Digital Policy Alert

# How Governments Address AI Safety

Tommaso Giardini and Nils Deeg

November 2026

An initiative of the
St.Gallen Endowment
for Prosperity through Trade

# How Governments Address AI Safety

In 2026, AI safety entered the headlines. Reports of AI intruding into external systems started with the OpenAI/Hugging Face incident and were followed promptly by similar disclosures from Anthropic, Google, and Meta. These incidents, alongside a wave of cyber incidents across Europe, raised political attention on AI safety. Moreover, both Anthropic and OpenAI cited cybersecurity as justification for delaying the release of advanced models, Mythos and Astra.

AI systems can pose risks to people, property and critical infrastructure, whether through malfunction, misuse or adversarial attack. While cyberattacks dominated headlines, governments built safety rules that span a system's whole lifecycle, including design and testing requirements during development, risk management, monitoring, and corrective measures following deployment, and outright prohibitions of certain systems that pose unacceptable risks.

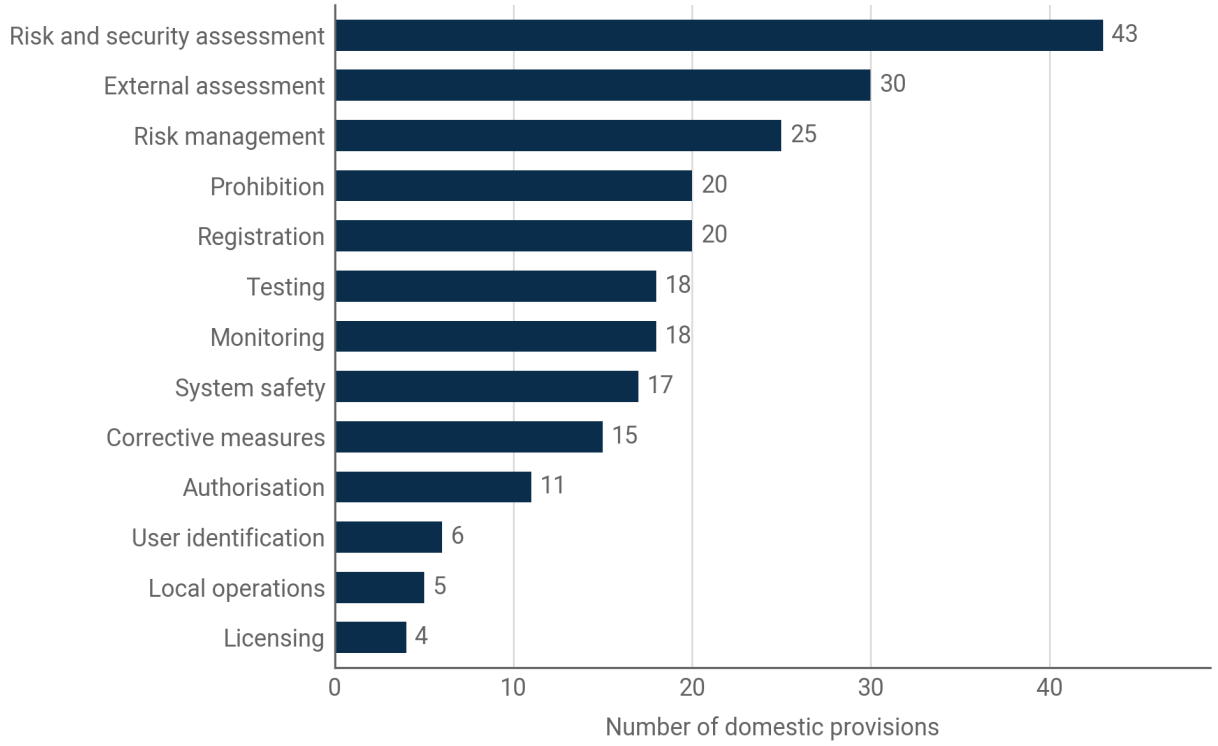
## Domestic Regulation

Safety is the most common concern in domestic AI regulation: The DPA has documented 498 domestic policy and enforcement developments on AI safety since January 2020. The DPA's systematic taxonomy is used to show the following trends in AI safety regulation:

- The US is the most active jurisdiction, with 86 federal and 59 state-level developments, followed by China (58), the EU (47) and Vietnam (26).
- The current status of developments is spread: 220 developments are in force or adopted, 243 are currently under deliberation and 35 are concluded, rejected, revoked or under investigation or litigation.
- Most safety rules are found in binding laws (214) or binding orders (117), compared to 126 non-binding guidelines. Only 41 enforcement cases and inquiries focus on AI safety.

