Digital Policy Alert

Human Oversight in International AI Frameworks

Nils Deeg

November 2026

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

Human Oversight in International AI Frameworks

Our text-level analysis of nine international AI frameworks shows that human oversight is found across eight frameworks, only being absent from the G7 Principles. A total of 16 provisions address human oversight, with various degrees of focus and actionability.

- Eight provisions focus on human oversight as the core concern, while eight provisions only touch upon human oversight tangentially.
- 15 provisions across eight frameworks provide actionable measures to pursue human oversight. Only one provision, in the Council of Europe Framework Convention, addresses human oversight in non-actionable language.
- The ASEAN Guide and UNESCO Recommendation have the most provisions on human oversight (4 each), all of which are actionable. Further, the UNESCO Recommendation also has the most core provisions on human oversight (3).

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

OECD AI Principles

Two OECD AI Principles address human oversight. Principle 1.2 (Respect for the rule of law, human rights and democratic values, including fairness and privacy) and Principle 1.4 (Robustness, security and safety) are both tangentially related to human oversight. Both Principles provide actionable measures.

- Principle 1.2 focuses on human rights and democratic values. To that end, it calls on AI actors to implement safeguards such as a "capacity for human agency and oversight" ((b)). The purpose of such mechanisms is also to address risks from uses outside the intended purpose and from intentional or unintentional misuse.
- Principle 1.4 focuses on safety. It suggests mechanisms ensuring that AI systems which risk causing undue harm or exhibit undesired behaviour can be "overridden, repaired, and/or decommissioned safely" ((b)).

UN General Assembly resolution

One provision in the UN General Assembly resolution A/RES/78/265, sub-paragraph 6(k), relates to human oversight, which has core relevance and is actionable. Its human oversight component encourages the promotion of oversight through review of automated decisions and related processes and, where appropriate and relevant, human decision-making alternatives. As elsewhere in the resolution, these measures are framed as encouragement to Member States and, where applicable, other stakeholders. Alongside calling for human oversight, the provision also calls for transparency measures, which are more closely covered in the transparency section.

Council of Europe Framework Convention

Two Convention Articles relate to human oversight, which is core to one and tangential to the other. However, only the tangential provision contains an actionable concern. The Convention is the only

instrument examined that contains “shall” obligations. These bind Parties, requiring them to adopt regulations in accordance with its provisions, rather than applying to AI actors directly.

Article 8, on transparency and oversight, is the core provision. It requires Parties to ensure that “adequate transparency and oversight requirements” are in place, tailored to specific contexts and risks. However, it does not identify any concrete oversight measure, and is therefore non-actionable for human oversight. Its one specified element, the identification of AI-generated content, concerns transparency.

Article 20, on digital literacy and skills, is tangentially related to human oversight. It requires Parties to encourage and promote digital literacy for all segments of the population. Its actionable element for human oversight is a call for the promotion of “specific expert skills” for those responsible for identifying, assessing, preventing and mitigating risks posed by AI systems.

G7 Hiroshima Process Code of Conduct

The Code of Conduct contains a single provision with a human oversight element, Provision 5, which is tangential and actionable. The human oversight element was absent from the G7 Principles, and therefore represents an addition that is unique to the Code of Conduct.

Provision 5, on AI governance and risk management policies, focuses on risk management. In service of that objective, it suggests that organisations put in place “appropriate organizational mechanisms” to develop, disclose and implement risk management and governance policies. It also suggests policies, procedures and training to ensure that staff are familiar with their duties and with the organisation’s risk management practices.

ASEAN Guide on AI Governance and Ethics

Four provisions in the ASEAN Guide address human oversight. It is the core concern of two provisions and tangential to two. All provisions contain an actionable human oversight element. The Guide is addressed to deployers and developers and offers worked examples, named tools and options scaled to organisational size.

