Digital Policy Alert

The Fabric of AI Governance

How International Frameworks and Domestic Rules Tackle Ten Common Concerns on AI

Tommaso Giardini and Nils Deeg

November 2026

An initiative of the

St.Gallen Endowment
for Prosperity through Trade

The Fabric of AI Governance

How Domestic Regulations and International Frameworks Tackle Ten Concerns on AI

As governments worldwide start regulating AI, this report reveals which policy concerns they prioritise and what regulatory requirements they impose to tackle each concern.

- We focus on ten concerns: transparency, safety, consumer protection, data governance, human oversight, discrimination and bias, information integrity, labour, intellectual property, and accountability. Governments currently prioritise transparency and safety in AI rulebooks.
- We distinguish 50 types of regulatory requirements that governments use to address these ten concerns. The most common regulatory requirements include disclosure, information sharing, and documentation (transparency), as well as risk and security assessment (safety).

We deconstruct AI governance at the international and domestic level, on two layers of detail:

- To give a broad overview of global activity, we use the Digital Policy Alert (DPA) database, which tracks over 28,000 global digital policy developments. We find 1,756 entries that target the AI sector specifically and address one of the ten concerns.
- To provide detailed insight on regulatory approaches, we use the full-text libraries on domestic AI regulation and international AI coordination on Clairk. At the domestic level, we analyse 84 AI laws and regulations, finding 454 provisions that establish 935 regulatory requirements. At the international level, we investigate nine international AI frameworks, identifying 203 mentions of the ten concerns.

The three sections below analyse how governments address each concern in more depth.

1. The first section briefly summarises our findings on the domestic and international level, showing whether the two levels are aligned, across all ten concerns.
2. The second section synthesises domestic and international activity for each concern.
3. The third section breaks down the language of international AI frameworks on each concern.

Beyond the report, readers can engage with our data in real time with unprecedented sophistication:

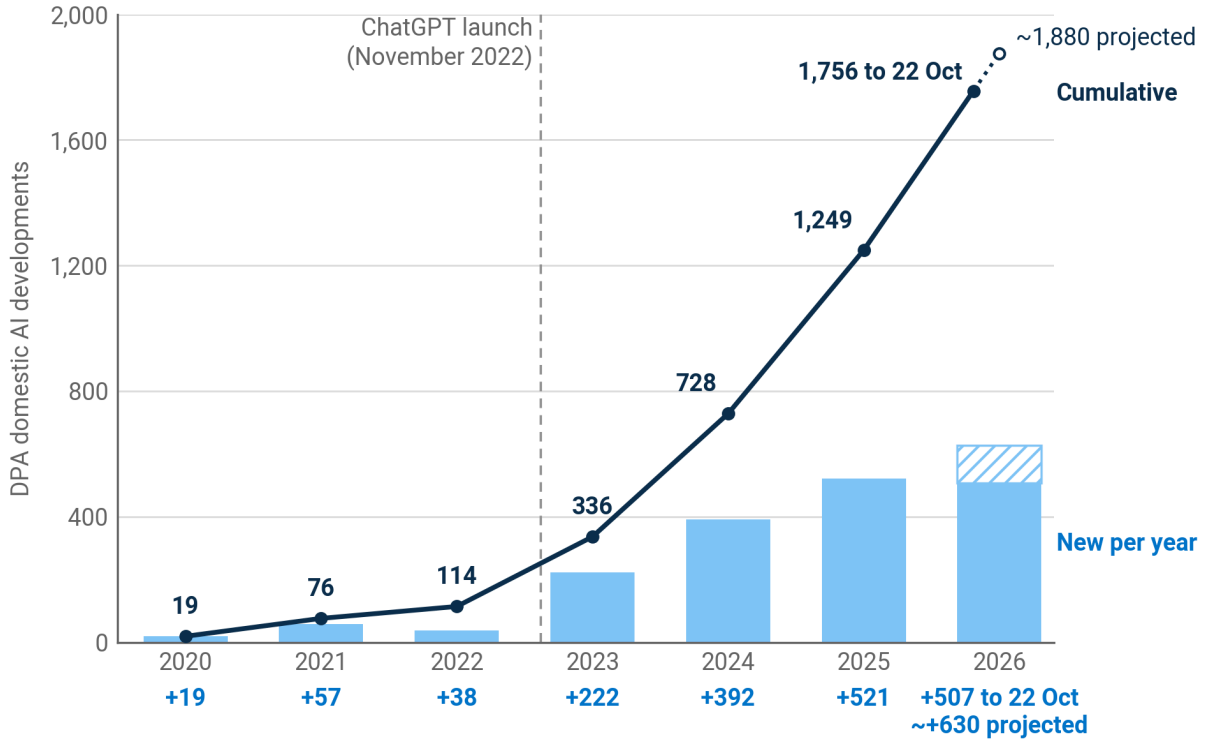
- The DPA's AI Governance Hub dedicates a page to each concern, showing both the relevant DPA entries and the provisions of domestic AI regulations and international AI frameworks in full text, structured using our taxonomy.
- The DPA's Activity Tracker provides daily updates on government regulation of the digital economy, beyond AI. Every entry provides a description and the official government source. Users can filter and download the dataset and set up tailored notifications.
- The St.Gallen Endowment's technology suite empowers users to interact with our data, including our text analysis tool Clairk, our chat system (drawing exclusively from verified data), and our API and MCP integrations, which feed our data directly into users' systems.

We thank the Patrick J. McGovern Foundation for the generous support that made this report and the Digital Policy Alert's AI Governance Hub possible. This report considers policy developments until 22 October 2026 but considers more recent real-life examples. It builds on this methodology.

Executive Summary

Which concerns do domestic AI regulations address?

AI regulation is accelerating rapidly: The DPA has documented 1,756 domestic policy and enforcement developments that target the AI sector and address one of the ten concerns, between 1 January 2020 and 22 October 2026. Notably, AI regulation has grown significantly since the launch of ChatGPT in November 2022. Governments have become more active every year, with total activity growing by a factor of 15 in only four years. In 2026, the DPA documented a record average of 12 weekly developments, totalling 507 just up to 22 October.

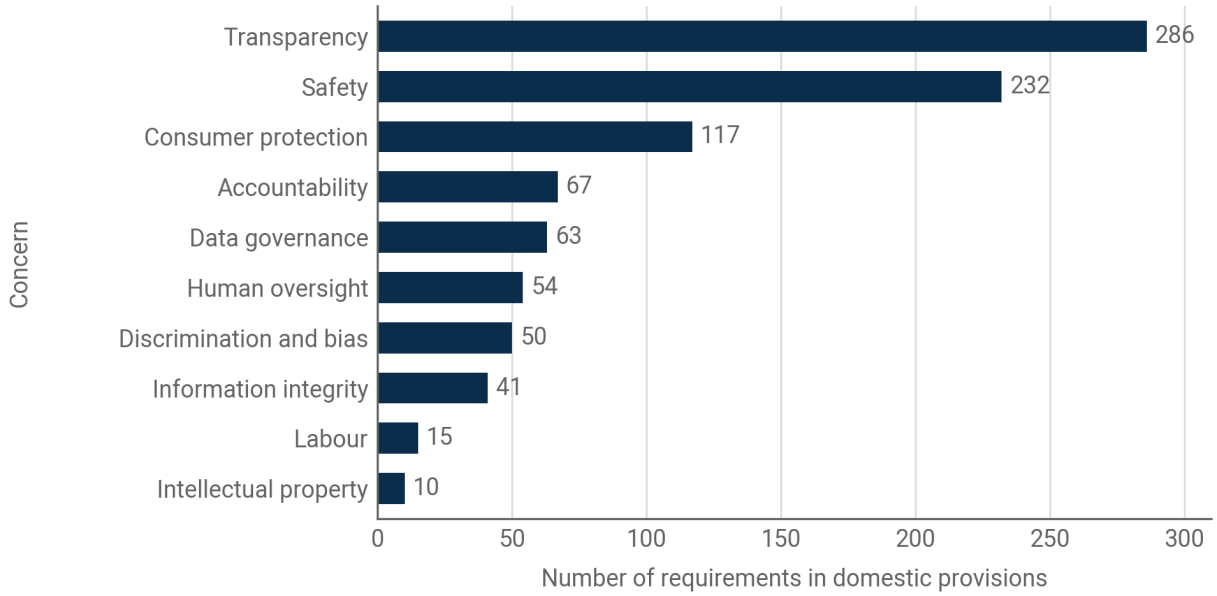


As governments worldwide tackle concerns raised by AI, differences emerge across borders. The DPA's systematic taxonomy reveals trends in the 1,756 developments.

- The United States is the most active jurisdiction, with 276 federal and 244 state-level developments, followed by the European Union (EU, 148), China (131), the United Kingdom (UK) and Republic of Korea (93 each). Other active governments include Australia (54), Vietnam (44), Argentina (42), and Brazil (40).
- 821 developments are under deliberation, while 759 are adopted or in force, and 176 are concluded or rejected. Pending developments are common in the US, at both the federal and state level (173 and 134), while the EU (100) and China (80) have more adopted ones.
- Laws and bills are the main legal form (695), followed by 419 non-binding guidelines, 348 enforcement cases and inquiries, and 294 binding orders. Compared to other jurisdictions, US states focus most often on laws (189), China on binding orders (65), and the UK on enforcement cases (37).

The policy concerns raised by AI are common across borders. This report focuses on ten AI concerns that governments are currently addressing: transparency, safety, consumer protection, data governance, human oversight, discrimination and bias, information integrity, labour, intellectual property, and accountability. First, we analyse which concern each of the 1,756 AI-related DPA entries addresses. Then, we analyse the text of 84 AI rulebooks across 19 jurisdictions to document the regulatory requirements governments use to tackle concerns.

The crux lies in the regulatory requirements, of which we have identified 50 different types, each addressing one of the ten concerns. Our text-level analysis identifies 454 unique provisions that establish one of these regulatory requirements, adding up to a total of 935, whereby the same provision can include different types of regulatory requirements. Transparency (286) and safety (232) account for more than half of all requirements, followed by consumer protection (117). Accountability (67[1]), data governance (63), human oversight (54), discrimination and bias (50), and information integrity (41) each draw several dozen. Labour (15) and intellectual property (10) draw the fewest, to date.[2]



How do international frameworks cover these concerns?

Different governments addressing the same concern creates an opportunity for international coordination. The DPA has tracked over 170 international developments that focus on AI and address one of the ten concerns. Coordination is growing, with 42 new frameworks between 1 January and 22 October 2026 alone. Most frameworks (167) are bi- or plurilateral and a majority are not binding (150).

As on the domestic level, we analyse which concern each framework addresses on the DPA and then go into text-level detail on select agreements. We focus on nine frameworks from various

[1] In accountability, we do not cover regulatory requirements but rather the targets and consequences thereof. See below.
[2] Less salient concerns, including competition and sustainability, are not in scope. The corpus of analysed texts does not cover guidelines, enforcement action, and non-AI-specific rulebooks. Expanding the corpus would increase counts for labour and intellectual property, as evidenced by the DPA counts (see below). Finally, our taxonomy of regulatory requirements will evolve as new approaches emerge.

international organisations (including UN bodies, the OECD, the Council of Europe, and ASEAN) that contain language guiding AI regulation or establishing principles that regulation could codify.[3]

Our analysis of the nine international frameworks identifies 203 mentions of the ten concerns. Beyond the mention of a concern, we analyse whether a provision focuses on a concern (core) or only references it (tangential), as well as whether a provision specifies an action that governments or firms should take to address the concern (actionable), or not.

Framework	Transparency	Safety	Data governance	Discrimination and bias	Human oversight	Intellectual property	Consumer protection	Accountability	Labour	Information integrity
ASEAN Expanded Guide (Generative AI)	A	A	A	A	A	A		A	N	N
ASEAN Guide	A	A	A	A	A	N	A	A	A	
Council of Europe Framework Convention	A	A	N	N	A		A	A		
G7 Hiroshima Code of Conduct	A	A	A	A	A	N	A			N
G7 Hiroshima Guiding Principles	A	A	A	A		N				
Joint Declaration	A	A	A	A	A	A	A	A	N	A
OECD AI Principles	A	A	N	N	A	N	A	A	N	N
UNESCO Recommendation	A	A	A	A	A	N	A	A	A	A
UNGA Resolution A/RES/78/265	A	A	A	A	A	N	A		N	

Core Tangential **A** actionable (at least one provision) **N** non-actionable only

This reveals which concerns the international frameworks mention, as well as their level of focus and actionability for each:

- Transparency (with 47 total mentions), safety (38), data governance (26), and discrimination and bias (22) appear in all nine frameworks. Human oversight (16) and intellectual property (12) appear in eight frameworks, consumer protection (11) in seven, accountability (17) and labour (7) in six, and information integrity (7) in five.
- Of the 203 concern mentions, a large part are actionable (158, 78%). Actionability is highest for human oversight (94%), safety (92%), consumer protection (91%), and transparency (89%). It is lowest for labour and information integrity (43% each) and especially for intellectual property, where only two of twelve provisions (17%) are actionable.
- In terms of focus, in over half (106, 52%) of the 203 mentions the concern at hand is the focus of the provision. Safety is the exception, with 29 (76%) core mentions. Most other concerns sit between 41% and 59%. Intellectual property is, again, at the lower end, with only one in twelve mentions found in a provision dedicated to intellectual property.

[3] As with the domestic level, our scope cannot cover all international frameworks that contain language on AI, including documents issued following AI Summits in Bletchley Park, Seoul, Paris, and New Delhi. Our selection of frameworks also includes two pairs of closely related documents, namely the G7 and the ASEAN, which can inflate counts.

These different characteristics of international AI frameworks also intersect: Transparency, for instance, is the only concern that every framework treats as core and backs with at least one actionable provision. Accountability appears in only six frameworks, but all six contain both a core and an actionable provision on it. In information integrity, more than half of the mentions are in dedicated provisions (57%), but fewer than half are actionable (43%). The frameworks also differ: The UNESCO Recommendation addresses all ten concerns, nine of them as core, while the OECD AI Principles also address all ten but treat only three as core.

