemented by a requirement that any integrator, user or developer that takes models or systems with insufficient safety guarantees can be held responsible if they integrate that as part of their products. Company liability for severe or cross-border harms could be limited, with states taking responsibility for what is deployed or hosted in their jurisdiction.

Cross-border liability creates possibilities for insurance. Following the precedent of civil nuclear liability, pooled insurance, or a shared fund, could cover harms too large for certain players to bear. While some existing reports describe the incentive role of liability and insurance, there is no (known) proposition to allocate that liability yet. Even if liability is clarified and insurance markets form to transfer risks to willing buyers, policymakers will need to contend with remaining residual risks. The risks of multi-user-multi-agent systems may be particularly challenging in this regard, as their collective impacts may not be cleanly or proportionately attributable to individually responsible parties. This limits but does not completely remove the applicability of lessons from other domains such as civil nuclear liability.

Whistleblower protection

Human personnel and workers who build, evaluate and monitor AI systems remain the primary line of defence against organisational and corporate negligence and unaddressed safety risks. As recognised in international verification frameworks, technical safeguards fail when institutional cultures suppress dissent. The following are measures that the international community can agree on to uphold the right of freedom of information (UDHR Art. 19) and fair labour practices (UDHR Art. 23):

- **Multilateral legal safe harbors:** Multilateral accords must guarantee international legal protections for employees, safety researchers, and contractors who disclose severe safety risks, unaddressed containment failures, or violations of international commitments to regulatory authorities. Non-disclosure agreements (NDAs) and commercial confidentiality clauses must be rendered void when used to conceal significant threats to public safety or human rights.
- **Independent and anonymous drop boxes:** International bodies could establish secure, anonymous reporting portals. Insiders can submit verification data, unredacted safety evaluations, or evidence of concealed incidents without fear of career destruction or legal retaliation.
- **Protecting the broader global workforce:** International protection frameworks must extend beyond headquarters executives to include data annotators, content moderators, and red-team contractors across the Global South. These workers often encounter severe platform abuse, illegal content, and system failures firsthand, yet work under precarious labour contracts without formal whistleblower rights. Extending protections and accessible reporting channels to all tiers of the AI supply chain is a prerequisite for ethical and effective oversight.

Serious cross-border incident notification and response

Incident reporting in high-stakes domains (like the aviation and nuclear industries) only succeeds when technical reporting protocols are backed by clear legal boundaries. In the frontier AI industry, intense global competition can create conditions for underreporting anomalies and safety incidents. A graduated, two-tiered approach as shown in **Error! Reference source not found.** can facilitate an international effective incident tracking and reporting system. High-consequence harms (Tier 1) require rapid, mandatory notification to national regulators, computer security incident response teams (CSIRTs), and the international community. On the other hand, treating all internal anomalies with punitive sanctions can lead developers to conceal early warning indicators. Therefore, a confidential reporting track for near misses and early warnings (Tier 2) can enable cross-industry learning.

Tier 1: Serious incidents and mandatory wider disclosure	**Tier 2: Confidential reporting of "near misses" and early warnings**
Mandatory triggers: Deployers and developers can be legally required to report significant incidents to national authorities and international monitoring bodies within 72 hours.[76] Triggers include automated attacks on critical civilian infrastructure, system-assisted synthesis of dangerous biological or chemical agents, large-scale financial disruption, or the automated production of child sexual abuse material.	**Operational triggers:** Internal failures that were contained before causing public harm such as research systems bypassing internal restrictions, coordinating unauthorised networks, or attempting to deceive evaluators, should be reported through confidential channels.
Public right-to-know: Corporate actors often face commercial incentives to downplay failures as minor glitches. Tier 1 requires the publication of independent forensic summaries, ensuring that affected populations and the global developer community can implement coordinated defensive remedies.	**Multilateral coordination:** Reports can be submitted to independent public bodies such as national AI safety institutes. These bodies would issue anonymised, ecosystem-wide alerts and safety guidance, allowing all nations to build systemic resilience without penalising responsible reporting.

