Digital Policy Alert

Safety in International AI Frameworks

Nils Deeg

November 2026

An initiative of the

St.Gallen Endowment
for Prosperity through Trade

Safety in International AI Frameworks

Our text-level analysis of nine international AI frameworks shows that all nine cover safety. A total of 38 provisions address safety, with various degrees of focus and actionability.

- 29 provisions focus on safety as the core concern, while 9 provisions only touch upon safety tangentially.
- 35 provisions across all nine frameworks provide actionable measures to pursue safety. Only three provisions address safety in non-actionable language: one each in the Council of Europe Framework Convention, the UNESCO Recommendation and the UN General Assembly resolution.
- The ASEAN Guide has the most provisions on safety (7), all of which are also actionable. The ASEAN Guide, along with the G7 Principles and Code of Conduct, also has the most core provisions (5 each).

We now outline the prevalence, focus, and actionability of relevant provisions in each international framework, before summarising their language.

OECD AI Principles

Three OECD AI Principles feature safety. Principle 1.4 (Robustness, security and safety) covers it as core concern, while Principles 1.2 (Respect for the rule of law, human rights and democratic values, including fairness and privacy) and 1.5 (Accountability) cover it tangentially. All three Principles contain actionable elements with regard to safety.

Principle 1.4 states that AI systems should be robust, secure and safe throughout their lifecycle, including under foreseeable misuse ((a)). Its actionable element is the call for mechanisms ensuring that systems which risk causing undue harm or exhibit undesired behaviour can be "overridden, repaired, and/or decommissioned safely" ((b)).

In the two tangential principles, safety serves other objectives:

- Principle 1.2, on human rights and democratic values, calls for safeguards addressing risks from uses outside the intended purpose and from intentional or unintentional misuse ((b)).
- Principle 1.5, on accountability, calls on AI actors to apply a "systematic risk management approach" to each phase of the lifecycle ((c)).

UN General Assembly resolution

Six provisions in the UN General Assembly resolution A/RES/78/265 feature safety, which is core to four of them and tangentially related to two. All except one are actionable. The measures are framed as encouragement to Member States and, where applicable, other stakeholders. Out of the core provisions:

- 6(b) encourages measures for the internationally interoperable identification, classification, evaluation, testing, prevention and mitigation of vulnerabilities and risks before deployment;

- 6(c) encourages feedback mechanisms through which end-users and third parties can discover and report technical vulnerabilities, misuses and incidents after deployment, so that they can be addressed;
- 6(f) encourages investment in safeguards including physical security, AI system security and lifecycle risk management.
- 6(e) focuses on the development of risk monitoring and management mechanisms and impact assessments. It also mentions measures touching other concerns, such as data security, privacy policies, impact assessments and their disclosure.

Out of the tangential provisions:

- 6(l) names risk and impact assessments as safeguards for the protection of human rights;
- 6(a) names safe, secure and trustworthy AI only as an objective of domestic regulatory frameworks, and is therefore non-actionable.

Council of Europe Framework Convention

Three articles in the Convention feature safety, which is core to all of them. Two of the three are actionable. The Convention is the only instrument examined that contains “shall” obligations. Convention articles bind state parties rather than AI actors directly, and several are softened by qualifiers.

Article 16, on the risk and impact management framework, focuses on risks to human rights, democracy and the rule of law, while also containing the Convention’s most detailed safety measures. It requires measures for the identification, assessment, prevention and mitigation of risks (16(1)), graduated by the severity and probability of impacts (16(2)(b)) and applied iteratively across the lifecycle (16(2)(d)). These measures include monitoring (16(2)(e)) and, where appropriate, testing before first use and after significant modification (16(2)(g)). Parties must also address adverse impacts (16(3)) and assess the need for a moratorium, ban or other measures for uses incompatible with human rights, democracy or the rule of law (16(4)). It is the only provision in the dataset to mention a moratorium.

Article 13, on safe innovation, calls upon Parties to enable controlled environments for developing, experimenting and testing AI systems under the supervision of competent authorities.