Are domestic AI rules and international AI frameworks aligned?

We find commonalities and differences between how governments address the ten concerns at the domestic and international level: Transparency and safety lead at both levels and together account for 55.4% of domestic and 41.9% of international concern mentions. Accountability, human oversight, and information integrity make up a similar share of mentions at both levels. A mismatch is visible in consumer protection, which is twice as salient at the domestic level. The opposite is visible in data governance, discrimination and bias, labour, and intellectual property, which appear more often at the international level.

The alignment at the requirement level is the main focus of this report. We provide a brief overview here and investigate each concern in detail in the dedicated sections below.

- In six of ten concerns, the most common domestic requirement is also most frequently mentioned in international frameworks. This includes disclosure for transparency, risk and security assessment for safety, general data protection for data governance, human-in-the-loop for human oversight, copyright for intellectual property, and worker protection for labour. In the other concerns, the requirement found most often in domestic rules is rarer or absent at the international level.
- Since the domestic level is generally more detailed than the international level, 10 domestic requirements across four concerns appear in none of the analysed international frameworks. Only two requirements appear internationally but not domestically.

The next section provides an overview of domestic and international activity, analysing alignment between the two levels, for each concern:

- Transparency
- Safety
- Consumer Protection
- Data Governance
- Human Oversight
- Discrimination and Bias
- Information Integrity
- Labour
- Intellectual Property
- Accountability

We then turn to the international level, analysing the text of nine international frameworks in detail.

How Governments Address Ten Common Concerns on AI

Transparency

In the second half of 2026, two leading AI labs, Anthropic and OpenAI, announced watermarking mechanisms for text written by their AI systems. Both firms highlighted the EU AI Act as the reason. Anthropic mentioned that it would apply watermarks globally at launch given the absence of a durable way to scope it by region.

Governments worldwide are concerned with the opacity of AI systems, because a lack of transparency can prevent users from making informed choices and government authorities from overseeing AI systems. Governments thus require AI firms to increase transparency about the use and functioning of AI systems, as well as the output of their systems, among others.

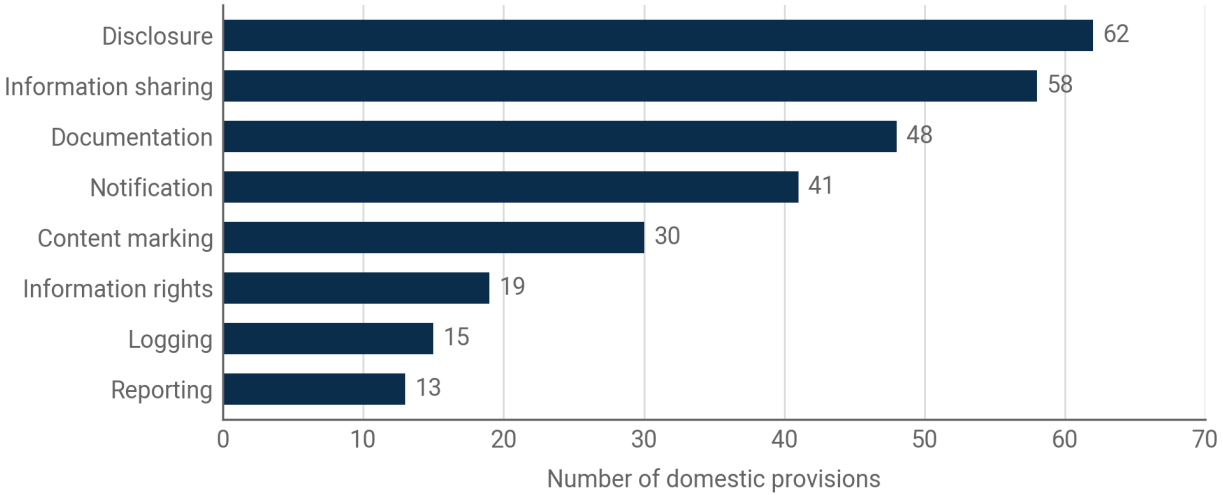
Domestic Regulation

Transparency is a common concern in domestic AI regulation: The DPA has documented 198 domestic policy and enforcement developments tackling AI transparency since January 2020. The DPA's systematic monitoring reveals the following trends in AI transparency regulation:

- The US is the most active jurisdiction, with 42 state-level and 29 federal developments, followed by China (19), the EU (16) and the Republic of Korea (12).
- The current status of developments is diverse: 79 developments are in force or adopted, 101 are currently under deliberation and 18 are concluded, rejected or revoked.
- A large majority of transparency rules are established in binding laws (123) or binding orders (36), compared to 32 non-binding guidelines and 7 enforcement cases and inquiries.

Recently, in August 2026, transparency obligations in the California AI Transparency Act and the EU AI Act took effect. In addition, Indonesia's Ministry of Communication and Digital Affairs announced that it was considering regulations regarding the labelling of AI-generated digital media content, while the Republic of Korea's Fair Trade Commission held a consultation on proposed guidelines on AI labelling and advertising.

Our text-level analysis of 84 rulebooks across 19 jurisdictions reveals how governments approach AI transparency in detail. We identified eight different regulatory requirements that appeared 286 times, namely: the prominent disclosure that AI is in use (62 provisions), sharing information on how AI systems work and should be used (58), documentation of information on AI systems and their development (48), notification of regulators or affected parties after incidents (41), marking AI-generated content through labels or watermarks (30), giving users information rights, such as an explanation of AI-assisted decisions (19), logging system inputs, outputs, and events (15), and periodic reporting on AI systems to regulators or the public (13).



International Coordination

Our text-level analysis of nine international AI frameworks shows that all cover transparency across 47 provisions. The frameworks address several of the regulatory requirements identified at the domestic level. The most commonly mentioned requirements are disclosure (9), documentation (8), and information sharing (8), followed by information rights (7). Content marking (6), notification (5), and reporting (4) are less common. Finally, logging is only explicitly mentioned in the ASEAN Guide.

Framework	Disclosure	Information sharing	Documentation	Information rights	Content marking	Notification	Reporting	Logging
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	■	■	–	■	■	–	■	–

International AI frameworks, however, differ in how they focus on transparency and how actionable their language is. Of the 47 provisions, 23 focus on transparency as the core concern, while 24 only touch upon transparency tangentially. 42 provisions provide actionable language, and only five

address transparency in non-actionable language (in the G7 documents, the Joint Declaration and the UN General Assembly resolution).

Alignment Between Domestic and International Action

Domestic regulations and international frameworks mostly address AI transparency in similar fashion. Disclosure, information sharing and documentation are the three most common requirements at both levels. The two levels diverge on logging and information rights: logging is a routine domestic requirement but appears in only one framework, while information rights rank sixth domestically yet appear in seven international frameworks. Regarding notification, content marking and reporting, the two levels are roughly aligned. The international language behind this overlap is largely concrete. Nearly all provisions are actionable, although only about half are mainly about transparency.

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

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.

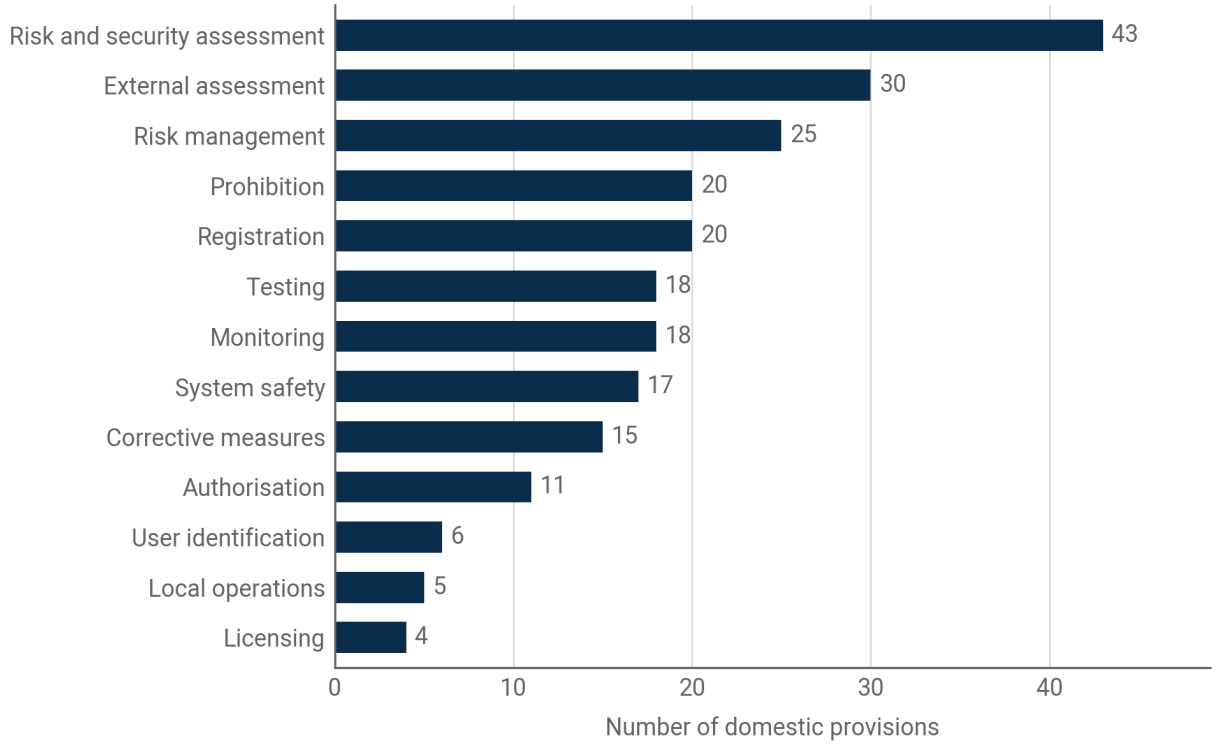
Domestic Regulation

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.

Consumer Protection

AI capabilities are advancing at a pace that is hard to follow for experts, let alone everyday consumers. At the same time, governments are reining in AI firms that exaggerate the capabilities of their AI systems: The US Federal Trade Commission prohibited DoNotPay from claiming that its "AI lawyer" service performs like a real lawyer without evidence, while the Italian consumer authority secured commitments from DeepSeek, Mistral, and Nova to warn users about hallucinations.

As AI enters more realms of our professional and personal life, protecting consumers is a concern shared across borders. Governments worry that AI can deceive users, manipulate their behaviour, and lead to addiction. They thus regulate AI firms' and AI systems' behaviours, while establishing safeguards to protect vulnerable groups, such as minors.

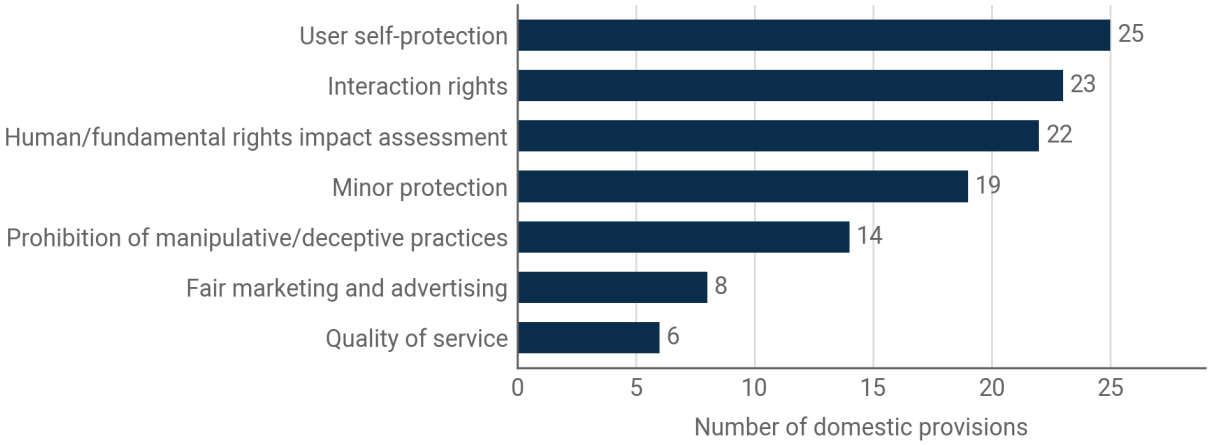
Domestic Regulation

Consumer protection is a common concern in AI regulation: The DPA has documented 242 domestic policy and enforcement developments on consumer protection in AI since January 2020. The DPA's systematic taxonomy reveals the following trends in AI consumer protection regulation:

- The US is the most active jurisdiction, with 65 state-level and 49 federal developments, followed by the EU (19), China (18) and the UK (12).
- The current status of developments is spread: 84 developments are in force or adopted, 124 are currently under deliberation and 34 are concluded, rejected, revoked or under investigation.
- Most consumer protection rules are found in binding laws (127) or binding orders (24), compared to 33 non-binding guidelines. Enforcement is frequent, with 58 enforcement cases and inquiries, the third-highest count of any concern.

Recent developments include the entry into force of China's interim measures on anthropomorphic AI interaction services, requiring providers to warn users who become reliant on the service and remind them after two hours of continuous use, an assessment by Brazil's Ministry of Justice (digital rights secretariat) of 13 AI companion services against the Consumer Protection Code and the children's digital statute, and the introduction of an Irish Bill barring AI companion services from misleading users into believing the system is human.

Our text-level analysis of 84 rulebooks across 19 jurisdictions reveals how governments address consumer protection in AI in detail. We find that governments use seven different regulatory requirements to pursue consumer protection, which appear a total of 117 times. The requirements span: safeguards to protect users against addiction, emotional dependency and self-harm (25 provisions, "user self-protection"), rights for users to opt out of AI, file complaints or request human review (23, "interaction rights"), human or fundamental rights impact assessments (22), minor protection through usage restrictions and parental controls (19), prohibitions of manipulative or deceptive practices such as subliminal techniques (14), fair marketing and advertising rules with regard to the capacities of AI systems (8), and quality of service rules ensuring stable and continuous service (6).



