AI Governance Handbook

A practical guide to AI regulation, governance, safety, and compliance.

By Khullani M. Abdullahi, J.D.

Version 2.0.0 · 2026-05-11

Techne AI

Introduction

Artificial intelligence is no longer an emerging technology. By May 2026, generative AI has been mainstream for more than three years; frontier models from Anthropic, Google, OpenAI, and xAI are routinely deployed inside regulated workflows; and the question facing every organisation has shifted from "should we govern AI?" to "how do we govern AI without slowing down what already works?"

AI Governance is the set of frameworks, controls, and processes an organisation uses to ensure that the AI systems it builds, buys, or deploys are lawful, safe, fair, and aligned with its values. This handbook is a practical primer covering five neighbouring disciplines — governance, safety, trustworthiness, responsible AI, and risk management — tailored for practitioners, compliance officers, executives, and policymakers.

How this edition is different

The previous edition (March 2025) was written before the second major wave of AI regulation. Since then:

- The European Union has begun enforcing the AI Act and, on **7 May 2026**, the Council and Parliament reached a provisional agreement on a Digital Omnibus that rebases the high-risk obligation deadline from August 2026 to **December 2027**.[1]
- The United States has replaced Executive Order 14110 with **Executive Order 14179**, published the America's AI Action Plan, and issued a December 2025 executive order targeting state-law obstruction of federal AI policy.[2]
- **Colorado** (SB 24-205, effective 30 June 2026), **Texas** (TRAIGA, effective 1 January 2026), and **California** (SB 53, SB 942, AB 2013) have enacted binding state laws.[3]
- **South Korea** and **Japan** have national AI statutes; **Canada's** Bill C-27/AIDA died on the order paper; **China's** synthetic-content labelling rules took effect September 2025.[4]
- **ISO/IEC 42005:2025** (AI impact assessment) and **ISO/IEC 42006:2025** (audit/certification body requirements) joined ISO/IEC 42001:2023.[5]
- The **NIST AI Safety Institute** was renamed the **Center for AI Standards and Innovation (CAISI)** in June 2025.[6]
- The largest US AI copyright case to date — Bartz v. Anthropic — settled in September 2025 for **$1.5 billion**.[7]

Two entirely new chapters cover US state laws and Copyright & IP. A third new chapter, Frontier Models, pulls together GPAI obligations, the EU Code of Practice, frontier-safety frameworks, and the international AI Safety Institute network.

How to read this handbook

This handbook is meant to be useful at three levels:

1. **As a reference.** Each chapter is self-contained. Compliance officers can jump to EU AI Act or US Federal; engineers can jump to Technical Safety; executives can jump to Audience Guidance.
2. **As a maturity assessment.** The six Maturity Models chapters provide seven-stage progressions for governance, safety, trust, responsible AI, risk, and compliance — useful for benchmarking and roadmap planning.
3. **As a primer.** Read it cover-to-cover for a comprehensive view of the field as of mid-2026.

A note on stability. Some material in this edition is fact (signed, in force, or published) and some is pending. We flag pending items inline. The most volatile area at the time of writing is the EU AI Act Omnibus, which has been agreed in principle but is awaiting formal adoption. Where the text says "PENDING," verify with the source linked in References.

Looking forward

Three trends will dominate the next eighteen months:

1. **Compliance as a product feature.** ISO/IEC 42006 unlocked credible third-party certification of AI management systems in 2025. Vendors that can produce a 42001 certificate and a GPAI Code of Practice signature are now winning enterprise procurement on those grounds alone.
2. **State-law fragmentation in the US.** With the federal preemption executive order signed in December 2025 but no statute attached to the NDAA, state laws will continue to multiply through 2026-2027 until either a federal AI statute passes or the courts resolve preemption challenges.
3. **Frontier-model governance maturing.** The CAISI testing agreements, the EU GPAI Code, the UK Blueprint, and Korea's frontier-safety track are converging on a shared template: pre-deployment evaluations, post-deployment incident reporting, and structured disclosure of capabilities, limitations, and known failure modes.