Two sections take human oversight as their core concern:

- Section 1, on internal governance structures and measures, suggests a multi-disciplinary central governing body with a representative composition, such as an AI Ethics Advisory Board or Ethics Committee. It also suggests escalating higher-risk use cases to a governing body with higher authority, illustrated with a three-level model, and setting clear roles and terms of reference for oversight bodies. Escalation of this kind appears in no other framework examined. The section further covers codes of conduct, oversight mechanisms to ensure governance guidelines are followed, periodic assessment of the governance model, and guidance and training for personnel, including external certification. It acknowledges that a central governing body may be too onerous for smaller companies and offers a risk-based alternative.
- Section 2, on determining the level of human involvement in AI-augmented decision-making, suggests grading human oversight by the probability and severity of harm. It distinguishes human-in-the-loop, human-over-the-loop and human-out-of-the-loop arrangements. It

suggests that high-risk systems adopt a human-in-the-loop approach in which humans can assume full control. Under human-over-the-loop, humans can intervene, take over or alter parameters during operation. It is the only provision to grade human involvement by risk and to warn against automation bias, or "rubber stamping risk".

In the two tangential provisions, human oversight serves other objectives:

- Principle 3, on security and safety, suggests precautions so that humans can intervene to prevent harm. It also suggests implementing an appropriate level of human intervention when deploying systems, such as mechanisms for drivers of autonomous vehicles to resume manual driving.
- The guidance on deployment and monitoring suggests processes to ensure that the appropriate level of management is aware of performance issues. It also suggests autonomous monitoring "to effectively scale human oversight".

ASEAN Expanded Guide on AI Governance and Ethics – Generative AI

The Expanded Guide on Generative AI builds on the 2024 Guide. One provision, Section 3.3, focuses on human oversight, which is of core relevance and actionable.

Section 3.3, on trusted development and deployment, combines safety techniques, model evaluation and the disclosure of safety practices with a human oversight component. That component has two parts. The first is an overarching governance structure of roles, responsibilities and oversight within each organisation in the value chain. The second is human moderation, listed among deployment guardrails specific to generative AI.

UNESCO Recommendation on the Ethics of AI

Four provisions feature human oversight: three are core and one is tangential, while all are actionable. The principles are addressed to AI actors, while the policy areas are addressed to Member States.

Human oversight is the core concern of three provisions:

- The principle 'Human oversight and determination' (paragraphs 35 and 36). It holds that the decision to cede control in limited contexts remains with humans and that, as a rule, "life and death decisions" should not be ceded to AI systems. It is the only provision in the dataset to rule out ceding such decisions. It also attributes ultimate responsibility to physical persons or legal entities and extends human oversight to inclusive public oversight.
- Policy Area 2 is a broad governance section. Its human oversight component suggests that public entities, companies and civil society organisations consider an independent AI Ethics Officer or another mechanism to oversee ethical impact assessment, auditing and continuous monitoring (paragraph 58). Its other references to human oversight, in the judiciary (paragraph 63) and in regulatory frameworks (paragraph 68), name it only as a principle.
- Policy Area 11, on health and social well-being, contains a human oversight component within broader material on health care, privacy, consent and research. It calls for final decisions on diagnosis and treatment to be taken by humans (paragraph 123(e)). It also calls

for patients, professionals or caregivers to be included as “domain experts” in development teams (paragraph 123(b)) and for review by an ethical research committee before clinical use (paragraph 123(f)).

Human oversight is tangential to Policy Area 1, on ethical impact assessment. In service of impact assessment, it calls on Member States and private companies to develop due diligence and oversight mechanisms (paragraphs 51 and 53).

Joint Declaration on AI, Freedom of Expression and Media Freedom

One provision relates to human oversight, which has tangential relevance. The provision includes actionable human oversight elements. The Declaration addresses states, companies, as well as financial institutions and investors.

The section addressed to the private sector combines human rights due diligence, data protection, remedies and media relations with a human oversight component. Companies should offer reviews of content governance decisions to be conducted with human oversight where greater contextual analysis is needed ((h)).

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