International Coordination

Our text-level analysis of nine international AI frameworks shows that seven cover consumer protection across 11 provisions, all except the G7 Principles and the ASEAN Expanded Guide on Generative AI.

International frameworks address only some consumer protection requirements. The most commonly mentioned requirement is human/fundamental rights impact assessment (4), followed by minor protection (3). The prohibition of manipulative or deceptive practices (2) and interaction rights (2) appear only in the UNESCO Recommendation and the ASEAN Guide. User self-protection (1) appears only in the UNESCO Recommendation. Fair marketing and advertising and quality of service are not mentioned in any framework.

Framework	Human/ fundamental rights impact assessment	Minor protection	Prohibition of manipulative/ deceptive practices	Interaction rights	User self-protection
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	■	–	–	–	–

There are differences in how international AI frameworks focus on consumer protection and how actionable their language is. Of the 11 provisions, six focus on consumer protection as the core

concern, while five only touch upon it tangentially.Ten provisions across all seven frameworks provide actionable language.

Alignment Between Domestic and International Action

Consumer protection features far more heavily in domestic rulebooks than in international frameworks. None of its requirements reaches more than four of the nine international frameworks. User self-protection and interaction rights, the two most common domestic requirements, show the sharpest contrast, appearing in only one and two international frameworks. Human rights impact assessments travel furthest, with four international frameworks covering them. Fair marketing and advertising and quality of service have no international counterpart at all. Finally, where international frameworks address consumer protection, their language is largely concrete. Nearly all provisions are actionable, and about half focus mainly on consumer protection.

To understand how governments tackle AI consumer protection in depth, readers can explore our AI Governance Hub's consumer protection 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.

Data Governance

Four months after ChatGPT's launch, the Italian data protection authority temporarily suspended it, prohibiting OpenAI from processing Italian users' data with immediate effect on 30 March 2023. The authority cited the lack of legal basis for the mass collection and storage of personal data used to train ChatGPT, missing age checks, and a data breach as reasons. After a month, ChatGPT was reinstated and the authority fined OpenAI EUR 15 million, which a court overturned in 2026. Concerns about data governance and AI, however, keep growing since.

AI models are trained on enormous volumes of data and continue learning as we use them, raising concerns about how data is collected, stored, and processed. Governments regulate the composition of training data, the security of data during usage, as well as the rights of individuals with regard to their personal information when interacting with AI systems.

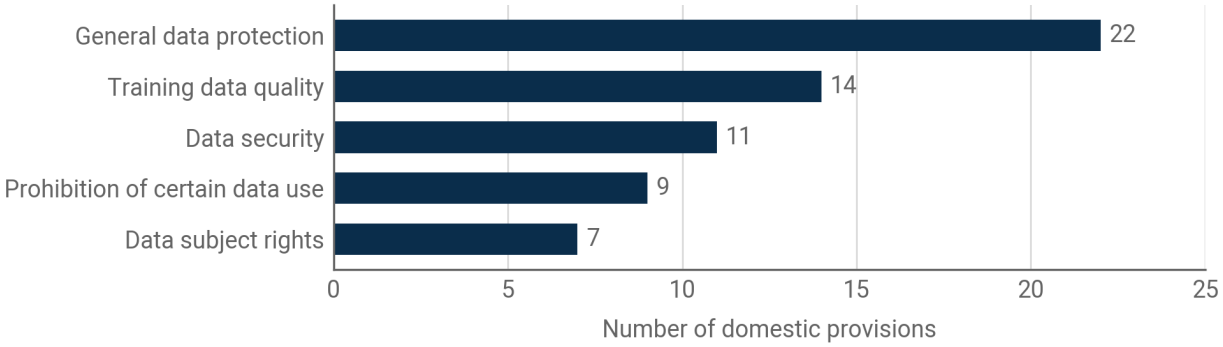
Domestic Regulation

Data governance is a common concern in AI regulation: The DPA has documented 367 domestic policy and enforcement developments on data governance in AI since January 2020. The DPA's systematic taxonomy reveals the following trends in AI data governance regulation:

- The Republic of Korea is the most active jurisdiction (42), followed by the US with 21 federal and 20 state-level developments, China (34), the EU (31) and the UK (21).
- Over half of all developments are in force or adopted (192), while 132 are currently under deliberation and 43 are concluded, rejected, revoked or under investigation or litigation.
- Non-binding guidelines and enforcement outnumber binding rules. There are 143 non-binding guidelines and 116 enforcement cases and inquiries, more enforcement than any other concern. Binding laws number 67 and binding orders 41.

Recently, in September 2026, the President of the Republic of Korea signed an amendment to the Personal Information Protection Act creating a personal information use exception for AI development. Further, UK lawmakers introduced the Personal Data (Digital Twins) Bill, which would require explicit consent before a digital twin of a person can be created. Oversight actions include a lawsuit by the Montana Attorney General against OpenAI for privacy violations in relation to ChatGPT and inquiries by the UK's Information Commissioner's Office into generative AI development and agentic AI systems.

Our text-level analysis of 84 rulebooks across 19 jurisdictions reveals how governments approach data governance in AI in detail. We have identified five regulatory requirements that appear a total of 63 times: general data protection rules on collecting and processing personal data in AI contexts (22 provisions), training data quality standards requiring relevant, representative and error-free datasets (14), data security measures against the loss of or unauthorised access to AI-related data (11), prohibitions of certain data use such as training on sensitive personal data (9), and rights of data subjects, including to access or delete personal information (7).



International Coordination

Our text-level analysis of nine international AI frameworks shows that all cover data governance across 26 provisions. International frameworks address different data governance requirements. General data protection is by far the most commonly mentioned requirement (8), only the OECD AI Principles do not address it. Training data quality (6) and data security (6) are also common. Prohibition of certain data use (3) and data subject rights (2) are rarer. Data subject rights are only mentioned in the UNESCO Recommendation and the ASEAN Guide.

Framework	General data protection	Training data quality	Data security	Prohibition of certain data use	Data subject rights
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	■	–	■	–	–

International AI frameworks are diverse in how they focus on data governance and how actionable their language is. Of the 26 provisions, 13 focus on data governance as the core concern, while 13 only touch upon data governance tangentially. 20 provisions across seven frameworks provide actionable language, and six address data governance in non-actionable language (two in the OECD AI Principles and one each in the Council of Europe Framework Convention, the G7 Code of Conduct, the UNESCO Recommendation and the UN General Assembly resolution).

Alignment Between Domestic and International Action

Domestic regulations and international frameworks are closely aligned on data governance. General data protection measures lead domestically and feature in eight of nine international frameworks. Training data quality and data security, the next most common domestic requirements, each appear in six. The gap lies in prohibitions of certain data uses and data subject rights. Domestic rulebooks contain nine and seven provisions on them, not far behind data security (11), but only three and two international frameworks respectively include them. Finally, the international language behind this overlap is mostly concrete: Most provisions are actionable and about half of the provisions are focused on data governance.

To understand how governments tackle AI data governance in depth, readers can explore our AI Governance Hub's data governance 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.

Human Oversight

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.

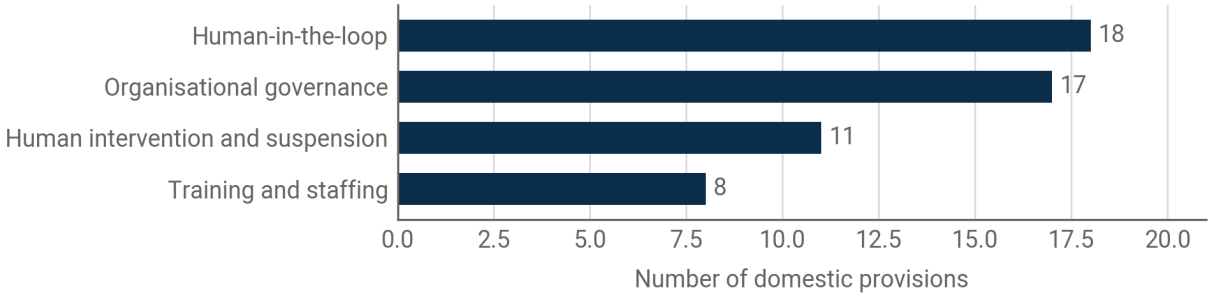
Domestic Regulation

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.

Discrimination and Bias

The risk that AI systems reproduce or amplify existing bias and lead to discrimination is particularly acute where AI informs high-stakes decisions, including areas such as housing, credit, hiring, and healthcare. Examples of discrimination cut across these areas, as evidenced by cases in the US: They include a settlement between the US Justice Department and Meta over claims that its housing advertising algorithm discriminated by race, sex and national origin, a settlement between the Massachusetts Attorney General and a student lender concerning allegations that its AI underwriting disadvantaged Black and Hispanic applicants, a US federal court allowing race, age and disability claims against Workday's AI hiring tools to proceed, and a pharmacy chain agreeing to a five-year ban on using facial recognition after the Federal Trade Commission suggested that it produced false matches more often in stores in Black and Asian neighbourhoods. Governments worldwide are aiming to prevent such cases and ensure redress when they occur.

Domestic Regulation

Discrimination and bias is an emerging concern in AI regulation: The DPA has documented 115 domestic policy and enforcement developments on discrimination and bias in AI since January 2020. The DPA's systematic taxonomy reveals the following trends in AI discrimination and bias regulation:

- The US is the most active jurisdiction, with 32 state-level and 13 federal developments, followed by Argentina (9), Saudi Arabia (7) and the EU (6).
- The current status of developments is spread: 44 developments are in force or adopted, 59 are currently under deliberation and 12 are concluded, rejected or revoked.
- Most discrimination and bias rules are found in binding laws (60) or binding orders (15), compared to 32 non-binding guidelines. Only 8 enforcement cases and inquiries focus on AI discrimination and bias.

Recently, California's Governor signed a bill on AI and health care services requiring both developers and deployers to make reasonable efforts to identify systems at risk of biased impacts when used in healthcare, while Colorado's Governor signed a Bill on automated decision-making in consequential areas that establishes liabilities under anti-discrimination laws for automated decisions. Further, the legislature of the Republic of Korea introduced a bill adding anti-discrimination standards to the AI Basic Act and Thailand's Electronic Transactions Development Agency closed a consultation on a proposed AI Act that would prohibit AI systems whose use results in broad discrimination.

Our text-level analysis of 84 rulebooks across 19 jurisdictions reveals how governments address discrimination and bias in AI in detail. We find that governments use three different regulatory requirements, appearing 50 times, to pursue non-discrimination and address bias in domestic rules, namely non-discrimination obligations requiring AI to be fair and banning discriminatory systems (24 provisions), mitigating bias (13), and detecting bias in datasets and outputs (13).

International Coordination

Our text-level analysis of nine international AI frameworks shows that all cover discrimination and bias across 22 provisions.

International frameworks address all three discrimination and bias requirements. Bias mitigation is the most commonly mentioned requirement (7), followed by non-discrimination (6) and bias detection (6). Five frameworks address all three requirements, while the OECD AI Principles address none.

Framework	Bias mitigation	Non-discrimination	Bias detection
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	■	■	■

International AI frameworks address discrimination and bias with various degrees of focus and actionability. Of the 22 provisions, nine focus on discrimination and bias as the core concern, while 13 only touch upon it tangentially. Also, 16 provisions across seven frameworks provide actionable language, and six address discrimination and bias in non-actionable language (in the Joint Declaration, the OECD AI Principles, the Council of Europe Framework Convention and the G7 Code of Conduct).

Alignment Between Domestic and International Action

Domestic regulations and international frameworks thus weigh discrimination and bias differently. Domestically, non-discrimination dominates, with nearly twice as many provisions as either bias mitigation or bias detection. Internationally, seven frameworks include bias mitigation, and six each cover non-discrimination and bias detection. Finally, the international language behind this overlap is mostly concrete but often indirect. Most provisions are actionable, but fewer than half focus on discrimination and bias.

To understand how governments tackle discrimination and bias in AI in depth, readers can explore our AI Governance Hub's discrimination and bias 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.

Information Integrity

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.

Labour

AI is set to reshape the nature of work, and first signs thereof are appearing In tech firms. Anthropic claimed in November 2026 that over 90% of the changes merged into its core codebase are now written by Claude, with human guidance. Salesforce's chief executive reportedly stated that support staff was cut from about 9,000 to about 5,000 “because I need less heads.” Klarna said that its AI assistant did the work of 700 full-time agents in 2024, although it later reversed course .

Governments are especially concerned about the displacement of roles in their workforce. For instance, in October 2026, China's State Council called on firms to explore an "unemployment buffer period" when deploying AI at scale, with advance notice periods and retraining programmes. At the same time, governments aim to address shifts in the daily reality of work, including worker surveillance and AI use in management and hiring.

Domestic Regulation

Labour is an uncommon concern in AI regulation to date: The DPA has documented 45 domestic policy and enforcement developments on AI and labour since January 2020. The DPA's systematic taxonomy reveals the following trends in AI labour regulation:

- The US is the most active jurisdiction, with 21 state-level and 9 federal developments, followed by the UK (3). Twelve other jurisdictions record 1 each.
- Most developments are currently under deliberation (28), while 11 are in force or adopted and 6 are concluded, rejected or revoked.
- A large majority of labour rules are found in binding laws (33), compared to 4 binding orders and 4 non-binding guidelines. Only 4 enforcement cases and inquiries focus on AI and labour.

Recently, California adopted laws on automated decision systems in employment and notice of technological displacement, while Italy’s implementing decree for the AI Act prohibits final employment decisions from being based solely on automated processing. In China, the Hangzhou Intermediate People's Court ruled in April 2026 that companies may not terminate workers if the only reason is to replace them with AI systems.

Our text-level analysis of 84 rulebooks across 19 jurisdictions shows how governments address labour in AI in detail. We show that governments use two regulatory requirements, appearing 15 times, namely: worker protections against AI-related harms in employment such as discriminatory hiring (9 provisions) and restrictions on worker monitoring and management through AI surveillance, scoring or ranking (6).