What this means for organisations: AI governance is no longer an optional uplift — it is the cost of doing business with AI in any meaningful market. The good news is that the building blocks are now well understood. The remainder of this handbook is a practical guide to assembling them.

---

1. Council of the EU. (2026, May 7). Artificial intelligence: Council and Parliament agree to simplify and streamline rules. ↩
2. The White House. (2025, January 23). Removing Barriers to American Leadership in Artificial Intelligence (EO 14179). ↩
3. See chapter on US State Laws for full text and citations. ↩
4. See chapter on International Regulation for full text and citations. ↩
5. ISO/IEC. 42005:2025 — AI System Impact Assessment; 42006:2025 — Requirements for Bodies Providing Audit and Certification of AI Management Systems. ↩
6. NIST. Center for AI Standards and Innovation (CAISI). ↩
7. Authors Guild. What Authors Need to Know About the Anthropic Settlement. ↩

Legal & Regulatory Frameworks

AI systems must comply with an increasingly dense landscape of laws, regulations, and standards. Where the 2025 edition of this handbook treated AI law as an emerging field, the 2026 edition treats it as established: dozens of binding instruments are in force across the EU, US states, the UK, Korea, Japan, China, and beyond, and the major international standards bodies have published certifiable AI management standards.

This chapter is organised by jurisdiction and instrument type. Use it as a reference; cross-references between sections highlight overlaps (e.g., GPAI obligations under the EU AI Act and the GPAI Code of Practice covered in Frontier Models).

What's in this chapter

- **International Standards (ISO)** — ISO/IEC 42001, 23894, 22989, and the 2025 additions of 42005 (impact assessment) and 42006 (audit/certification body requirements).
- **EU AI Act** — the world's first horizontal AI statute, now in phased enforcement and reshaped by the May 2026 Digital Omnibus.
- **US Federal** — EO 14179, the America's AI Action Plan, OMB memos M-25-21 and M-25-22, CAISI, NIST AI RMF, and the December 2025 preemption executive order.
- **US State Laws** — Colorado SB 24-205, Texas TRAIGA, California SB 53 / AB 2013 / SB 942, Utah SB 226, Tennessee ELVIS Act, NYC Local Law 144.
- **International** — United Kingdom, Canada, South Korea, Japan, China, Brazil, Singapore, plus OECD and UNESCO principles.
- **Sectoral** — FDA AI/ML medical-device guidance, OCC and Federal Reserve model-risk management, CFPB, EEOC employment AI.
- **Copyright & IP** — Bartz v. Anthropic, Kadrey v. Meta, training-data disclosure laws, and where the law is still unsettled.

Reading guide

Organisations active in **multiple jurisdictions** should start with ISO, then layer the most binding regime on top: EU AI Act for any product or service touching EU users; US Federal plus the relevant state laws for US deployment; the International chapter for region-specific obligations. Sector-specific overlays (financial services, healthcare, employment) come last.

Organisations operating in a **single jurisdiction** can read just the relevant national chapter, the ISO section (for management-system architecture), and the Sectoral section if applicable.

International Standards (ISO/IEC)

Three new ISO/IEC standards joined the AI governance toolkit during 2025: **ISO/IEC 42005** (AI System Impact Assessment) and **ISO/IEC 42006** (Requirements for Bodies Providing Audit and Certification of AI Management Systems). With 42006 in place, third-party certification of AI management systems under 42001 became credible during the second half of 2025 — **Anthropic** received the first such certification in January 2025; **IBM Granite**, **UiPath**, and **Changi Airport** followed.[1]

ISO/IEC 42001:2023 – AI Management Systems

Published in December 2023, ISO/IEC 42001 is the first global standard for AI management systems.[2] It is the AI-specific analogue of ISO 9001 (quality) and ISO 27001 (information security): a **certifiable management-system standard** that requires organisations to establish an AI governance policy, senior-leadership commitment, risk management processes, resource allocation, and operational controls covering the AI lifecycle.