Article 12, on reliability, requires Parties to take “as appropriate” measures to promote reliability and trust in outputs. It names requirements on “adequate quality and security” only as a possible measure, and is therefore non-actionable.

G7 Hiroshima Process Guiding Principles

Five principles are about safety, which is a core concern of all of them. Further, all five principles are actionable with regard to safety. The Principles are voluntary and addressed to organisations developing advanced AI systems:

- Principle 1, on identifying, evaluating and mitigating risks across the lifecycle, suggests diverse internal and independent external testing, including red-teaming, and mitigation of identified risks and vulnerabilities.

- Principle 2, on vulnerabilities, incidents and misuse after deployment, suggests monitoring for vulnerabilities, incidents, emerging risks and misuse, taking action to address them, and facilitating third-party and user discovery of issues.
- Principle 6, on security controls, suggests physical security, cybersecurity and insider threat safeguards, including securing model weights, algorithms and servers through operational security and access controls. The G7 documents are the only frameworks in the dataset to address the security of model weights.
- Principle 5, on governance and risk management policies, suggests accountability and governance processes to evaluate and mitigate risks throughout the lifecycle.
- Principle 8, on research and investment in mitigation, recommends research and investment into mitigation tools.

G7 Hiroshima Process Code of Conduct

The Code of Conduct builds on the G7 Principles, adding detail in several cases. Six provisions of the Code feature safety, as the core concern of five and tangential to one. All provisions are actionable with regard to safety. For the five core provisions the Code shares with the Principles, the Code's relevance and actionability tags are identical. Its added value therefore lies in elaboration:

- Provision 1 adds that testing should take place in secure environments at several checkpoints, in particular before deployment and placement on the market. It lists risks to which testing should attend: chemical, biological, radiological and nuclear risks; offensive cyber capabilities; risks to health and safety, including control of physical systems and interference with critical infrastructure; self-replication; societal risks; threats to democratic values; and chain reactions affecting up to an entire city or community. It is the only provision in the dataset to name chemical, biological, radiological and nuclear risks or self-replication.
- Provision 2 adds bounty systems, contests and prizes to incentivise the responsible disclosure of weaknesses.
- Provision 5 includes additional details on the risk management framework that should be implemented across the AI lifecycle.
- Provision 6 adds a cybersecurity risk assessment, cybersecurity policies, secure environments with limited access for model weights, a vulnerability management process with regular review, and an insider threat detection programme.
- Provision 8 adds a commitment to proactively manage risks of advanced AI systems.

One provision contains a safety element with no tagged counterpart in the Principles. Provision 10, on international technical standards, is tangential and actionable: it encourages contributions to standards that organisations can use when developing their testing methodologies and cybersecurity policies.

ASEAN Guide on AI Governance and Ethics

Seven provisions in the ASEAN Guide are relevant to safety. Safety is the core concern of five provisions and tangential to two. All seven provisions contain an actionable safety element. The Guide is addressed to deployers and developers. It follows the AI lifecycle, from project definition through modelling and outcome analysis to deployment and monitoring.

Five provisions take safety as their core concern:

- Principle 3, on security and safety, suggests risk assessments and relevant testing or certification before deployment, and mechanisms for systems to disengage safely. On security, it names AI-specific attacks, such as data poisoning, model inversion, dataset tampering and byzantine attacks in federated learning. It suggests authentication, encryption, software updates, access management and incident response plans, a minimum list of security tests such as penetration testing, and business continuity and disaster recovery plans. The two ASEAN guides are the only frameworks in the dataset to name specific attacks on AI systems.
- Principle 7, on robustness and reliability, suggests resilience to unexpected inputs and stressful conditions, access control for critical systems, mitigation of unreliable performance and rigorous testing before deployment.
- Section 2, on determining the level of human involvement, suggests a risk management approach at every lifecycle stage. Risks are assessed by the probability and severity of harm, and risk management and mitigation plans are continually reviewed.
- The guidance on deployment and monitoring suggests a secure and resilient architecture and performance monitoring with defined metrics and thresholds. It also suggests revalidation and drift monitoring, testing whether systems generalise well and "fail gracefully", and regular model tuning followed by renewed testing.
- The guidance on modelling suggests adversarial testing before deployment and close monitoring of continuously learning systems for unintended behaviour. Its robustness component sits alongside material on explainability and auditability.