International Coordination

Our text-level analysis of nine international AI frameworks shows that six cover labour across seven provisions, all except the G7 Principles, the G7 Code of Conduct and the Council of Europe Framework Convention.

International frameworks primarily focus on worker protections. In this regard, three frameworks, the UNESCO Recommendation, the ASEAN Guide, and the UN General Assembly resolution, mention measures such as workforce impact assessments, job redesign, information for employees and transition support for displaced workers. None of the frameworks address restrictions on using AI to monitor, score or manage workers.

International AI frameworks rarely focus on labour or address it in actionable language. Of the seven provisions, three focus on labour as the core concern (in the UNESCO Recommendation, the ASEAN Guide, and the UN General Assembly resolution), while four only touch upon it tangentially. Three provisions across two frameworks provide actionable language (the ASEAN Guide and the UNESCO Recommendation), and four address labour in non-actionable language, including the UN General Assembly resolution's core provision on labour markets.

Alignment Between Domestic and International Action

To date, labour gets little attention in either domestic rulebooks or international frameworks. Worker protection is the more common domestic requirement and appears in three international frameworks. Worker monitoring and management is found only in domestic rulebooks. In addition, the international language behind this overlap is mostly not concrete. Fewer than half of the provisions are actionable, all of them in the ASEAN Guide and the UNESCO Recommendation, and fewer than half are mainly about labour.

To understand how governments tackle labour and AI in depth, readers can explore our AI Governance Hub's labour 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.

Intellectual Property

In July 2026, a US federal court approved a USD 1.5 billion settlement between Anthropic and authors whose work it had downloaded from piracy websites (Library Genesis and Pirate Library Mirror). The settlement, the largest recorded copyright settlement to date, covered close to 500,000 works, at about USD 3,000 per work.

How intellectual property rights evolve is essential for the business model of AI firms, the creative sector, and AI users. Courts and legislators in many jurisdictions are discussing whether existing intellectual property frameworks are adequate, or whether new rules are needed to address the distinctive features of AI training and AI-driven creation of works. This includes concerns regarding the training of AI models using copyrighted material, the appropriate compensation of creators of AI training material, and ownership rights over AI outputs and prompts.

Domestic Regulation

Intellectual property is not yet a common concern in AI regulation: The DPA has documented 95 domestic policy and enforcement developments on intellectual property and AI since January 2020. The DPA's systematic taxonomy reveals the following trends in AI intellectual property regulation:

- The US is the most active jurisdiction, with 21 federal and 3 state-level developments, followed by the EU (12), the UK (11), and Japan and the Republic of Korea (5 each).
- The current status of developments is spread: 33 developments are in force or adopted, 47 are currently under deliberation and 15 are concluded, rejected, revoked or under investigation.
- Intellectual property is the only concern where enforcement is the most frequent legal form: 41 enforcement cases and inquiries, compared to 29 binding laws, 15 non-binding guidelines and 10 binding orders.

For example, in October 2026, Italy's legislative decree implementing the EU AI Act entered into force, including trade secret protection for data, algorithms, and mathematical methods used to train AI. Further, a bill creating a rebuttable presumption that AI systems have used cultural works protected by copyright was introduced to the Belgian parliament. Beyond legislation, Japan adopted its principle code on intellectual property protection in generative AI in August 2026, which asks generative AI businesses to publish training data sources and to disclose whether certain works were used to rights holders on request. Courts are also moving: on the same day in November 2026, a São Paulo court barred Meta from scraping Brazilian press associations' members' content to train its AI and the Court of Justice of the EU's Advocate General issued an opinion that chatbots reusing press articles in real time may infringe publishers' rights.

Our text-level analysis of 84 rulebooks across 19 jurisdictions reveals that governments address intellectual property using provisions on copyright at the domestic level (10 provisions). These provisions are found in enacted AI rulebooks in China, El Salvador, the EU, Japan, and the US state Arkansas, as well as legislative proposals in Brazil and Mexico.

International Coordination

Our text-level analysis of nine international AI frameworks shows that eight cover intellectual property across 12 provisions, all except the Council of Europe Framework Convention.

International frameworks address different regulatory requirements regarding intellectual property. Five frameworks, the G7 Code of Conduct, the UNESCO Recommendation, the ASEAN Expanded Guide, the Joint Declaration, and the UN General Assembly resolution, refer to copyright, covering the use of copyright-protected content in AI training and the ownership of AI-generated works. Only a single framework, the Joint Declaration, refers to remunerating rights-holders whose content is used for training. None of the frameworks address patents.

Provisions in international frameworks rarely focus on intellectual property or address it in actionable language. Of the 12 provisions, a single provision in the UN General Assembly resolution focuses on intellectual property as the core concern, while 11 only touch upon it tangentially. Only two provisions provide actionable language (in the ASEAN Expanded Guide and the Joint Declaration), and ten address intellectual property in non-actionable language.

Alignment Between Domestic and International Action

On intellectual property, domestic regulations and international frameworks both focus on copyright. It is the only requirement domestic rulebooks use and the most frequent international reference, appearing in five frameworks. One requirement exists only at the international level: the Joint Declaration calls for compensating rights-holders whose content is used for AI training. Patents appear at neither level. Finally, the international language behind this overlap is rarely concrete. Only two provisions are actionable, and only one is mainly about intellectual property.

To understand how governments tackle intellectual property and AI in depth, readers can explore our AI Governance Hub's intellectual property 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.

Accountability

Accountability, in our taxonomy, does not encompass specific regulatory requirements, but rather sets out the responsibility for compliance, the consequences of non-compliance, and the oversight mechanisms to ensure compliance.

- The responsibility for compliance is a core element of accountability, since it describes who governments target with regulatory requirements. Targets are typically determined based on AI systems' risk level, the AI systems' technological specificities, and AI firms' role in the AI value chain.
- The consequences of non-compliance with a regulatory requirement are a strong indicator of how a government creates accountability. Governments are currently experimenting with a spectrum of provision types, from high-level principles, to non-binding or voluntary commitments, to binding requirements that carry a sanction threat.
- Oversight mechanisms determine how governments create accountability in practice: Which existing or new government authorities are in charge of overseeing AI firms and what powers do these authorities have to enforce regulatory requirements?

Domestic Regulation

Accountability is a common concern in domestic AI regulation: The DPA has documented 309 domestic policy and enforcement developments on accountability in AI since January 2020. The DPA's systematic taxonomy reveals the following trends in AI accountability regulation:

- The US is the most active jurisdiction, with 40 federal and 23 state-level developments, followed by the EU (27), China (16) and the Republic of Korea (14).
- Adopted or in-force developments narrowly outnumber pending ones: 153 are in force or adopted, 143 are currently under deliberation and 13 are concluded or rejected.
- Binding laws lead (133), ahead of non-binding guidelines (81) and binding orders (74). Only 21 enforcement cases and inquiries focus on AI accountability.

Recent developments focus mainly on oversight. Governments worldwide are establishing AI-specific government agencies, such as Australia's announced AI office, AI regulator, and AI safety institute, or other oversight structures, such as Japan's AI and Digital Reform Promotion Council. Governments also establish sanctions regimes, including the entry into force of the enforcement and penalty provisions of Poland's Act on AI Systems and the introduction of a Finnish bill amending the existing enforcement rules on AI.

Our text-level analysis of 84 rulebooks across 19 jurisdictions shows 67 provisions in which governments provide for investigative powers and sanctions with regard to AI systems. In these provisions, governments set out the legal framework for government oversight of AI, either by creating new regulatory bodies or by expanding existing powers, as well as legal consequences for not observing regulatory requirements.[4]

[4] We do not provide detailed provisions on the allocation of responsibility on the basis of risk level, AI capabilities, and the role of AI actors within the value chain, despite this element forming an important part of accountability. We aim to map out these approaches to assigning accountability in our future analysis on AI governance. Finally, at this point we do not focus on measures that relate to the general governance of the AI ecosystem insofar as they are not directly tied to regulatory requirements for private-sector AI actors.

International Coordination

Our text-level analysis of nine international AI frameworks shows that six frameworks cover accountability across 17 provisions. The international frameworks that address accountability are the OECD AI Principles, the UNESCO Recommendation, the Council of Europe Framework Convention, the ASEAN Guides, and the Joint Declaration. The UN General Assembly resolution, the G7 Principles and the G7 Code of Conduct do not cover accountability, as we have conceptualised it.

The frameworks cover a range of elements of accountability, with all frameworks referring to the allocation of responsibility among AI actors, or designating a responsible person within organisations. The UNESCO Recommendation and the Joint Declaration also refer to the establishment of institutions for the oversight and enforcement of AI rules. Three frameworks, the Joint Declaration, the UNESCO Recommendation, and the Council of Europe Framework Convention, address the need to provide remedies or redress.

International AI frameworks, however, differ in how they focus on accountability and how actionable their language is. Of the 17 provisions, 10 focus on accountability as the core concern, while seven only touch upon it tangentially. 12 provisions across six frameworks provide actionable language, and five address accountability in non-actionable language.

Alignment Between Domestic and International Action

Given the narrow scope of our current analysis on accountability, we do not draw definitive conclusions on the alignment between domestic regulations and international frameworks. In summary, the 67 domestic provisions designate which governmental bodies are responsible for oversight and set out the applicable sanctions for violations of regulatory requirements. While three international frameworks stress the need for creating effective remedies or redress, only two international frameworks refer to oversight institutions and sanctions, possibly since such questions of concrete implementation are generally left to domestic regulators. International frameworks do address the issue of how to allocate responsibility among AI actors. As stated above, domestic frameworks adopt varying approaches for classifying and allocating responsibility to AI actors and systems, which will be a focus of our upcoming work on AI governance.

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

The Language of International AI Frameworks

This section analyses the provisions of international AI frameworks addressing each concern, including an overview of their focus and actionability, and a summary of relevant passages:

- Transparency
- Safety
- Consumer Protection
- Data Governance
- Human Oversight
- Discrimination and Bias
- Information Integrity
- Labour
- Intellectual Property
- Accountability

Transparency

Our text-level analysis of nine international AI frameworks shows that all nine cover transparency, across a total of 47 provisions with various degrees of focus and actionability.

- 23 provisions focus on transparency as the core concern while 24 provisions only touch upon transparency tangentially.
- 42 provisions provide actionable language for governments to pursue transparency. Only five provisions, in the G7 documents, the Joint Declaration, and the UN General Assembly resolution, address transparency in non-actionable language.
- The ASEAN Guide, G7 Code of Conduct, and G7 Principles have the most provisions on transparency (8 each). The ASEAN Guide also has the most actionable provisions (8), while the UN General Assembly resolution has the most core provisions (4).

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 transparency. Principle 1.3 (Transparency and Explainability) covers it as core concern and Principle 1.5 (Accountability) covers it tangentially. Both principles provide actionable measures for governments.

Principle 1.3 focuses on transparency and specifies four ways in which AI actors should provide information:

- general understanding of capabilities and limitations;
- awareness of interactions with AI systems, including in the workplace;
- plain-language information on the data, factors and logic behind an output; and
- information enabling adversely affected persons to challenge an output.

Principle 1.5 focuses on accountability but includes actionable language on transparency. Specifically, the Principle states that AI actors should ensure the traceability of datasets, processes and decisions across the lifecycle, which in turn enables the analysis of outputs and "responses to inquiry".

UN General Assembly resolution

Six provisions in the UN General Assembly resolution A/RES/78/265 address transparency. Four provisions are core and two are tangential. All except one are actionable. Out of the core provisions:

- 6(d) encourages raising public awareness of AI systems' core functions, capabilities, limitations and domains of appropriate civil use;
- 6(g) encourages the development of content authentication and provenance mechanisms;
- 6(k) encourages the provision of notice and explanation for systems that make or support decisions affecting end-users;
- 6(n) encourages information-sharing among entities across the AI lifecycle.

Out of the tangential provisions:

- 6(e) encourages the disclosure of risk monitoring and management mechanisms, privacy policies and impact assessments;
- Sub-paragraph 6(j) mentions transparency and reporting requirements only as a context in which privacy must be safeguarded, and is therefore non-actionable.

All measures are framed as encouragement to Member States and, where applicable, other stakeholders. Sub-paragraphs 6(d) and 6(g) closely mirror the wording of G7 Principles 3 and 7.

Council of Europe Framework Convention

Four articles in the Council of Europe Framework Convention are related to transparency. The concern is core to one of them and tangential to three of them. All are actionable. The Convention is the only instrument examined that contains "shall" obligations. These bind Parties rather than AI actors directly, and several are softened by qualifiers.

Article 8, on transparency and oversight, is the core provision requiring parties to adopt transparency measures. It leaves the substance of transparency requirements to Parties, requiring only that they be "tailored to the specific contexts and risks". Its one specified element is the mention of the identification of AI-generated content as a relevant activity in the AI lifecycle.

Three provisions refer to transparency tangentially:

- Article 14(2) requires documentation of information on systems that could significantly affect human rights, access for authorised bodies and communication to affected persons. The provided information must allow affected persons to contest the decision and, where relevant, the use of the system itself.
- Article 15(2) requires notification that a person is interacting with AI rather than a human. This is the Convention's weakest transparency obligation, since Parties need only "seek to ensure" it "as appropriate for the context".
- Articles 16(2)(f) and 16(3) prescribe the documentation of risks and adverse impacts, including on human rights.

G7 Hiroshima Process Guiding Principles

Eight Guiding Principles feature transparency. Transparency is a core concern for three and tangential to five others. Six Principles include an actionable element with regard to transparency, while two are non-actionable. The Principles are voluntary and addressed to organisations developing advanced AI systems.

Transparency is core to three principles:

- Principle 3, on the public reporting of system capabilities, suggests transparency reports for "all new significant releases", a clarity standard (enabling deployers and users to interpret output) and underlying documentation. While the G7 Principles do not specify contents for the report, the Code of Conduct does.

- Principle 4, on responsible information sharing and incident reporting, suggests sharing evaluation reports, safety and security risks, dangerous capabilities and attempts to circumvent safeguards.
- Principle 7 suggests the development of content authentication and provenance mechanisms.