42001 is sector-agnostic and proportionate to organisation size. Achieving certification demonstrates to regulators and customers that AI projects are managed against recognised baselines for ethics, transparency, accountability, bias mitigation, safety, and privacy.

What changed in 2025-2026. With ISO/IEC 42006 now defining requirements for the bodies that audit and certify AIMS, the certification pathway is real rather than theoretical. Certification is increasingly relevant for procurement — enterprise buyers and regulators are starting to ask for it as a baseline.

ISO/IEC 23894:2023 – AI Risk Management

ISO/IEC 23894 is the AI-specific companion to ISO 31000 (generic risk management).[2:1] It guides organisations through identifying risks across the AI lifecycle — from data collection and model training to deployment and monitoring — assessing severity, and treating risks with appropriate controls.

Together, 23894 (risk management) and 42001 (management system) form a coherent toolkit: 42001 establishes the governance structure, 23894 provides the risk-specific procedures.

ISO/IEC 22989:2022 – AI Concepts and Terminology

ISO/IEC 22989 is the foundational terminology standard for AI. Where ambiguity matters — in contracts, in compliance documentation, in incident reporting — aligning on 22989 definitions of AI system, AI agent, model, training, and similar terms reduces downstream disputes. Definitions used in this handbook's Glossary are aligned with 22989 where the standard speaks.

ISO/IEC 42005:2025 – AI System Impact Assessment

Published in May 2025, **ISO/IEC 42005:2025** is the first international standard dedicated to AI impact assessment.[3] It provides lifecycle guidance for assessing how an AI system affects individuals, groups, and society — covering risk identification, stakeholder analysis, evaluation of impacts on fundamental rights, and documentation.

42005 is not certifiable on its own but feeds into 42001 compliance: an AI management system that conforms to 42001 will use 42005 as the operational guide for impact assessments. It also aligns with the **Fundamental Rights Impact Assessment** required under Article 27 of the EU AI Act for certain high-risk

systems, making 42005 a natural choice for organisations needing a single methodology that satisfies both standards-based and regulatory regimes.

ISO/IEC 42006:2025 — Audit and Certification

Published in 2025, **ISO/IEC 42006** defines the requirements for bodies that audit and certify AI management systems against 42001.[4] In practical terms, 42006 is what makes 42001 certification credible: it ensures that certification bodies have appropriate competence, independence, and process, so that a 42001 certificate from one accredited body means the same thing as a 42001 certificate from another.

The combined effect of 42001 + 42005 + 42006 in 2025 is that AI management-system certification now has the same architectural completeness as quality management (ISO 9001 + 9000 family + 17021) or information security (ISO 27001 + 27000 family + 17021). Expect certification to become a procurement default in regulated sectors during 2026–2027.

Other relevant ISO/IEC standards

- **ISO/IEC 38507:2022** — Governance implications of the use of AI by organisations. A governance-board-level companion to 42001.
- **ISO/IEC 5338:2023** — AI system life cycle processes. Practical lifecycle reference, particularly useful for engineering teams.
- **ISO/IEC TR 24028:2020** — Trustworthiness in AI. Technical report on trustworthy AI characteristics.
- **ISO/IEC TR 24368:2022** — Overview of ethical and societal concerns. Useful framing for ethics teams.

How to use these standards

For organisations starting from scratch, a typical path is:

1. Adopt **22989** terminology in internal documentation.
2. Build the management system to **42001**, using **38507** for board-level governance hooks.
3. Use **23894** to design the risk management process inside 42001.
4. Use **42005** as the operational guide for impact assessments.
5. Pursue **42006-accredited certification** when ready for external assurance.