The guidance on project governance and problem statement definition is tangential to safety. In service of project governance, it suggests risk-based assessments before any data collection or modelling, covering foreseeable accidental or malicious misuse. Its mitigation measures include regular monitoring and maintenance, procedures for previously unknown risks and safe disengagement mechanisms. It also suggests action to resolve non-compliance by third-party developers or vendors.

The guidance on outcome analysis is also tangential to safety. It suggests confirming that pre-deployment purposes are actually fulfilled by a system and conducting acceptance tests that may include security and performance evaluations.

ASEAN Expanded Guide on AI Governance and Ethics – Generative AI

The Expanded Guide on Generative AI builds on the 2024 Guide, adapting its underlying principles to the generative AI context. Four provisions in the Expanded Guide are about safety, and it is the core concern of all four. All four are actionable:

- Section 1.3 restates Security and Safety as a principle and mentions safety elements specific to generative AI. These include fine-tuning techniques such as reinforcement learning from human feedback, red-teaming to expose unsafe behaviours, and protection against prompt injection and data poisoning. They also include dynamic testing for unforeseen behaviours, measurable benchmarks for acceptable outcomes, and guardrails

against out-of-context prompts. The section acknowledges that reliable common practices for evaluating generative AI safety still require further research.
- Section 3.3, on trusted development and deployment, suggests integrating safety best practices across the generative AI lifecycle, naming practices and guardrails, such as adversarial testing, and continuous system monitoring.
- Section 3.5, on testing and assurance, suggests third-party testing and, eventually, formal auditing to complement internal evaluation. As areas for ASEAN to explore, it proposes regional benchmarks and testing tools reflecting ASEAN's languages and cultures, and an agreed list of preferred evaluation tools. It illustrates automated benchmarking and red-teaming with an open-source toolkit.
- Section 3.6, on security, addresses generative AI-specific threats such as adversarial machine learning. It proposes vulnerability detection programmes such as bug bounties and ethical hacking, security knowledge-sharing, and security-by-design guidelines aligned with existing international standards. Bug bounties also appear in G7 Code Provision 2.

UNESCO Recommendation on the Ethics of AI

Three provisions in the UNESCO Recommendation feature safety: two core and one tangential. Two of the three are actionable.

Safety is the core concern of the principle 'Safety and security' (paragraph 27). It is non-actionable: harms and vulnerabilities are to be avoided, addressed, prevented and eliminated, though no safety-specific mechanism to do so is named.

Safety is also core to Policy Area 2, a broad governance section addressed to Member States. Its safety component includes:

- certification mechanisms with different levels of audit and a regular monitoring component, requiring re-certification if necessary (paragraph 56);
- international standards describing "measurable, testable levels" of safety, and research on safety and security risks (paragraph 64);
- self-assessment by public authorities, with a prohibition of use where adoption would violate human rights obligations (paragraph 57).

Safety is tangential to Policy Area 1, on ethical impact assessment. In service of that assessment, it calls for frameworks to identify risks and appropriate prevention, mitigation and monitoring measures (paragraph 50). It also calls for broad testing, including in real-world conditions, of systems identified as potential risks to human rights before market release (paragraph 51), monitoring of all lifecycle phases (paragraph 52), and external review of AI systems (paragraph 53).

Joint Declaration on AI, Freedom of Expression and Media Freedom

One provision, on journalists and media actors, is tangentially about safety, as well as being actionable. In service of responsible journalism, it calls on journalists and media actors to perform a "systematic risk assessment" as a precondition for deploying AI systems for journalistic purposes ((b)).

Digital Policy Alert

Contact
Written by Nils Deeg
Please send questions and suggestions to nils.deeg@sgept.org

The independent Digital Policy Alert is a pillar of the Swiss-based
St. Gallen Endowment for Prosperity Through Trade

St.Gallen Endowment
for Prosperity through Trade