Transparency is tangentially related to five further principles:

- Principle 2, on vulnerability identification and mitigation, encourages the documentation of reported incidents.
- Principle 5, on governance and risk management policies, encourages the disclosure of such policies, including privacy policies.
- Principle 10, on the development of technical standards, explicitly mentions the need to develop standards for watermarking.
- Principle 1, on appropriate measures to address risks, is non-actionable, simply mentioning that developers should aim to enable traceability of datasets, processes, and decisions.
- Principle 11, on data input and protection measures, is also non-actionable, stating that appropriate transparency of training datasets should be supported.

G7 Hiroshima Process Code of Conduct

The Code of Conduct builds on the G7 Principles, adding additional detail in several cases. Transparency appears in the same eight provisions, as the core concern of three provisions and tangential to five. Seven of eight provisions contain an actionable element on transparency, with Provision 1 receiving an actionable element that was not present in the Principles. Apart from this, the added value of the Code of Conduct lies in elaboration:

- Provision 3 specifies what transparency reports should contain: evaluations of safety, security, societal and human rights risks; capabilities and significant limitations; effects on safety and society, including bias, discrimination and privacy; and red-teaming results. It adds instructions for use and technical documentation, all to be kept up to date.
- Provision 4 extends information sharing to the public and to public authorities. It adds lifecycle documentation for systems posing significant risks and mechanisms for shared standards. It also provides that reporting "should safeguard intellectual property rights".
- Provision 2 adds bounty systems, contests and prizes as incentives for responsible disclosure.
- Provision 10 extends standards work to content authentication, provenance and public reporting, and to interoperable standards for distinguishing AI-generated content.
- Provision 1, on risk identification and mitigation, adds an actionable transparency element by requiring regularly updated technical documentation to support risk testing.
- Provisions 5, 7, and 11 substantially reproduce the transparency elements of the Principles. Provision 7 adds a commitment to research on provenance.

ASEAN Guide on AI Governance and Ethics

Eight provisions across the ASEAN Guide focus on transparency. Three provisions have transparency as a core concern, and five provisions are tangentially related to transparency. All eight

provisions contain an actionable transparency element. The Guide is addressed to deployers and developers. Its distinguishing feature is practical specificity, with worked examples, named tools and fallback options where full transparency is not feasible.

Three provisions take transparency as their core concern:

- Principle 1 focuses on Transparency and Explainability. It sets out several actionable measures. These include disclosure of AI use, of its role in decisions, of the data used and of its purpose, and explanations ranging from simple text to heatmaps. Where models cannot be explained, it offers alternatives: outcome-based explanations, documentation of repeatability, audit trails, a "black box recorder" capturing input data, records of data provenance, and AI Model Cards. The black box recorder is the only explicit logging-type measure in the dataset.
- Section 4 centres on Stakeholder interaction and communication. It suggests chatbot disclosure requirements and disclaimers that user inputs may be used for training.
- The section on 'Modelling' guidance has a strong focus on explainability. Its actionable measures include:
    - documentation of training, model selection and automated machine-learning parameters;
    - expected-behaviour descriptions in product documentation;
    - contractual transparency obligations on developers and joint audits;
    - Use of explainability tools such as AI Verify, SHAP and LIME;
    - implicit and counterfactual explanations for lay users.

The section also flags fraud detection and proprietary trading algorithms as cases where disclosure may be inappropriate.

In the five tangential provisions, transparency serves other principles:

- With regard to security and safety, risks, limitations and safeguards should be made known to users.
- Internal documentation of errors can ensure accountability and integrity.
- Robustness and reliability can be improved through documentation of data sources and lineage.
- Determining the level of human involvement in AI-augmented decisions is facilitated by recording system and user behaviour for auditing purposes.
- Proper outcome analysis can be conducted if an organisation has documented the intended purpose of an AI system and established communication channels between technical teams and stakeholders.

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. Four provisions feature transparency: it is core to three provisions and tangential to one further provision. All transparency elements are actionable.

Three provisions take transparency as their core concern:

- Section 3.4 contains the most detailed treatment of incident reporting in the dataset. It covers internal channels, external reporting to regulators, users, customers and value chain participants, vulnerability reporting, severity tiering, and interaction with existing breach notification regimes.
- Section 3.7 covers labelling, watermarking and cryptographic provenance.
- Section 1.3 restates all seven principles. Its transparency component is the Transparency and Explainability principle plus one element under privacy and data governance. The definitions reproduce the 2024 Guide. The new actionable elements are a mention of user awareness of interacting with generative AI and of how data is used, and transparency about training data sources and pre-training processing. Otherwise the section is largely diagnostic: it openly acknowledges that explanation and traceability have inherent limits for generative AI.

One provision is tangential:

- Section 3.3 combines safety and evaluation guidance with a transparency component, namely the disclosure of safety practices to downstream deployers and end users, likened to "food or ingredient labels", and proposed common disclosure elements for developers.

UNESCO Recommendation on the Ethics of AI

Three provisions in the UNESCO Recommendation feature transparency. All three are core and actionable.

Transparency is the core concern of the principle 'Transparency and explainability'. It is largely conceptual and offers the most developed rights-based rationale in the dataset, linking transparency to liability regimes, fair trial and effective remedy. It nevertheless contains concrete measures:

- informing people when decisions are AI-based and when products or services use AI;
- a right to request explanatory information and to access the reasons for decisions;
- submissions to a designated staff member able to review and correct decisions;
- meaningful explanations for decisions whose impact is not temporary, easily reversible or low risk.

Its most far-reaching suggested measure, not mentioned in any other framework, is the sharing of code or datasets where there are serious threats of adverse human rights impacts.

Transparency is also core to Policy Area 9, on communication and information. The Policy Area's transparency component refers to the transparency of online communication and information operators and prompt notification of the reasons for content removal or other treatment. It sits alongside material on access to information, literacy and media policy.

Policy Area 2 is a broad governance section. Its transparency component includes clear transparency and explainability requirements (paragraph 70), mandatory disclosure of stereotyping in AI outcomes and data (paragraph 66), and international standards setting "measurable, testable levels" of transparency (paragraph 64). It also covers auditability and traceability for redress (paragraph 55) and transparent self-assessment of AI systems by public authorities (paragraph 57).

Joint Declaration on AI, Freedom of Expression and Media Freedom

Four provisions of the Joint Declaration feature transparency: it appears in two as a core concern and in two as a tangential concern. Three provisions include actionable transparency elements. The Declaration addresses states, companies, as well as financial institutions and investors.

Transparency is a core concern of Principle 6, on transparency, accountability, and effective remedies. Its actionable component is its call for "meaningful information and access to data".

Further, transparency is also core to the section addressed to the private sector, which calls on AI companies to publish due diligence assessments ((a)); implement industry standards for the attribution and labelling of synthetic content ((e)); and provide meaningful information on system functioning, traceability, explainability, capabilities and limitations, and affected persons' rights ((f)), as well as, specifically, ensuring the explainability of AI in content moderation, curation and targeted advertising, paired with an appeal avenue ((h)).

Transparency is tangential to the section on state regulation of companies, which calls for transparency requirements aimed at bias correction, harm identification, liability and remedies ((a)(iii)); public availability of oversight and testing findings as "epistemic counterpower" ((b)(i)); and data protection laws enabling transparency in data collection and use ((c)).

Finally, transparency is also tangential to Principle 2, on a healthy, pluralistic, and diverse information environment. It only names transparency as a goal of regulation, without specifying any actionable measure.

Safety

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

- 29 provisions focus on safety as the core concern, while 9 provisions only touch upon safety tangentially.
- 35 provisions across all nine frameworks provide actionable measures to pursue safety. Only three provisions address safety in non-actionable language: one each in the Council of Europe Framework Convention, the UNESCO Recommendation and the UN General Assembly resolution.
- The ASEAN Guide has the most provisions on safety (7), all of which are also actionable. The ASEAN Guide, along with the G7 Principles and Code of Conduct, also has the most core provisions (5 each).

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

OECD AI Principles

Three OECD AI Principles feature safety. Principle 1.4 (Robustness, security and safety) covers it as core concern, while Principles 1.2 (Respect for the rule of law, human rights and democratic values, including fairness and privacy) and 1.5 (Accountability) cover it tangentially. All three Principles contain actionable elements with regard to safety.

Principle 1.4 states that AI systems should be robust, secure and safe throughout their lifecycle, including under foreseeable misuse ((a)). Its actionable element is the call for mechanisms ensuring that systems which risk causing undue harm or exhibit undesired behaviour can be "overridden, repaired, and/or decommissioned safely" ((b)).

In the two tangential principles, safety serves other objectives:

- Principle 1.2, on human rights and democratic values, calls for safeguards addressing risks from uses outside the intended purpose and from intentional or unintentional misuse ((b)).
- Principle 1.5, on accountability, calls on AI actors to apply a "systematic risk management approach" to each phase of the lifecycle ((c)).

UN General Assembly resolution

Six provisions in the UN General Assembly resolution A/RES/78/265 feature safety, which is core to four of them and tangentially related to two. All except one are actionable. The measures are framed as encouragement to Member States and, where applicable, other stakeholders. Out of the core provisions:

- 6(b) encourages measures for the internationally interoperable identification, classification, evaluation, testing, prevention and mitigation of vulnerabilities and risks before deployment;

- 6(c) encourages feedback mechanisms through which end-users and third parties can discover and report technical vulnerabilities, misuses and incidents after deployment, so that they can be addressed;
- 6(f) encourages investment in safeguards including physical security, AI system security and lifecycle risk management.
- 6(e) focuses on the development of risk monitoring and management mechanisms and impact assessments. It also mentions measures touching other concerns, such as data security, privacy policies, impact assessments and their disclosure.

Out of the tangential provisions:

- 6(l) names risk and impact assessments as safeguards for the protection of human rights;
- 6(a) names safe, secure and trustworthy AI only as an objective of domestic regulatory frameworks, and is therefore non-actionable.

Council of Europe Framework Convention

Three articles in the Convention feature safety, which is core to all of them. Two of the three are actionable. The Convention is the only instrument examined that contains “shall” obligations. Convention articles bind state parties rather than AI actors directly, and several are softened by qualifiers.

Article 16, on the risk and impact management framework, focuses on risks to human rights, democracy and the rule of law, while also containing the Convention’s most detailed safety measures. It requires measures for the identification, assessment, prevention and mitigation of risks (16(1)), graduated by the severity and probability of impacts (16(2)(b)) and applied iteratively across the lifecycle (16(2)(d)). These measures include monitoring (16(2)(e)) and, where appropriate, testing before first use and after significant modification (16(2)(g)). Parties must also address adverse impacts (16(3)) and assess the need for a moratorium, ban or other measures for uses incompatible with human rights, democracy or the rule of law (16(4)). It is the only provision in the dataset to mention a moratorium.

Article 13, on safe innovation, calls upon Parties to enable controlled environments for developing, experimenting and testing AI systems under the supervision of competent authorities.

Article 12, on reliability, requires Parties to take “as appropriate” measures to promote reliability and trust in outputs. It names requirements on “adequate quality and security” only as a possible measure, and is therefore non-actionable.

G7 Hiroshima Process Guiding Principles

Five principles are about safety, which is a core concern of all of them. Further, all five principles are actionable with regard to safety. The Principles are voluntary and addressed to organisations developing advanced AI systems:

- Principle 1, on identifying, evaluating and mitigating risks across the lifecycle, suggests diverse internal and independent external testing, including red-teaming, and mitigation of identified risks and vulnerabilities.

- Principle 2, on vulnerabilities, incidents and misuse after deployment, suggests monitoring for vulnerabilities, incidents, emerging risks and misuse, taking action to address them, and facilitating third-party and user discovery of issues.
- Principle 6, on security controls, suggests physical security, cybersecurity and insider threat safeguards, including securing model weights, algorithms and servers through operational security and access controls. The G7 documents are the only frameworks in the dataset to address the security of model weights.
- Principle 5, on governance and risk management policies, suggests accountability and governance processes to evaluate and mitigate risks throughout the lifecycle.
- Principle 8, on research and investment in mitigation, recommends research and investment into mitigation tools.

G7 Hiroshima Process Code of Conduct

The Code of Conduct builds on the G7 Principles, adding detail in several cases. Six provisions of the Code feature safety, as the core concern of five and tangential to one. All provisions are actionable with regard to safety. For the five core provisions the Code shares with the Principles, the Code's relevance and actionability tags are identical. Its added value therefore lies in elaboration:

- Provision 1 adds that testing should take place in secure environments at several checkpoints, in particular before deployment and placement on the market. It lists risks to which testing should attend: chemical, biological, radiological and nuclear risks; offensive cyber capabilities; risks to health and safety, including control of physical systems and interference with critical infrastructure; self-replication; societal risks; threats to democratic values; and chain reactions affecting up to an entire city or community. It is the only provision in the dataset to name chemical, biological, radiological and nuclear risks or self-replication.
- Provision 2 adds bounty systems, contests and prizes to incentivise the responsible disclosure of weaknesses.
- Provision 5 includes additional details on the risk management framework that should be implemented across the AI lifecycle.
- Provision 6 adds a cybersecurity risk assessment, cybersecurity policies, secure environments with limited access for model weights, a vulnerability management process with regular review, and an insider threat detection programme.
- Provision 8 adds a commitment to proactively manage risks of advanced AI systems.

One provision contains a safety element with no tagged counterpart in the Principles. Provision 10, on international technical standards, is tangential and actionable: it encourages contributions to standards that organisations can use when developing their testing methodologies and cybersecurity policies.

ASEAN Guide on AI Governance and Ethics

Seven provisions in the ASEAN Guide are relevant to safety. Safety is the core concern of five provisions and tangential to two. All seven provisions contain an actionable safety element. The Guide is addressed to deployers and developers. It follows the AI lifecycle, from project definition through modelling and outcome analysis to deployment and monitoring.