For organisations already certified to **ISO 27001** or **ISO 9001**, 42001 is intentionally compatible: shared management-system clauses mean an integrated management system covering quality, security, and AI is feasible and often the most efficient implementation.

---

1. Anthropic. (2025, January). ISO/IEC 42001 certification. See also reporting on IBM Granite, UiPath, and Changi Airport 2025 certifications. ↩
2. Osler, Hoskin & Harcourt LLP. The role of ISO/IEC 42001 in AI governance. ↩ ↩
3. ISO/IEC. 42005:2025 — Information technology — Artificial intelligence — AI System Impact Assessment. ↩
4. ISO/IEC. 42006:2025 — Information technology — Artificial intelligence — Requirements for bodies providing audit and certification of AI management systems. ↩

EU AI Act

The EU AI Act (**Regulation (EU) 2024/1689**) entered into force on **1 August 2024**. It is the world's first horizontal AI statute, taking a risk-based approach that classifies AI systems into four tiers with corresponding obligations.[1]

The Act has been enforced in phases since February 2025. **On 7 May 2026, the Council and Parliament reached a provisional agreement on a "Digital Omnibus on AI"** that rebases several deadlines and adds new prohibitions. The Omnibus changes are described in detail below; readers should note that the Omnibus is **PENDING formal adoption** at the time of this writing.[2]

Risk classification

Tier	Description	Examples
Unacceptable	Banned AI practices under Article 5	Government social scoring, manipulative AI exploiting vulnerable groups, untargeted scraping for facial recognition databases, emotion recognition in workplaces and schools, real-time remote biometric identification in public spaces (with narrow law-enforcement exceptions), CSAM/NCII generation (added by Omnibus)
High	Systems subject to conformity assessment, technical documentation, risk management, transparency, human oversight	AI in education, employment, essential services, law enforcement, migration, justice, biometric categorisation, critical infrastructure, medical devices
Limited	Transparency obligations only	Chatbots, deepfakes, AI-generated content
Minimal	No new obligations beyond existing law	Spam filters, AI in video games, most enterprise tooling

Figure: The EU AI Act risk pyramid — Unacceptable (banned, with CSAM/NCII added in the May 2026 Omnibus), High Risk (heavily regulated, deadline rebased to 2 December 2027), Limited Risk (transparency obligations from 2 August 2026), Minimal Risk (largely unregulated).[^babl]

Figure: EU AI Act enforcement timeline reflecting the 7 May 2026 Digital Omnibus political agreement — high-risk Annex III obligations move from 2 December 2026 to 2 December 2027; Annex I obligations move from 2 August 2027 to 2 August 2028.

Phased timeline (current, including Omnibus)

Date	Event	Status
1 August 2024	Regulation enters into force	In force
2 February 2025	Article 5 prohibitions and AI-literacy obligations (Art. 4) apply	**In force**
4 February 2025	Commission Guidelines on Prohibited AI Practices	Published[3]
10 July 2025	GPAI Code of Practice published (Transparency, Copyright, Safety & Security)	Published[4]
2 August 2025	GPAI obligations (Arts. 53, 55), AI Office governance, penalties	**In force**
19 November 2025	Commission proposes the Digital Omnibus on AI	Published
7 May 2026	Council/Parliament provisional agreement on Omnibus	**Pending formal adoption**
2 August 2026	Article 50(2) synthetic-content marking obligations	In force (transition to 2 December 2026)
2 December 2026	Original high-risk Annex III deadline	**Rebased to 2 December 2027 by Omnibus**
2 December 2027	High-risk Annex III obligations (new Omnibus date)	Future
2 August 2027	Article 6(1) Annex I obligations (original)	**Rebased to 2 August 2028 by Omnibus**
2 August 2028	Annex I obligations (new Omnibus date)	Future

What's actually live today (May 2026)

Three categories of obligation are enforceable now:

