Digital Policy Alert

# How Governments Address Information Integrity and AI

Tommaso Giardini and Nils Deeg

November 2026

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

# How Governments Address Information Integrity and AI

In January 2026, users on X (formerly Twitter) started asking its Grok chatbot to undress people in photos, including minors. Governments worldwide reacted promptly: Within a month, Australia, Brazil, Canada, the EU, Hong Kong, India, Indonesia, Malaysia, the Philippines, the Republic of Korea, the UK, and the US (at the federal and state level) took action, ranging from banning the app, to opening investigations, to demanding explanations. Beyond the generation of prohibited content, AI facilitates the impersonation of public figures. In November 2026, China's Cyberspace Administration reported removing over 47,000 items in a campaign against AI impersonation.

AI can shape public discourse, including through AI-driven feeds and generative AI systems that produce and amplify information at scale. Governments have responded by walking a thin line between setting guardrails around AI-generated content, including the moderation of certain content, and maintaining freedom of expression, by protecting users' lawful speech.

## Domestic Regulation

Information integrity is a common concern in AI regulation: The DPA has documented 187 domestic policy and enforcement developments on information integrity in AI since January 2020. The DPA's systematic taxonomy reveals the following trends in AI information integrity regulation:

- The US is the most active jurisdiction, with 39 federal and 35 state-level developments, followed by Australia and the UK (14 each), China (13), Argentina and Brazil (9 each), and the EU (8).
- Over half of all developments are currently under deliberation (104), while 60 are in force or adopted and 23 are concluded, rejected, revoked or under investigation.
- Information integrity rules are most often found in binding laws (78), followed by 64 enforcement cases and inquiries, the second-highest enforcement count of any concern. There are 25 non-binding guidelines and 20 binding orders.

For example, the US TAKE IT DOWN Act's notice-and-removal duties entered into force in May 2026: Platforms that host user-generated content and make it publicly accessible must implement a mechanism enabling identifiable individuals to request the removal of non-consensual intimate visual depictions. In October 2026, Brazil's Superior Electoral Court required Meta to apply identical distribution criteria to political content relating to candidates in the presidential election, including with regard to content generation with the Meta AI tool. The UK's Ofcom, meanwhile, narrowed its Online Safety Act investigation into X to focus on the platform's assessment of illegal content risks, reporting tools, and handling of user reports.

Our text-level analysis of 84 rulebooks across 19 jurisdictions shows how governments address information integrity in AI in detail. We find that governments use three regulatory requirements to pursue information integrity, which appear a total of 41 times in domestic provisions, namely: generation restrictions on producing certain AI content such as non-consensual deepfakes (29 provisions), content moderation obligations to remove prohibited AI-generated content (11), and reporting channels for users to flag illegal or harmful AI-generated content (1).

## International Coordination

Our text-level analysis of nine international AI frameworks shows that five cover information integrity across seven provisions, all except the G7 Principles, the Council of Europe Framework Convention, the ASEAN Guide and the UN General Assembly resolution.

International frameworks address different information integrity requirements. Two frameworks, the UNESCO Recommendation and the Joint Declaration, refer to measures protecting users' lawful speech from unjustified removal, through notification and appeal mechanisms, while only the Joint Declaration refers to the restriction of the generation of certain types of content with AI. None of the frameworks address obligations to moderate AI-generated content or channels for users to report it. The emphasis of international frameworks thus lies on protecting expression rather than on restricting content, driven by the Joint Declaration.

There are differences in how these international frameworks focus on information integrity and how actionable their language is. Of the seven provisions, four focus on information integrity as the core concern, while three only touch upon it tangentially. Three provisions across two frameworks (the UNESCO Recommendation and the Joint Declaration) provide actionable language. Four frameworks (the OECD AI Principles, the G7 Code of Conduct, the ASEAN Expanded Guide and the Joint Declaration) address information integrity in non-actionable language.

## Alignment Between Domestic and International Action

Domestic regulations and international frameworks share only one information integrity requirement. Content generation restrictions, the most common domestic requirement, appear in a single international framework. Content moderation and content reporting exist in domestic rulebooks alone. The international language behind this overlap is mostly not concrete. Fewer than half of the provisions are actionable, all of them in the UNESCO Recommendation and the Joint Declaration, although information integrity is the core subject of more than half of the provisions.

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