Five provisions take safety as their core concern:

- Principle 3, on security and safety, suggests risk assessments and relevant testing or certification before deployment, and mechanisms for systems to disengage safely. On security, it names AI-specific attacks, such as data poisoning, model inversion, dataset tampering and byzantine attacks in federated learning. It suggests authentication, encryption, software updates, access management and incident response plans, a minimum list of security tests such as penetration testing, and business continuity and disaster recovery plans. The two ASEAN guides are the only frameworks in the dataset to name specific attacks on AI systems.
- Principle 7, on robustness and reliability, suggests resilience to unexpected inputs and stressful conditions, access control for critical systems, mitigation of unreliable performance and rigorous testing before deployment.
- Section 2, on determining the level of human involvement, suggests a risk management approach at every lifecycle stage. Risks are assessed by the probability and severity of harm, and risk management and mitigation plans are continually reviewed.
- The guidance on deployment and monitoring suggests a secure and resilient architecture and performance monitoring with defined metrics and thresholds. It also suggests revalidation and drift monitoring, testing whether systems generalise well and "fail gracefully", and regular model tuning followed by renewed testing.
- The guidance on modelling suggests adversarial testing before deployment and close monitoring of continuously learning systems for unintended behaviour. Its robustness component sits alongside material on explainability and auditability.

The guidance on project governance and problem statement definition is tangential to safety. In service of project governance, it suggests risk-based assessments before any data collection or modelling, covering foreseeable accidental or malicious misuse. Its mitigation measures include regular monitoring and maintenance, procedures for previously unknown risks and safe disengagement mechanisms. It also suggests action to resolve non-compliance by third-party developers or vendors.

The guidance on outcome analysis is also tangential to safety. It suggests confirming that pre-deployment purposes are actually fulfilled by a system and conducting acceptance tests that may include security and performance evaluations.

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. Four provisions in the Expanded Guide are about safety, and it is the core concern of all four. All four are actionable:

- Section 1.3 restates Security and Safety as a principle and mentions safety elements specific to generative AI. These include fine-tuning techniques such as reinforcement learning from human feedback, red-teaming to expose unsafe behaviours, and protection against prompt injection and data poisoning. They also include dynamic testing for unforeseen behaviours, measurable benchmarks for acceptable outcomes, and guardrails

against out-of-context prompts. The section acknowledges that reliable common practices for evaluating generative AI safety still require further research.
- Section 3.3, on trusted development and deployment, suggests integrating safety best practices across the generative AI lifecycle, naming practices and guardrails, such as adversarial testing, and continuous system monitoring.
- Section 3.5, on testing and assurance, suggests third-party testing and, eventually, formal auditing to complement internal evaluation. As areas for ASEAN to explore, it proposes regional benchmarks and testing tools reflecting ASEAN's languages and cultures, and an agreed list of preferred evaluation tools. It illustrates automated benchmarking and red-teaming with an open-source toolkit.
- Section 3.6, on security, addresses generative AI-specific threats such as adversarial machine learning. It proposes vulnerability detection programmes such as bug bounties and ethical hacking, security knowledge-sharing, and security-by-design guidelines aligned with existing international standards. Bug bounties also appear in G7 Code Provision 2.

UNESCO Recommendation on the Ethics of AI

Three provisions in the UNESCO Recommendation feature safety: two core and one tangential. Two of the three are actionable.

Safety is the core concern of the principle 'Safety and security' (paragraph 27). It is non-actionable: harms and vulnerabilities are to be avoided, addressed, prevented and eliminated, though no safety-specific mechanism to do so is named.

Safety is also core to Policy Area 2, a broad governance section addressed to Member States. Its safety component includes:

- certification mechanisms with different levels of audit and a regular monitoring component, requiring re-certification if necessary (paragraph 56);
- international standards describing "measurable, testable levels" of safety, and research on safety and security risks (paragraph 64);
- self-assessment by public authorities, with a prohibition of use where adoption would violate human rights obligations (paragraph 57).

Safety is tangential to Policy Area 1, on ethical impact assessment. In service of that assessment, it calls for frameworks to identify risks and appropriate prevention, mitigation and monitoring measures (paragraph 50). It also calls for broad testing, including in real-world conditions, of systems identified as potential risks to human rights before market release (paragraph 51), monitoring of all lifecycle phases (paragraph 52), and external review of AI systems (paragraph 53).

Joint Declaration on AI, Freedom of Expression and Media Freedom

One provision, on journalists and media actors, is tangentially about safety, as well as being actionable. In service of responsible journalism, it calls on journalists and media actors to perform a "systematic risk assessment" as a precondition for deploying AI systems for journalistic purposes ((b)).

Consumer Protection

Our text-level analysis of nine international AI frameworks shows that seven frameworks cover consumer protection, all except the G7 Principles and the ASEAN Expanded Guide on Generative AI. A total of 11 provisions address consumer protection across seven frameworks, with various degrees of focus and actionability.

- Six provisions focus on consumer protection as the core concern, while five provisions only touch upon consumer protection tangentially.
- 10 provisions across all seven frameworks provide actionable measures to pursue consumer protection. Only one provision, in the Council of Europe Framework Convention, addresses consumer protection in non-actionable language.
- The frameworks with the most provisions on consumer protection are the ASEAN Guide, Council of Europe Framework Convention, Joint Declaration, and the UNESCO Recommendation, which all have two provisions each. Additionally, all consumer protection provisions in the ASEAN Guide, Joint Declaration, and UNESCO Recommendation are also actionable. Further, both of the ASEAN Guide's provisions are core.

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 consumer protection. Principle 1.3 (Transparency and explainability) covers it tangentially and is actionable.

Principle 1.3's consumer protection element is the focus on enabling those adversely affected by an AI system to challenge an output ((iv)).

UN General Assembly resolution

One provision in the UN General Assembly resolution A/RES/78/265, sub-paragraph 6(l), is of core relevance to consumer protection. The provision is actionable.

Sub-paragraph 6(l) encourages investment in developing and implementing safeguards such as risk and impact assessments to mitigate the impact on human rights and fundamental freedoms.

Council of Europe Framework Convention

Two Convention Articles, one core and one tangential, address consumer protection. Article 16 is actionable, while Article 18 is non-actionable.

Article 16, on risk and impact management frameworks, is the core provision. It requires the adoption of risk assessment and management frameworks, which should evaluate risks posed by AI systems with regard to impacts on human rights, democracy, and the rule of law.

Article 18, on the rights of persons with disabilities and of children, is tangentially relevant to consumer protection. It requires Parties to take "due account" of specific needs and vulnerabilities in

relation to the rights of such persons. However, it does not specify any protective measures and is therefore non-actionable.

G7 Hiroshima Process Code of Conduct

One provision in the Code of Conduct, Provision 8, relates to consumer protection, which is tangential and actionable.

Provision 8, on prioritising research and investment to mitigate societal, safety and security risks, focuses on safety research. Its consumer protection element is the commitment to prioritise research on key risks including "protecting children and vulnerable groups".

ASEAN Guide on AI Governance and Ethics

Two provisions in the ASEAN Guide address consumer protection, both of which have core relevance and are actionable. The Guide is addressed to deployers and developers and offers worked examples and decision factors.

- Principle 4, on human-centricity, is a core provision combining material on labour and job redesign with a consumer protection component. It suggests that AI systems should not be used to "sway or deceive users" into decisions that do not benefit them. It also suggests that dark patterns be avoided and that systems not take advantage of vulnerable individuals. It is the only provision in the dataset to address dark patterns.
- Section 4, on stakeholder interaction and communication, combines disclosure, employee communication, consent to training-data use and supplier relationships with a consumer protection component. It suggests feedback channels and mechanisms for aggrieved individuals, which can also allow users to correct inaccurate personal data. It further suggests considering whether users should be able to opt out of the AI-enabled service, and obtaining feedback from affected users where they cannot. It is the only provision in the dataset to list factors for deciding on an opt-out: the degree of risk, reversibility, alternative decision-making mechanisms, the complexity of maintaining parallel systems, and technical feasibility.

UNESCO Recommendation on the Ethics of AI

Two provisions relate to consumer protection, one of which is of core and another of tangential relevance. Both are actionable. The policy area is addressed to Member States.

The core provision is Policy Area 11, on health and social well-being. It combines health care policy, privacy and consent, and human decision-making in diagnosis and treatment with a consumer protection component. It covers four elements of consumer protection, more than any other provision in the dataset:

- Users should be able to "effectively refuse" interaction with an AI system imitating human or animal characteristics and request human intervention (paragraph 127).
- Special attention should be paid to the possibility of AI being used to manipulate and abuse human cognitive biases (paragraph 125).

- Member States should establish research on harms to mental health such as depression, social isolation and addiction, and on their regulation (paragraph 124) and develop guidelines for human-robot interaction, including chatbots and companion robots for children and adults (paragraph 125).
- Due attention should be paid to children and young people in four areas: the regulation of health and mental health AI (paragraph 122), awareness of anthropomorphisation (paragraph 128), research on psychological and cognitive impacts (paragraph 129), and mechanisms to engage them in decision-making on AI (paragraph 130).

The tangential provision, Policy Area 2, on ethical governance and stewardship, calls on Member States to carry out self-assessments of AI systems with regard to whether adoption would result in violations of human rights obligations (paragraph 57). Further, it calls on Member States to encourage public entities, private sector companies and civil society organisations to add mechanisms, such as an independent AI Ethics Officer, to oversee ethical impact assessment (paragraph 58).

Joint Declaration on AI, Freedom of Expression and Media Freedom

Two provisions address consumer protection. It has core relevance in one provision and tangential relevance in another. Both provisions are actionable. The Declaration addresses states, companies, as well as financial institutions and investors. Its consumer protection element in both provisions concerns human rights due diligence, and it is the only framework in the dataset to address this element.

The core section addressed to the private sector calls on companies to conduct human rights due diligence at every layer of the AI lifecycle, to make it publicly available and to subject it to external scrutiny ((a)). It extends due diligence to financial institutions and investors funding AI ((a) of the list addressed to them), the only framework examined to do so. The section's other material is unrelated, including staff diversity and training, data protection, content governance appeals and relations with media organisations.

The section on state regulation of companies, which is tangential, calls on states to introduce regulations requiring human rights due diligence at every stage of the AI lifecycle, subject to external, independent scrutiny ((a)(i)). The due diligence should include a specific assessment for marginalised persons or groups more susceptible to negative mental health impacts, opinion manipulation, and online violence and harassment. The section's other material on transparency, oversight institutions, data protection, competition and linguistic diversity is unrelated to consumer protection.

Data Governance

Our text-level analysis of nine international AI frameworks shows that data governance is a universal concern across all nine frameworks. A total of 26 provisions address data governance, with various degrees of focus and actionability.

- 13 provisions focus on data governance as the core concern, while 13 provisions only touch upon data governance tangentially.
- 20 provisions across seven frameworks provide actionable measures to pursue data governance. Only 6 provisions address data governance in non-actionable language: two in the OECD AI Principles and one each in the Council of Europe Framework Convention, the G7 Code of Conduct, the UNESCO Recommendation and the UN General Assembly resolution. In the OECD Principles and the Council of Europe Framework Convention, data governance appears only in non-actionable language.
- The ASEAN Guide, UNESCO Recommendation, G7 Code of Conduct, and Joint Declaration have the most provisions on data governance, with four provisions each. In both the ASEAN Guide and the Joint Declaration, all provisions are also actionable. The frameworks with the most core provisions are the ASEAN Guide, UNESCO Recommendation, Joint Declaration, ASEAN Expanded Guide, and the UN General Assembly resolution (2 each).

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 data governance. Principle 1.2 (Respect for the rule of law, human rights and democratic values, including fairness and privacy) and Principle 1.5 (Accountability) both cover it tangentially, and neither is actionable.

- Principle 1.2 names privacy and data protection only as one of many values AI actors should respect (a).
- Principle 1.5 names privacy only as one category of risk to manage (c). Section (b) also mentions the importance of traceability with regard to datasets, though we consider this primarily a transparency measure.

UN General Assembly resolution

Two provisions in the UN General Assembly resolution A/RES/78/265 feature data governance as a core concern. One of the two is actionable. Both are framed as encouragement to Member States and, where applicable, other stakeholders.

Sub-paragraph 6(e) is actionable. Its data governance component is the development and implementation of mechanisms for securing data, including personal data protection and privacy policies. These sit alongside risk monitoring and management mechanisms and impact assessments, which are covered in the safety section.

Sub-paragraph 6(j) has data governance as its core concern but is non-actionable. It calls for safeguarding privacy and the protection of personal data when testing and evaluating systems and in transparency and reporting requirements, without identifying any specific data governance measure.

Council of Europe Framework Convention

A single article of the Convention, Article 11, is about data governance. The article is core and non-actionable. The Convention is the only instrument examined that contains "shall" obligations. These bind Parties rather than AI actors directly.

Article 11, on privacy and personal data protection, requires Parties to ensure that privacy rights and personal data are protected, including through applicable laws, standards and frameworks ((a)), and that "effective guarantees and safeguards" are in place for individuals ((b)). It does not identify any specific measure and is therefore non-actionable.

G7 Hiroshima Process Guiding Principles

Three principles feature data governance. It is the core concern of one and tangential to two. All three are actionable. The Principles are voluntary and addressed to organisations developing advanced AI systems.

Principle 11, on data input measures and protections for personal data and intellectual property, has data governance as its core concern. It suggests measures to manage data quality, including training data and data collection, to mitigate harmful biases.

Data governance is tangential to two further principles:

- Principle 5, on governance and risk management policies, encourages the disclosure of privacy policies covering personal data, user prompts and system outputs.
- Principle 6, on security controls, states that security measures taken by actors could include securing datasets.

G7 Hiroshima Process Code of Conduct

The Code of Conduct builds on the G7 Principles, adding detail in several cases. Four provisions address data governance, which is the core concern of one and tangential to three. Three of the four are actionable. For the three provisions it shares with the Principles, the Code's relevance and actionability tags are identical.

Core Provision 11 provides further detail by mentioning the possibility of employing privacy-preserving training techniques, and testing and fine-tuning to ensure that systems do not divulge confidential or sensitive data. It also adds safeguards to respect privacy rights. The risk of systems disclosing personal data is also addressed in the ASEAN Expanded Guide's Section 1.3.