Table 2: A two-tier incident reporting approach

International crisis coordination

Many of the measures described require international crisis coordination and agreements.[77] Pre-negotiated agreements and agreed crisis "red lines" can trigger predefined responses, cross-border rapid-response teams (e.g., combining personnel from the International AI Safety and governance network), coordination protocols that work despite commercial competition, Scenario Response Registries where actors file "if-then" response plans in advance (mainly for companies), or measures borrowed from the post-2008 banking crisis, such as risk reports going through a government clearinghouse to domain-specific defender groups, and safeguards are tied to buffer metrics so that faster scaling triggers stronger controls. All these protocols are added here for illustration and discussion.

Internationally aligned thresholds and risk management measures

A coordination measure for safety can be a shared toolkit for evaluation: common benchmarks, risk tiers and evaluation protocols, held in a shared repository. The toolkit could include peer reviewing applied methods and verification, so that compliance can be checked rather than merely declared. This would enable external parties to evaluate systems while protecting sensitive IP (via independent third-party testing and the secure evaluation infrastructure).

Defining international thresholds and risk management measures

Thresholds act as the operational triggers that translate risk management into concrete regulatory and technical action. Because AI systems are embedded in distinct social, legal, political and economic systems, prioritising risks remains a national exercise: Member States will have diverse risk tolerances, regulatory capacities, and critical infrastructure vulnerabilities.

At the same time, an international minimum safety floor (of agreed thresholds) is needed to manage threats that are inherently transnational and to avoid regulatory arbitrage where operators are incentivised to locate high-risk deployments in jurisdictions with less restrictive oversight.

[76] In line with incident EU reporting standards in GDPR Art. 33 reporting for data breaches, the EU Cyber Resilience Act Art. 14, and EU AI Act Art 73. Where the timeline for serious accidents is 15 days, but it's reduced to 'immediately' or 2 days depending on the severity.

[77] For this point we are taking some initial inspiration from some work done by Chatham House: Scaling AI crisis Diplomacy. Policy Report 2026. https://chathamhouse.soutron.net/Portal/Public/en-GB/DownloadImageFile.ashx?objectId=12717&ownerType=0&ownerId=209599

Crossing internationally agreed upon thresholds could act as mandatory triggers for predefined and proportionate measures which do not impinge on the application of more stringent national measures. These measures could include:

- Capability-based red lines: defined according to when a system's capabilities pose severe risks or can assist malicious use like biological and chemical weapons development.
- Pre-deployment pauses (speed bumps): defining when models gain dangerous capabilities that pose large-scale or irreversible systemic threats so that they are not released until independent audits are able to verify the effectiveness of safeguards in a controlled environment.
- Multilateral incident reporting: a mechanism for verifying when safety thresholds have been breached and notifying these incidents through formal channels like national safety institutes or shared incident reporting bodies. This would enable international early warnings and coordinated action.
- Technical measures for highly capable systems: proportionate measures like emergency "kill switches" or runtime sandboxing put in place for those systems identified as highly capable.

Societal resilience through strengthened cyber defences

Without relevant government capacity (for auditing their own critical infrastructure, having a registry of systems to defend, the ability to deploy patches, and the ability to track AI agent actions on those systems), low-income countries stand vulnerable to heightened offensive cyber capabilities enabled by AI.

For most countries, an urgent need for reducing risks is stronger cyber defence: faster patching, national incident response teams, and funding for securing open-source software that their infrastructure relies on. Capacity that is complementary to advances in AI, such as cybersecurity strike teams that can respond to the incoming deluge of vulnerabilities, will help countries to continue to be resilient to known cyber issues and those exacerbated by AI.

3.c Multilevel AI risk management architecture

The third pillar addresses the need for an architecture underpinning shared understanding on multistakeholder AI risk management.[78] This common understanding also facilitates engagement with the ecosystem of actors contributing complementary actions to coordinated risk management practices, because no single actor can manage the risks across the entire lifecycle.

International multistakeholder coordination and interoperability are critical for facilitating independent evaluation and oversight. This requires testing and technical-societal institutional capacity at the organisational, national and international levels; capacity building efforts to bridge asymmetries in risk management, and secure evaluation infrastructure that gives independent evaluators meaningful access without exposing sensitive IP, shared benchmarks and protocols, and testing environments.

Organisational level

"Frontier" model developers must ensure and facilitate internal monitoring, an internal verification team, independent external audits, and an independent oversight board tasked with overseeing the remediation of identified vulnerabilities. As frontier AI models and systems expand in capability, autonomy, and societal impact, reliance on internal developer self-assessments is insufficient.