Recent developments include the EU's requests for information to OpenAI and Anthropic regarding cybersecurity incidents and the White House's voluntary Accord on Super Intelligence with six AI firms and Gold Eagle Initiative on AI cybersecurity clearinghouse rules. Over the past years, declarations including language on AI safety were issued following the AI Summits in Bletchley Park, Seoul, Paris, and Delhi.

Our text-level analysis of 84 rulebooks across 19 jurisdictions reveals how governments address AI safety in detail. We found that governments use thirteen different regulatory requirements that appear a total of 232 times: requiring risk and security assessments of AI systems (43 provisions), external assessments by government bodies or third parties (30), risk management measures (25), prohibitions of certain types of AI systems (20), public registration of AI systems (20), testing for safe functioning and conformity with standards (18), monitoring system behaviour post-deployment (18), system safety measures to prevent unauthorised access and ensure reliable functioning (17), taking corrective measures if issues arise after deployment (15), authorisation of AI systems before market entry or deployment (11), user identification (6), local operations requirements for foreign firms (5), and licensing of AI firms or systems (4).




## International Coordination

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

International frameworks address different safety requirements. The most commonly mentioned requirements are risk and security assessment (8) and risk management (8), followed by testing and monitoring (7 each). Corrective measures and system safety are also common (6 each), while external assessment (5) appears in just over half of the frameworks. Prohibition is only mentioned by the UNESCO Recommendation and the Council of Europe Framework Convention (2). Registration, licensing, authorisation, local operations and user identification are not mentioned in any of the analysed international frameworks.

International AI frameworks differ in how they focus on safety and how actionable their language is. Of the 38 provisions, 29 focus on safety as the core concern, while nine only touch upon safety tangentially. 35 provisions across all nine frameworks provide actionable language, and only three address safety in non-actionable language (one each in the Council of Europe Framework Convention, the UNESCO Recommendation, and the UN General Assembly resolution).

| Framework | Risk management | Risk and security assessment | Testing | Monitoring | System safety | Corrective measures | External assessment | Prohibition |
|---|---|---|---|---|---|---|---|---|
| ASEAN Expanded Guide (Generative AI) | ■ | ■ | ■ | ■ | ■ | – | ■ | – |
| ASEAN Guide | ■ | ■ | ■ | ■ | ■ | ■ | ■ | – |
| Council of Europe Framework Convention | ■ | ■ | ■ | ■ | ■ | ■ | – | ■ |
| G7 Hiroshima Code of Conduct | ■ | ■ | ■ | ■ | ■ | ■ | ■ | – |
| G7 Hiroshima Guiding Principles | ■ | ■ | ■ | ■ | ■ | ■ | ■ | – |
| Joint Declaration | – | ■ | – | – | – | – | – | – |
| OECD AI Principles | ■ | – | – | – | – | ■ | – | – |
| UNESCO Recommendation | ■ | ■ | ■ | ■ | – | – | ■ | ■ |
| UNGA Resolution | ■ | ■ | ■ | ■ | ■ | ■ | – | – |

## Alignment Between Domestic and International Action

Risk and security assessment is the main requirement of AI safety at the domestic and international level. No requirement is more common domestically, and eight of nine international frameworks include it. Risk management is just as widespread internationally. Testing, system safety, corrective measures and monitoring also line up well. Gaps between the two levels are visible in prohibition and external assessment. Prohibition is among the four most common domestic requirements, yet only two international frameworks contain it, while external assessment ranks second at home but features in five international frameworks. Registration, licensing, authorisation, local operations and user identification are purely domestic requirements. The international language behind this overlap is largely concrete: Nearly all provisions are actionable, and three in four focus mainly on safety.

To understand how governments tackle AI safety in depth, readers can explore our AI Governance Hub's safety page. It shows the relevant DPA entries and the full-text provisions of domestic AI regulations and international AI frameworks, structured using our taxonomy. The page draws data from the DPA's Activity Tracker, which provides daily updates on digital policy and provides tailored notifications, as well as our full-text analysis tool Clairk.

Digital Policy Alert

**Contact**
Written by Tommaso Giardini and 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