Tangential provisions 5 and 6 reproduce the Principles' data governance elements, namely privacy policies and the securing of datasets.

One provision contains a data governance element with no tagged counterpart in the Principles. Provision 1, on risk identification and testing, is tangential and non-actionable: violations of privacy and data protection appear as possible risks to which testing should attend.

ASEAN Guide on AI Governance and Ethics

Four provisions in the ASEAN Guide feature data governance. It is the core concern of two provisions and tangential to two. All four provisions contain an actionable data governance element. The Guide is addressed to deployers and developers and offers worked examples and named techniques.

Two provisions take data governance as their core concern:

- Principle 5, on privacy and data governance, suggests protocols governing who can access data and when. It calls for data to be collected, stored, generated and deleted in compliance with data protection laws, and lists six national data protection laws in ASEAN. It also suggests consent or another legal basis before personal data is collected, used or disclosed, and not gathering unnecessary or irrelevant data. It recommends data protection and governance frameworks, data protection impact assessments (noting that these are narrower than an AI risk assessment) and privacy by design. It is the only provision in the dataset to name specific privacy-enhancing technologies, namely differential privacy and zero-knowledge proofs.
- The guidance on data collection and processing covers representative datasets, guidelines and quality assurance for data labelling, and five factors affecting data quality. It also covers a data provenance record, storage standards and retention policies, and encryption, access controls and anonymisation. Further, it suggests data protection impact assessments, compatibility of data use with its intended purpose, and minimising sensitive or personal data in development datasets. Developers and deployers should "reasonably avoid" using protected characteristics such as gender or ethnicity to drive decisions, while using them to assess outcomes for bias. It is the only provision in the dataset to call for due diligence on third-party sources of personal data.

In the two tangential provisions, data governance serves other principles:

- Principle 2, on fairness and equity, suggests diverse and representative training datasets and bias mitigation during data collection and pre-processing.
- Section 4, on stakeholder interaction and communication, suggests informing users and obtaining their informed consent before using their data for training, and safeguarding the security of user data. It also suggests feedback channels through which users can correct inaccurate personal data. It is the only provision in the dataset that provides for the correction of personal data.

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. Two provisions are about data governance as a core concern. Both are actionable.

Section 1.3 restates the seven principles of the 2024 Guide, reproducing its definition of privacy and data governance. The new elements specific to generative AI address the scraping of internet data and the risk that models "memorise" and release training data. To meet these, the section suggests data minimisation and anonymisation, safeguards within the model to refuse to disclose personal data, and a legal basis such as consent. It also suggests data traceability and internal verification of data accuracy and integrity. Under its security principle, it adds secure data collection and storage, a response plan for data breaches and safeguards against data poisoning.

Section 3.2, on data, takes data governance as its core concern. It notes that internet-sourced training data requires attention to usage rights, provenance, accuracy and personal data protection. As areas for ASEAN to explore, it proposes a common regional approach to personal data protection and regional guidelines on data handling, storage and governance. It also proposes measurable regional benchmarks on data quality and guidance on cross-border data flows related to AI. Its proposals are directed at ASEAN as a convening body rather than at organisations. Its material on data sharing and regional language models is unrelated.

UNESCO Recommendation on the Ethics of AI

Four provisions are related to data governance, of which two are core and a further two tangential. Three of the four are actionable.

Two provisions take data governance as their core concern:

- The principle 'Right to privacy, and data protection' (paragraphs 32 to 34) calls for data to be collected, used, shared, archived and deleted consistently with international law (paragraph 32). It calls for data protection frameworks requiring a legitimate aim and a valid legal basis, including informed consent, and for data subjects to be able to exercise their rights (paragraph 33). It also calls for privacy impact assessments and privacy by design (paragraph 34).
- Policy Area 3, on data policy, is addressed to Member States. It calls for data governance strategies that continually evaluate the quality of training data and ensure data security (paragraph 71), and for privacy impact assessments and privacy by design (paragraph 72). It calls for a framework giving individuals rights over their personal data, including the ability to access and erase it, safeguards for sensitive data, and protection where data are used for micro-targeted advertising or transferred across borders (paragraph 73). It further calls for full security for sensitive data such as biometric, genetic and health data (paragraph 74), and for quality datasets built on a valid legal basis with annotation standards (paragraph 76).

In the two tangential provisions, data governance serves other objectives:

- Policy Area 11, on health and social well-being, calls for compliance with data protection requirements in health AI (paragraph 123(c)). It also calls for informed consent to the use and analysis of personal data without preventing access to health care (paragraph 123(d)).
- The principle 'Safety and security' (paragraph 27) names privacy-protective data access frameworks and quality training data only as enablers of safe AI, and is therefore non-actionable.

Joint Declaration on AI, Freedom of Expression and Media Freedom

Four provisions in the Joint Declaration are relevant to data governance. Two are core provisions and two are tangential. All four are actionable. The Declaration addresses states, companies, financial institutions and investors, as well as journalists and media actors.

Data governance is a core concern of Principle 3, on respecting privacy and data protection. Its actionable component is the call to respect international data protection standards throughout the lifecycle, "including consent, purpose limitation".

The section on state regulation of companies is also core. Its data governance component calls for strong data protection laws that enable transparency in data collection and use ((c)). These laws should limit "extractive data mining practices" and pervasive tracking and targeting of individuals online, and ensure that data extraction for AI training is based on consent. The section's other material on due diligence, oversight institutions, competition and linguistic diversity is unrelated.

In the two tangential provisions, data governance serves other objectives:

- The section addressed to the private sector calls on companies to develop and train AI systems in full respect of data protection ((d)), to develop standards for fair data extraction ((e)), and to adopt clear privacy and data protection policies ((g)).
- The section on journalists and media actors calls on them, when choosing an AI provider, to rigorously evaluate the diversity, fairness and quality of its data and its compliance with privacy and data protection ((c)).

Human Oversight

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)).

Discrimination and Bias

Our text-level analysis of nine international AI frameworks shows that discrimination and bias is a universal concern across all nine frameworks. A total of 22 provisions address discrimination and bias, with various degrees of focus and actionability.

- Nine provisions focus on discrimination and bias as the core concern, while 13 provisions only touch upon discrimination and bias tangentially.
- 16 provisions across seven frameworks provide actionable measures to pursue non-discrimination and address bias. Six provisions address discrimination and bias in non-actionable language: two each in the Joint Declaration and the OECD AI Principles, and one each in the Council of Europe Framework Convention and the G7 Code of Conduct. In the OECD Principles and the Council of Europe Framework Convention, discrimination and bias appears only in non-actionable language.
- The Joint Declaration has the most provisions on discrimination and bias (5), while the most core provisions per framework are found in the UNESCO Recommendation, ASEAN Guide, and the Joint Declaration (2 each). Both the UNESCO Recommendation and the ASEAN Guide have the most actionable provisions (4 each).

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 discrimination and bias. Principle 1.1 (Inclusive growth, sustainable development and well-being) and Principle 1.2 (Respect for the rule of law, human rights and democratic values, including fairness and privacy) both cover it tangentially, and neither is actionable.

- Principle 1.1 names the inclusion of underrepresented populations and the reduction of economic, social, gender and other inequalities as beneficial outcomes for stakeholders to pursue.
- Principle 1.2 lists non-discrimination, equality and fairness among the values AI actors should respect ((a)). It does not name any specific measure against bias.

UN General Assembly resolution

One provision in the UN General Assembly resolution A/RES/78/265, sub-paragraph 6(h), addresses discrimination and bias as a core concern. The provision is actionable. It encourages internationally interoperable frameworks, practices and standards for training and testing AI systems, to protect individuals from discrimination and bias. It also encourages avoiding the reinforcement or perpetuation of discriminatory or biased outcomes throughout the lifecycle, including by “analysing and mitigating bias encoded in datasets”.

Council of Europe Framework Convention

One Convention article features discrimination and bias as a core concern. The article is non-actionable with regard to discrimination and bias.

Article 10, on equality and non-discrimination, requires Parties to adopt measures ensuring that activities within the AI lifecycle respect equality, including gender equality, and the prohibition of discrimination (10(1)). It also requires measures aimed at overcoming inequalities to achieve fair, just and equitable outcomes (10(2)). It does not identify any specific measure.

G7 Hiroshima Process Guiding Principles

A single principle, Principle 11, addresses discrimination and bias. The Principle is tangential and actionable. The Principles are voluntary and addressed to organisations developing advanced AI systems.

Principle 11, on data input measures, suggests managing data quality, including training data and data collection, "to mitigate against harmful biases".

G7 Hiroshima Process Code of Conduct

The Code of Conduct builds on the G7 Principles, adding detail in several cases. Three Code provisions address discrimination and bias, all of them tangential. Two of the three are actionable. Provision 11 shares its relevance, actionability and requirement tags with Principle 11, being both tangential and actionable, and reproduces its call to manage data quality to mitigate harmful biases.

Two provisions contain a bias element with no tagged counterpart in the Principles:

- Provision 1, on risk identification and mitigation, is actionable. Its list of risks to which testing should attend includes the ways in which systems "can give rise to harmful bias and discrimination", and it calls for mitigation measures to address identified risks.
- Provision 8, on research, is non-actionable. It names avoiding harmful bias only as a research priority.

ASEAN Guide on AI Governance and Ethics

Four provisions in the ASEAN Guide are related to discrimination and bias. It is the core concern of two provisions and tangential to two. All four provisions contain an actionable element. The Guide is addressed to deployers and developers and offers worked examples, named techniques and testing tools.

Principle 2, on fairness and equity, is a core provision. It suggests safeguards so that algorithmic decisions do not exacerbate existing discriminatory impacts, such as human checks on algorithms and their outputs. It also suggests regular testing to detect bias and adjustments to rectify imbalances, and diverse and representative training datasets. Bias should be mitigated during data collection and pre-processing, training and inference. Its examples include screening job applications, predicting creditworthiness and an education-sector dataset representing students of different genders and ethnicities.

The guidance on data collection and processing has a large bias component alongside material on data quality and data protection, which is covered in the data governance section:

- It sets out six types of bias to watch for: representation, societal, labelling, measurement, activity and proxy bias. It is the only provision in the dataset to set out a typology of bias.
- It suggests heterogeneous and representative datasets. Developers and deployers should "reasonably avoid" using protected characteristics such as gender or ethnicity to drive decisions, while using them to assess outcomes for bias.
- It suggests avoiding the premature removal of data attributes that help identify bias, and providing labelling guidelines and unconscious-bias training for data labellers.
- It suggests testing with data from different demographic groups, citing a fairness testing framework, and regular evaluation and correction of systematic disadvantage.
- After deployment, it suggests periodic review of bias and fairness metrics. It also addresses reinforcement bias in continuously learning systems, the only provision in the dataset to do so.

In the two tangential provisions, fairness testing serves other objectives:

- Principle 1, on transparency and explainability, lists counterfactual fairness testing, which checks that a system's decisions stay the same in a counterfactual world, as a way to demonstrate repeatability.
- The guidance on outcome analysis suggests fairness testing before deployment to ensure the system does not discriminate against certain demographics of users.

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, addresses discrimination and bias. The provision is both core and actionable.

Section 1.3 restates the 2024 Guide's fairness and equity principle and adds considerations specific to generative AI. Foundation models trained on internet-sourced data may reflect social biases and over-represent well-connected populations and common languages. Bias can also be learned during fine-tuning or picked up from user inputs. Organisations are encouraged to attend to all sources of bias, including a biased organisational context, and to test the representativeness of outputs across demographics.

UNESCO Recommendation on the Ethics of AI

Four provisions in the UNESCO Recommendation address discrimination and bias: two core and two tangential. All four are actionable.

Two provisions take discrimination and bias as their core concern:

- The principle 'Fairness and non-discrimination' (paragraphs 28 to 30) calls on AI actors to safeguard fairness and non-discrimination of any kind (paragraph 28). It also calls on them to "make all reasonable efforts" to minimise and avoid reinforcing or perpetuating discriminatory or biased outcomes throughout the lifecycle, and for effective remedy against

biased algorithmic determination (paragraph 29). Its calls on Member States to promote inclusive access and tackle digital divides (paragraphs 28 and 30) concern inclusion more broadly.
- Policy Area 6, on gender, calls on Member States to ensure that gender stereotyping and discriminatory biases are not translated into AI systems, and to "identify and proactively redress these" (paragraph 90). It also calls on them to ensure that AI does not exacerbate existing gender gaps (paragraph 89) and that ethical impact assessments include a gender perspective (paragraph 87). It is the only provision in the dataset devoted to a single ground of discrimination. Its funding, education and workforce diversity measures (paragraphs 88 and 91 to 93) are unrelated.

In the two tangential provisions, discrimination and bias serves other objectives:

- Policy Area 11, on health and social well-being, calls for oversight to "minimize and mitigate bias" in health prediction, detection and treatment solutions (paragraph 123(a)).
- Policy Area 1, on ethical impact assessment, covers impacts on marginalised and vulnerable people.

Joint Declaration on AI, Freedom of Expression and Media Freedom

Five Joint Declaration provisions feature discrimination and bias: two core and three tangential. Three of the five are actionable. The Declaration addresses states, companies, financial institutions and investors, as well as journalists and media actors.

The section addressed to states is core and actionable. Its anti-discrimination component asks states to put in place safeguards against bias, discrimination and gender-based violence ((g)). These include legislation against AI-generated online violence and independent oversight bodies with equality expertise to monitor discriminatory effects of AI on minority groups.

Principle 4, on equality and non-discrimination, is a core provision but non-actionable. It states that AI systems should be designed, developed, trained and deployed "in ways that mitigate bias and discrimination", ensure inclusivity and reflect diverse cultural contexts, without naming a mechanism.

Out of the three tangential provisions:

- The section addressed to the private sector calls on companies to hire diverse staff to help lessen bias across the AI lifecycle ((b)) and to adopt clear policies guaranteeing data fairness and inclusiveness ((g)).
- The section on journalists and media actors calls on them, when choosing an AI provider, to rigorously evaluate the diversity, fairness and quality of its data ((c)).
- The section on state regulation of companies names the identification and correction of biases only as a purpose of transparency requirements ((a)(iii)), and is therefore non-actionable.

