Digital Policy Alert

How Governments Address Human Oversight of AI

Tommaso Giardini and Nils Deeg

November 2026

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

How Governments Address Human Oversight of AI

In June 2025, Anthropic tested 16 leading AI models: Each model read an inbox including information that an executive planned to shut it down that day and that he was having an affair. Most models decided to blackmail the executive. The test occurred in a fictional setup and Anthropic claimed not to have seen such behaviour in real-world deployment. But as AI systems become increasingly capable, a core concern for governments and firms is ensuring that humans retain control over AI decisions, particularly in high-risk contexts. The design of effective human oversight mechanisms is imperative, but complicated by the speed, scale, and opacity at which cutting-edge AI systems operate.

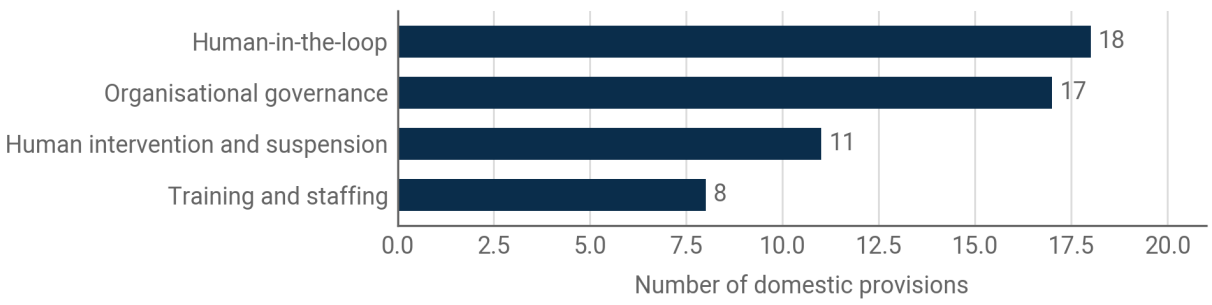
Domestic Regulation

Human oversight is not yet a common concern in AI regulation: The DPA has documented 62 domestic policy and enforcement developments on human oversight of AI since January 2020. The DPA's systematic taxonomy reveals the following trends in AI human oversight regulation:

- The US is the most active jurisdiction, with 7 state-level and 2 federal developments, followed by the EU (7), China (6), and Saudi Arabia and Singapore (3 each).
- Over half of all developments are in force or adopted (33), while 27 are currently under deliberation and 2 are concluded or revoked.
- Human oversight rules are split almost evenly between binding laws (21), non-binding guidelines (20) and binding orders (18). Only 3 enforcement cases and inquiries focus on human oversight of AI.

Recently, China mandated organisational oversight mechanisms in its Interim Measures for Administration of Anthropomorphic AI Interaction Services, while Illinois mandated internal governance structures in its AI Safety Measures Act. Further, an AI Emergency Button Act requiring advanced AI systems to incorporate a human shut-down capability was introduced to both chambers of the US Congress while, in Australia, a Bill mandating human system shutdown capacities was introduced to Parliament.

Our text-level analysis of 84 rulebooks across 19 jurisdictions reveals how governments address human oversight in AI. We find that governments use four different regulatory requirements that appear 54 times: requirements to have a human-in-the-loop to supervise the functioning of an AI system (18 provisions), implementing organisational governance structures, roles and procedures aimed at enabling effective supervision of AI (17), human intervention and suspension capabilities to override outputs or halt system functioning (11), and providing appropriate training and staffing for employees working with AI (8).



International Coordination

Our text-level analysis of nine international AI frameworks shows that eight cover human oversight across 16 provisions, all except the G7 Principles.

International frameworks address all four human oversight requirements identified at the domestic level: The most commonly mentioned requirement is human-in-the-loop (6), followed by organisational governance (5) and training and staffing (5). Human intervention and suspension (3) appears less often. The ASEAN Guide is the only framework to address all four requirements, while the G7 Principles contain no human oversight provision.

Framework	Human-in-the-loop	Organisational governance	Training and staffing	Human intervention and suspension
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	■	–	–	■

Differences are visible in how strongly international frameworks focus on human oversight and how actionable their language is. Of the 16 provisions, eight focus on human oversight as the core concern, while eight only touch upon it tangentially. 15 provisions across eight frameworks provide actionable language, and only one, in the Council of Europe Framework Convention, addresses human oversight in non-actionable language.

Alignment Between Domestic and International Action

Domestic regulations and international frameworks address human oversight similarly. Human-in-the-loop is most common on both levels, followed by organisational governance. Domestically, human intervention ranks third and training fourth, while internationally, the order is reversed. No requirement is common at one level and rare at the other. The international language is largely concrete: Nearly all provisions are actionable, though only half focus on human oversight.

To understand how governments tackle AI human oversight in depth, readers can explore our AI Governance Hub's human oversight 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