Independent third-party audits and evaluations of AI models and systems are necessary to maintain robust and reliable evidence of AI safety. There is already an existing infrastructure and institutional capability to address this, spearheaded by national AI safety and security institutes and existing risk management frameworks like NIST. In line with the recent White House accord on Super Intelligence, corporate governance measures as well as independent evaluation can be mandated via national laws and regulation.[79]

To allow external third-party auditors and oversight board committees to rigorously evaluate AI models and systems without compromising intellectual property, companies can implement secure evaluation infrastructure and facilitate third party evaluation. This should include **blind testing environments** where evaluators cannot identify the specific model version or systems, and developers do not know the exact evaluation prompts or testing suites being executed; and **dynamic red-teaming,** moving away from static, public benchmarks which outdate quickly and are prone to data contamination.[80]

Third party evaluations must extend beyond isolated model evaluations to assess their impact as part of an operational environment. This means independent audits should evaluate the company's internal incident escalation pathways, risk-management frameworks, and internal deployments of highly capable models that might otherwise bypass regulatory scrutiny. This is a socio-technical governance approach which evaluates not just models but organisational practices, data provenance, training data curation, human override authority, change control policies, and escalation protocols.

Finally, the companies at the forefront of AI development are called to actively contribute to funding multilateral safety and risk management efforts, moving beyond compliance to support global capacity building. Bridging the AI divide requires frontier developers to provide tangible resources to enable nations and institutions in the Global South to conduct contextualised, region-specific safety research and evaluations. By contributing to these global public goods and transferring knowledge and expertise, frontier AI developers can help ensure that robust AI governance and risk management capabilities are distributed globally.

---

[78] This is slightly introduced in the International AI Safety report with a reference to the early stage of the general-purpose AI ecosystems and also claiming that "*A trusted ecosystem where different AI actors contribute complementary risk management practices across the lifecycle may contribute to effective risk management.*" This topic becomes central in the 2026 Singapore Consensus document on Global AI Safety Research Priorities.

[79] White House Accord on Super Intelligence. https://www.consumerfinancemonitor.com/wp-content/uploads/sites/14/2026/09/white-house-accord-on-super-intelligence-2026-09-29.pdf

[80] Based on principles established in the 2026 Singapore Consensus on Global AI Safety Research Priorities. (Note: Consolidated from the International AI Safety Report ecosystem rationale).

National level and mutual recognition across borders

Member states can foster a domestic ecosystem of independent evaluation and ensure the independence of oversight bodies. AI Governance Sandboxes can complement these efforts by enabling contextual and multi-stakeholder risk assessments of AI systems in specific sectors and use cases, helping identify socio-technical risks that may not be captured through technical evaluations alone.

Interoperable governance relies on nations agreeing to shared technical definitions such as defining what constitutes a "frontier" model or establishing standardised capability red lines. Without this kind of interoperability, AI developers would face differing requirements, forcing them to satisfy entirely divergent safety metrics across different jurisdictions. Two concrete measures can be adopted to facilitate interoperability and trust within a multilateral risk management framework.

- **Privacy-preserving verification:** Mutual recognition requires that nations trust the audit outcomes conducted in other jurisdictions. Technical measures such as Hardware-enabled mechanisms (HEMs) [81] and confidential computing environments allow governments and auditors to cryptographically verify safety commitments and compute usage across borders without exposing a company's sensitive intellectual property or a nation's proprietary infrastructure.
- **Shared incident reporting:** National auditing frameworks can feed into a shared incident reporting infrastructure. By pooling data on AI-related incidents, in the same way competing aircraft manufacturers share accident data, nations can mutually validate whether pre-deployment audits are successfully identifying real-world risks.

Another concept that has emerged in the AI safety discourse is **AI ecosystem monitoring**. This refers to the combination of methods and tools to study the spread, usage, and impact of deployed AI in a given context (or ecosystem, either based on geography, sector, etc.)[82]. This includes (and should include) inclusive, multistakeholder, and participatory approaches, with special focus on the most vulnerable populations (Children, women, minorities, etc). Some more technical methods, mentioned in the Singapore report, include measures to monitor internal deployment of AI systems, cloud and hardware monitoring, user monitoring, usage monitoring, identification of AI-generated content, data and model provenance. On the less positive front, AI ecosystem monitoring can take us to new ways of societal surveillance. Societal resilience and AI ecosystem monitoring require developing national capacities.