Information Integrity

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.

Labour

Our text-level analysis of nine international AI frameworks shows that labour is found across six of the nine frameworks. It is not part of either of the G7 documents, the Principles and the Code of Conduct, or the Council of Europe Framework Convention. A total of seven provisions address labour, with various degrees of focus and actionability.

- Three provisions, in the ASEAN Guide, the UN General Assembly resolution, and the UNESCO Recommendation, focus on labour as the core concern, while four provisions only touch upon labour tangentially.
- Three provisions across two frameworks, the ASEAN Guide and the UNESCO Recommendation, provide actionable measures to address the impact of AI on workers. The remaining four provisions address labour in non-actionable language, including the UN General Assembly resolution's core provision on labour markets.
- The ASEAN Guide has the most provisions on labour (2), all of which are also actionable.

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 labour. Principle 1.5 (Accountability) covers it tangentially and is non-actionable: section (c) names labour rights as one of several risk categories that should be addressed by risk management.

UN General Assembly resolution

One provision in the UN General Assembly resolution A/RES/78/265, sub-paragraph 6(q), addresses labour. The provision is core and non-actionable. It encourages research and international cooperation to develop measures for identifying and assessing the impacts of AI deployment on labour markets. It also encourages support for mitigating potential negative consequences for workforces, and programmes on digital training and capacity building. It gives particular attention to developing and least developed countries, but names no concrete measure to protect workers.

ASEAN Guide on AI Governance and Ethics

Two provisions in the ASEAN Guide relate to labour. Labour is of core relevance in one provision and tangential to one. Both are actionable. The Guide's labour measures are addressed to deployers, making it the only framework in the dataset with actionable labour measures addressed to organisations rather than states.

- Principle 4, on human-centricity, is core. It states that an extension of the principle of human-centricity is ensuring that the adoption and large-scale deployment of AI does not "unduly disrupt labour and job prospects without proper assessment". It encourages deployers to carry out impact assessments for a systematic, stakeholder-based review and to consider how jobs can be redesigned to incorporate AI, pointing to national guidance on job redesign.

- Section 4, on stakeholder interaction and communication, is tangential. It calls on deployers to communicate adequately with employees about the AI systems deployed and how they change workflows, and to provide training and education so employees can work with AI. It also calls on deployers to redesign jobs so that AI augments workers' existing skills rather than rendering them redundant.

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, relates to labour tangentially, and without naming an actionable measure. Under its human-centricity principle, it names "employment disruptions" only as a possible consequence of an approach that is not human-centric. It does not restate the 2024 Guide's call for workforce impact assessments and job redesign.

UNESCO Recommendation on the Ethics of AI

One provision, Policy Area 10, addresses labour as a core concern. The provision names actionable measures. The policy area is addressed to Member States.

Policy Area 10, on economy and labour, calls on Member States to work with companies, civil society, workers and unions to ensure a "fair transition for at-risk employees" (paragraph 118). This includes upskilling and reskilling programmes, mechanisms to retain employees during transition periods and "safety net" programmes for those who cannot be retrained. It is the only provision in the dataset to set out transition measures for workers displaced by AI. The same paragraph calls for enhanced social protection, public funding for these programmes and a review of tax regimes to counteract unemployment caused by AI-based automation. The policy area also calls for assessing the impact of AI on labour markets and education requirements (paragraph 116), collaboration on skills (paragraph 117) and research on local labour impacts (paragraph 119). Its material on competition (paragraph 120) is unrelated.

Joint Declaration on AI, Freedom of Expression and Media Freedom

One provision, Principle 5, relates to labour tangentially. It is non-actionable. Principle 5, on equitable access to opportunities, states that equitable access to AI can help people meet their "labour, economic, cultural and social needs". The Principle goes on to name closing the digital divide and promoting digital literacy as priorities, and does not name any specific measure for workers.

Intellectual Property

Our text-level analysis of nine international AI frameworks shows that intellectual property is found across eight of the nine frameworks, though mostly in passing. Intellectual property is absent from the Council of Europe Framework Convention. A total of 12 provisions address intellectual property, with various degrees of focus and actionability.

- Only a single provision, in the UN General Assembly resolution, focuses on intellectual property as the core concern, while 11 provisions touch upon intellectual property tangentially.
- Two provisions across two frameworks, the ASEAN Expanded Guide and the Joint Declaration, provide actionable measures to address intellectual property. The remaining 10 provisions address intellectual property in non-actionable language.
- The G7 Code of Conduct has the most provisions on intellectual property (4).

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 intellectual property. Principle 1.5 (Accountability) covers it tangentially and is non-actionable: section (c) names intellectual property rights as one of several risk categories to address through risk management.

UN General Assembly resolution

One provision in the UN General Assembly resolution A/RES/78/265, sub-paragraph 6(i), addresses intellectual property. The provision is core and non-actionable. It encourages, where appropriate and relevant, "appropriate safeguards to respect intellectual property rights, including copyright-protected content, while promoting innovation", without naming any specific safeguarding measure. It is the only provision in the dataset with intellectual property as its core concern.

G7 Hiroshima Process Guiding Principles

One principle, Principle 11, relates to intellectual property tangentially, and without proposing an actionable measure. Its title refers to protections for intellectual property, but its text addresses only data quality, the transparency of training datasets and compliance with applicable legal frameworks.

G7 Hiroshima Process Code of Conduct

The Code of Conduct builds on the G7 Principles, adding detail in several cases. Four provisions address intellectual property. All are tangential and non-actionable. Provision 11 shares its relevance and actionability tags with Principle 11. However, the Code's version adds an encouragement to implement "appropriate safeguards" to respect intellectual property rights, "including copyright-protected content", without specifying any measure.

Three provisions contain an intellectual property reference with no counterpart in the Principles:

- Provision 4, on information sharing, states that reporting on incidents should safeguard intellectual property rights.
- Provision 6, on security controls, states that insider threat detection should be consistent with protections for organisations' "most valuable intellectual property and trade secrets", including by limiting access to proprietary and unreleased model weights.
- Provision 8, on research, names safeguarding intellectual property rights as one of many research priorities.

ASEAN Guide on AI Governance and Ethics

One provision, Principle 1, concerns intellectual property tangentially and without actionable measures. It notes only that auditability does not require making confidential information about business models or intellectual property publicly available.

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 actionable provision, Section 1.3, relates to intellectual property tangentially.

Under its security principle, Section 1.3 states that developers should "adopt measures to safeguard any retained copyrighted works used in training". It is the only provision in the dataset to address copyrighted works retained after training. Under its privacy principle, it notes that scraping training data from the internet raises intellectual property issues, without proposing a measure.

UNESCO Recommendation on the Ethics of AI

One provision addresses intellectual property. The provision is tangential and non-actionable. Policy Area 7, on culture, calls on Member States to foster research on whether or how to protect works created by means of AI with intellectual property rights (paragraph 99). It also calls on them to assess how AI affects the rights and interests of owners whose works are used to research, develop, train or implement AI applications. It is the only provision in the dataset to raise the issue of potential intellectual property protections for AI-created works.

Joint Declaration on AI, Freedom of Expression and Media Freedom

Two provisions relate to intellectual property, both of which are tangential and one of which is actionable.

Subsection (c) of the section on state regulation of companies calls on states to adopt data protection, copyright and similar rules ensuring that "data extraction for AI training is based inter alia on consent and compensation". It is the only provision in the dataset to call for compensating rights-holders whose content is used for AI training.

Section (d) of the section on the private sector states that AI systems should be developed and trained in full respect of copyright, without proposing any further concrete measures.

Accountability

Our text-level analysis of nine international AI frameworks shows that accountability is found across six of the nine frameworks. Accountability is absent from the UN General Assembly resolution and both the G7 Principles and Code of Conduct. A total of 17 provisions address accountability, with various degrees of focus and actionability.

- 10 provisions focus on accountability as the core concern, while seven provisions only touch upon accountability tangentially.
- 12 provisions across six frameworks provide actionable measures to establish accountability. Five provisions address accountability in non-actionable language.
- The Joint Declaration has the most overall provisions on accountability (5), while the UNESCO Recommendation has the most core provisions (3). The UNESCO Recommendation, along with the ASEAN Guide, also has the most actionable provisions (3 each).

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 accountability. Principle 1.5 (Accountability) covers it as core concern and is actionable. Paragraph (a) calls for AI actors to be accountable for the proper functioning of AI systems and for respecting the OECD principles, "based on their roles, the context" and the state of the art. Paragraphs (b) and (c) name traceability and systematic risk management as possible mechanisms, though these are covered more fully in the transparency and safety sections.

Council of Europe Framework Convention

Three Convention Articles address accountability. Articles 9 and 14 are core, while Article 15 is tangential. Only Article 14 is actionable. The Convention is the only instrument examined that contains "shall" obligations. These bind Parties rather than AI actors directly.

Two articles take accountability as their core concern:

- Article 9, on accountability and responsibility, requires Parties to adopt measures "to ensure accountability and responsibility" for adverse impacts on human rights, democracy and the rule of law. It identifies no specific mechanism to achieve this and is therefore non-actionable.
- Article 14, on remedies, is actionable. The Article requires accessible and effective remedies for human rights violations resulting from AI systems, including by providing for the effective possibility to lodge a complaint with competent authorities.

Article 15, on procedural safeguards, is tangential and non-actionable, since it requires effective procedural guarantees, safeguards and rights only in general terms.

ASEAN Guide on AI Governance and Ethics

Three provisions in the ASEAN Guide relate to accountability. It is the core concern of one provision and tangential to two. All three are actionable. The Guide is addressed to deployers and developers and concentrates on allocating responsibility within and between organisations.

Principle 6, on accountability and integrity, is the core provision. It states that deployers "should be accountable for decisions made by AI systems" and for compliance with applicable laws and AI ethics. It defines AI actors as any natural or legal person involved in at least one stage of the AI lifecycle; it is the only provision in the dataset to define AI actors. It suggests that organisations adopt clear reporting structures setting out roles and responsibilities across the lifecycle. Responsible individuals should take mitigating action after a malfunction or misuse, and errors or unethical outcomes should be documented and corrected. It is the only provision in the dataset to place accountability for AI decisions specifically on deployers.

In the two tangential provisions, accountability serves other objectives:

- The guidance on project governance suggests that roles and responsibilities be clearly defined "with the accompanying accountability mechanisms" before a project progresses. It also suggests that deployers should take steps to resolve or mitigate potential non-compliance by third-party developers or vendors.
- Section 4, on stakeholder interaction and communication, suggests that deployers may contractually oblige developers to provide relevant information on data, model development and safeguards, distributing obligations along the value chain.

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. Two provisions feature accountability as a core concern while naming actionable measures.

Like the 2024 Guide, Section 1.3 of the Expanded Guide names accountability and integrity as a principle. It mentions that control of generative AI may shift between deployers, developers and users across "multiple layers in the tech stack", which affects how responsibility is shared. It calls for "clear, adaptable terms and conditions" to manage accountability as systems evolve.

Section 3.1 on Accountability states that developers, cloud service providers, deployers, and users have a shared responsibility for ethical AI use, transparency, and trust, and suggests shared responsibility frameworks as a possible way for determining accountability.

UNESCO Recommendation on the Ethics of AI

Three provisions are focused on accountability: all three are core, and all are actionable.

Two principles take accountability as their core concern:

- The principle 'Human oversight and determination' (paragraphs 35 and 36) calls on Member States to ensure that ethical and legal responsibility for any stage of the AI lifecycle, and in

cases of remedy, can always be attributed to physical persons or existing legal entities (paragraph 35). It adds that AI "can never replace ultimate human responsibility and accountability" (paragraph 36).
- The principle 'Responsibility and accountability' (paragraphs 42 and 43) states that ethical responsibility and liability for decisions and actions based on AI should always be attributable to AI actors according to their role in the lifecycle (paragraph 42). It calls for oversight, impact assessment, audit and due diligence mechanisms, "including whistle-blowers' protection", to ensure accountability (paragraph 43). It is the only provision in the dataset to mention whistle-blower protection.

Policy Area 2, on ethical governance and stewardship, is also core. It is the only provision in the dataset to address responsibility, liability and oversight institutions together:

- Member States should ensure that governance mechanisms include the possibility of mitigation and redress across borders (paragraph 54).
- Member States should ensure that harms caused through AI systems are investigated and redressed, through strong enforcement mechanisms, remedial actions and remediation by companies, and should strengthen their institutional capacities to this end (paragraph 55).
- Member States should develop regulatory frameworks for accountability and responsibility, and liability frameworks where necessary. Ultimate responsibility should lie with natural or legal persons, and "AI systems should not be given legal personality themselves" (paragraph 68). It is the only provision in the dataset to rule out legal personality for AI.

Joint Declaration on AI, Freedom of Expression and Media Freedom

Five provisions address accountability, one as a core concern and four tangentially. Two of the five are actionable. The Declaration addresses states, companies, financial institutions and investors, as well as journalists and media actors.

Principle 6, on transparency, accountability and effective remedies, is the core provision but non-actionable. It names accountability, including effective remedies, as a key principle, but the only measure it specifies, meaningful information and access to data, concerns transparency.

Out of the four tangential provisions:

- The section on journalists and media actors calls on them to identify someone in the organisation who is "clearly accountable" for the implementation and outcomes of AI systems ((a)). It is the only provision in the dataset to call for designating a single accountable person.
- The section addressed to the private sector calls on companies to comply with liability regimes at local, regional and international levels ((j)), as well as ensuring the availability of effective remedies for freedom of expression ((h)).
- Principle 2, on a pluralistic information environment, names regulation to ensure transparency and accountability only in general terms, and is therefore non-actionable.
- The section on state regulation mentions accountability by calling on states to "establish robust institutions for oversight and accountability of AI" ((b)), as well as calling for transparency requirements that help "attribute liability, and provide for effective remedies" ((a)(iii)). It does not suggest specific actionable measures with regard to accountability.

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