Digital Policy Alert

Information Integrity in International AI Frameworks

Nils Deeg

November 2026

An initiative of the

St.Gallen Endowment
for Prosperity through Trade

Information Integrity in International AI Frameworks

Our text-level analysis of nine international AI frameworks shows that five of the nine frameworks cover information integrity. It does not appear in the G7 Principles, the Council of Europe Framework Convention, the ASEAN Guide, and the UN General Assembly resolution. A total of seven provisions address information integrity, with various degrees of focus and actionability.

- Four provisions focus on information integrity as the core concern, while three provisions only touch upon information integrity tangentially.
- Three provisions across two frameworks, the UNESCO Recommendation and the Joint Declaration, provide actionable measures to pursue information integrity. The remaining four provisions, in the OECD AI Principles, the G7 Code of Conduct, the ASEAN Expanded Guide and the Joint Declaration, address information integrity in non-actionable language.
- The Joint Declaration has the most overall provisions on information integrity (3), all of which are core, as well as the most actionable provisions of any framework (2).

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

OECD AI Principles

One OECD AI Principle addresses information integrity. Principle 1.4 (Robustness, security and safety) covers it tangentially and is non-actionable. It calls for mechanisms, where technically feasible, "to bolster information integrity while ensuring respect for freedom of expression" ((c)), without identifying any specific measures.

G7 Hiroshima Process Code of Conduct

The Code of Conduct builds on the G7 Principles, adding detail in several cases. One provision of the Code addresses information integrity. The provision is tangential and non-actionable and has no counterpart in the Principles. Provision 1, on risk identification and mitigation, lists the facilitation of disinformation as one of several risks to which testing should attend, and names disinformation as a research topic.

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. One provision, Section 1.3, tangentially addresses information integrity. It does not name an actionable measure. Under its privacy principle, it notes that malicious actors may use generative AI to create inauthentic images, videos and text, but describes this only as a risk. Its references to fine-tuning that minimises unsafe outputs and to guardrails against out-of-context prompts are framed as safety and robustness practices, not as restrictions on generating specific content.

UNESCO Recommendation on the Ethics of AI

One provision, Policy Area 9, relates to information integrity as a core concern. The provision is also actionable with regard to information integrity. The policy area is addressed to Member States.

Policy Area 9, on communication and information, calls on Member States to ensure that AI actors respect freedom of expression and access to information in automated content generation, moderation and curation (paragraph 113). Frameworks, including regulation, should ensure transparency of online communication and information operators and users' access to a diversity of viewpoints. They should also provide "prompt notification" to users of the reasons for removal or other treatment of their content, and appeal mechanisms that allow users to seek redress. The policy area also calls for AI to improve access to information (paragraph 112), for media and information literacy to counter disinformation, misinformation and hate speech (paragraph 114), and for enabling environments for media (paragraph 115).

Joint Declaration on AI, Freedom of Expression and Media Freedom

Three provisions focus on information integrity as a core concern. Two of the three are actionable. Freedom of expression is the Declaration's overall subject, and it addresses states, companies, financial institutions and investors, as well as journalists and media actors.

The section addressed to states is core and actionable. It is the only provision in the dataset to call for restrictions on the use of AI to generate certain types of content, and to protect political commentary and satire:

- Paragraph (g)(ii) calls on states to adopt or implement legislation to "prohibit, investigate and prosecute" AI-generated online violence against women, minorities and other groups at risk, and ensure its implementation by law enforcement, the judiciary and private actors.
- Paragraph (c) states that restrictions on the right to seek, receive and impart information enacted in response to AI-related risks must be prescribed by law, pursue a legitimate aim and be necessary and proportionate. Such restrictions should not suppress political commentary, satire or other legitimate expression under international law.
- Paragraph (d) states that public authorities should refrain from using AI to spread hate speech, disinformation and propaganda.

The section addressed to the private sector is core and actionable. Paragraph (h) calls for clarity, explainability and accessibility on the use of AI for content moderation, content curation and targeted advertising. Further, organisations should provide an avenue for appealing content governance decisions promptly, within a specified time limit. Further, paragraph (i) calls for accessible and effective remedies for violations of freedom of expression.

Principle 2, on the promotion of a healthy, pluralistic and diverse information environment, is core but non-actionable. It states that AI should enhance public trust in "the reliability and integrity of the information order". It also states that the concentration of corporate power poses a risk to pluralism and should be mitigated through human rights-based regulation, without naming more specific measures.

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