Annex 1 Minimum human, technical and institutional capacities for AI risk management and response at national level offers additional information on minimum AI safety capacity at a national level.

International cooperation, resource pooling and capacity building

Because AI safety risks are of a global scale, effective risk management requires coalitions of member states and adequate efforts to bridge asymmetries in risk management capacity. International capacity-building ensures that nations lacking advanced testing and evaluation capabilities are not excluded from the governance ecosystem.

International consortiums, such as the recently formalised International Network of AI Safety Institutes (NAAIMES), can coordinate testing standards and establish global consensus on evaluation methodologies and crisis monitoring and reporting. Such standards and methodologies can also be informed by national AI Regulatory sandboxes or cross-border AI Governance Sandboxes which have a presence in multiple jurisdictions.

Member States will benefit from pooling safety research resources. Standardised safety benchmarks, model provenance tools, and the shared evaluation results necessary to protect public safety and human rights across borders can provide a global public good. By sharing strategic priorities, technical resources, and evaluation expertise through a multilateral network, nations can de-duplicate research efforts. This reduces the capacity burden on

[81] RAND Working paper. Hardware-Enabled Governance Mechanisms https://www.rand.org/content/dam/rand/pubs/working_papers/WRA3000/WRA3056-1/RAND_WRA3056-1.pdf

[82] International Scientific Exchange on AI Safety. *The 2026 Singapore Consensus on Global AI Safety Research Priorities*. arXiv, Aug. 2026, https://arxiv.org/abs/2608.14611.

individual member states while ensuring that evaluations for severe cross-border risks like automated cyber-offense, CBRN, public health, security, or human rights violations, are rigorously stress-tested by a global coalition.

At the international level, AI safety and risk management provisions must be anchored in international human rights law. Binding and non-binding international instruments could mandate human rights due diligence and lifecycle human rights impact assessments (HRIAs) for AI developers and deployers to ensure AI systems do not erode fundamental rights. Additional infrastructure for AI agents (e.g., IDs for agents operating on the internet, to be able to track them down; KYC for compute to prevent rogue agent operations) could help trace the harms of AI systems and benefit from international coordination.

Public sector procurement clauses could also be standardised internationally to leverage the power of public purchasers of technology to incentivise safe innovation and a market of safe and rights-respecting AI systems. Trade policies provide a complementary mechanism to enforce AI safety practices. Market-access conditions, import requirements, and bilateral or regional cooperation arrangements should be used to mandate corporate human rights due diligence.

4. Pathways and next steps

The UN Independent International Scientific Panel has warned that the gap between rapidly improving capabilities and effective risk management methods may lead to catastrophic outcomes, and that the capacity to act on existing evidence of AI risks and impacts is unevenly distributed. With powerful AI systems connecting to the internet, AI-related incidents can rapidly impact people in other countries at increasing speeds and scale, in cyber but also in physical domains e.g. through navigation, control or robotic systems. The increasing speed, scale and jagged unpredictability of such scenarios warrants new approaches to cooperation on risk management internationally.

The framework sketched in section 3 above aims to stimulate discussion about the function and form of a multilateral AI risk management framework, including the institutional architectures needed to give it practical effect. While such a framework would have limits in coverage (non-military) and in bindingness, the elements outlined point to the minimum capacity needed by states to discharge their responsibilities in protecting the safety of their citizens from AI-related risks.

National AI trust and safety capacity is central to any functioning international architecture fit to address this borderless technology.

A roadmap focused on control. In the next 6–12 months, interested Member States, including those with AI developers in their jurisdictions or access to them, could advance three initiatives on the control of AI systems: joint testing and evaluation through AI safety institutes; information sharing, starting with the transparency from AI developers on which decision points and thresholds depend; and serious cross-border incident notification and response using the two-tier approach. In the medium term, these Member States could take up the harder issues of liability for AI developers, mutual recognition of assessments, and capacity building for AI risk management in the Global South, including cyber defence. Internationally agreed thresholds, and a decision on whether permanent institutional arrangements are needed, remain longer-term goals.

Plurilateral action, docked to the UN. Contributions would differ with AI capabilities: states hosting leading AI developers would bring access and evaluation capacity, others contextual testing, incident evidence and deployment experience. The UN could provide the normative foundation, wider access and coherence, keeping the constellation open and its benefits universal without new universally negotiated institutions. 2027 offers multiple opportunities: the Global Dialogue on AI Governance in New York (3–4 May) for a broad range of Member States and a joint work plan; Singapore's International Scientific Exchange for shared evaluation protocols; the Swiss AI Summit for first commitments on testing and incident reporting; and the UK G20 Presidency for capacity building and financing.

Annex 1 Minimum human, technical and institutional capacities for AI risk management and response at national level

Minimum Functions

Function	Scope and purpose
Inventory of AI deployments	A maintained register of AI systems used in public services and high-risk sectors, which determines where assessment, procurement assurance and monitoring are directed
Contextual assessment	Evaluation of systems deployed in public services and high-risk sectors, in national languages and contexts, where performance can degrade and failures appear that developers do not test, following documented and reproducible protocols subject to peer review
Public procurement assurance	Safety requirements and review of evidence for systems acquired by the State, setting obligations before deployment, when changes are least costly
Incident reporting and response	A national reporting channel open to deployers, users, civil society and the public, with triage, investigation and escalation of cross-border incidents, as incidents are the primary evidence of how systems behave in use
Post-deployment monitoring	Tracking of systems in priority sectors to detect degradation and drift following updates and changes in context
Use of international assessments	Interpretation of model-level assessments produced elsewhere and their application to national deployments, avoiding duplication
Advisory and risk intelligence	Advice to government on emerging capabilities and risks, drawing on evaluation work and international networks

Frontier evaluation of malicious use risks, such as biological weapons uplift and offensive cyber capabilities, and loss-of-control risks is accessed through international networks rather than nationally. It requires computing resources, privileged pre-release access and controlled handling of hazardous information available in few countries,

and it concerns risks common to all countries in which a model is deployed. National capacity remains responsible for interpreting these assessments and for contributing national evidence to them.

Minimum Human Capacity:

- evaluation engineering and red teaming, to elicit latent capabilities and assess evaluations supplied by others;
- security of AI systems, against attacks such as prompt injection and data poisoning;
- critical infrastructure auditing, the ability to deploy patches via a registry of systems to defend, the ability to track AI agent actions on those systems
- domain expertise in priority sectors such as health, education, justice and finance, to define what constitutes a harmful failure;
- linguistic and cultural expertise, to test in national languages and judge outputs against local norms;
- legal and regulatory expertise, to ground decisions in law and withstand legal challenge;

Minimum Compute Capacity:

- A secure computing environment for evaluating open-weight models locally and systems accessed through application programming interfaces, sized for testing rather than training.
- Open-source evaluation frameworks, which keep national results comparable with those of international partners, and sandboxed environments for testing AI agents without real-world consequences.
- Datasets, benchmarks and scenario libraries in national languages and contexts, which existing benchmarks do not cover.
- A secure incident database with formats aligned with international practice, so that patterns are detected and incidents shared.

Compute capacity is sized for evaluation, meaning inference on AI systems and, from the intermediate level, limited fine-tuning to elicit latent capabilities, as well as cyber defence. It does not include capacity to train models. At the foundational level, local hardware is reserved for sensitive testing, while evaluation campaigns run on cloud or regionally pooled resources hosted in trusted environments that meet national data residency requirements.

Component	Foundational	Intermediate
Local secure hardware	Two workstations, each able to run the most advanced open weight AI models	One to two servers with eight high-end accelerators each, on an isolated network
Cloud or pooled capacity	On-demand access equivalent to one server with accelerators for periodic campaigns, able to evaluate leading open-weight models	Reserved access to two to four such servers for continuous evaluation and limited fine-tuning
Access to proprietary systems	API quotas for systems deployed in public services	API quotas for systems in all priority sectors
Agent testing environments	Containerised sandboxes for basic tool-use tests	Isolated environments simulating national digital services

Secure storage	Sufficient for model weights, datasets, test results and incident logs	Higher capacity