1. **Article 5 prohibitions** (since February 2025). Untargeted scraping for facial-recognition databases, social scoring by public authorities, workplace and educational emotion recognition, and the other banned practices in Article 5 must have ceased. Penalties reach **EUR 35 million or 7% of global annual turnover**, whichever is higher.[5]
2. **AI literacy** under Article 4 (since February 2025). Providers and deployers must take measures to ensure sufficient AI literacy among staff dealing with AI systems.
3. **General-purpose AI (GPAI) obligations** under Articles 53 and 55 (since August 2025). Providers of GPAI models must maintain technical documentation, publish a sufficiently detailed summary of training content, comply with Union copyright law (including Article 4(3) of the CDSM Directive on text and data mining), and — for models posing systemic risk — conduct adversarial testing, track and report serious incidents, and ensure cybersecurity protection.

The GPAI Code of Practice

Published on **10 July 2025**, the **EU GPAI Code of Practice** is a voluntary instrument intended to demonstrate adherence to the GPAI obligations.[4:1] It has three chapters — **Transparency**, **Copyright**, and **Safety & Security** — and was endorsed on 1 August 2025.

Signatories include **Google**, **Microsoft**, **OpenAI**, and **Anthropic**. **Meta** declined to sign, citing concerns about scope and legal uncertainty.

For non-signatories, compliance with the underlying obligations is assessed directly under the Act. The Code is therefore a safe-harbour-like mechanism: signing it does not exempt a provider from the Act, but it provides a structured way to demonstrate compliance and reduces enforcement risk.

See Frontier Models for the substantive obligations the Code operationalises and the parallel CAISI testing agreements in the United States.

The May 2026 Digital Omnibus (PENDING)

On **7 May 2026**, the Council and Parliament reached a provisional political agreement on the Digital Omnibus on AI, a Commission proposal published 19 November 2025 to "simplify and streamline" the Act.[2:1] The agreement is **pending formal adoption** by both institutions before becoming law; the most important changes are:

- **High-risk Annex III deadline moved from 2 December 2026 to 2 December 2027** (16-month delay).
- **Article 6(1) Annex I deadline moved from 2 August 2027 to 2 August 2028** (12-month delay).
- **New Article 5 prohibition** for AI systems generating child sexual abuse material (CSAM) or non-consensual intimate imagery (NCII).
- **Article 50(2)** synthetic-content marking obligations remain at 2 August 2026, with a transitional period until **2 December 2026** during which providers are not subject to enforcement for technically demonstrable best-efforts compliance.
- Administrative and reporting simplifications for SMEs and for providers operating under the AI Office's GPAI framework.

Practical implication. Organisations that built compliance roadmaps around the August 2026 high-risk deadline now have an additional 16 months. However, the prohibitions, AI-literacy obligations, GPAI duties, and Article 50 synthetic-content marking obligations are unchanged or accelerated. Do not assume the

Omnibus means "the Act is delayed"; it means certain high-risk and Annex I duties are delayed while the rest of the Act continues to enforce on its original schedule.

High-risk obligations (Article 6 et seq.)

For systems classified high-risk under Article 6 (whether by inclusion in Annex III or by Article 6(1) Annex I), providers must:

- Establish a **risk management system** covering the entire AI lifecycle (Art. 9).
- Use training, validation, and test data meeting **data and data-governance requirements** (Art. 10).
- Produce **technical documentation** demonstrating compliance (Art. 11).
- Implement **logging** sufficient to ensure traceability (Art. 12).
- Provide **transparency** and instructions for use to deployers (Art. 13).
- Enable **human oversight** (Art. 14).
- Meet appropriate levels of **accuracy, robustness, and cybersecurity** (Art. 15).
- Undergo **conformity assessment** before placing on the market (Arts. 43, 44).
- Register in the **EU database** (Art. 71).
- Conduct a **Fundamental Rights Impact Assessment** (Art. 27, deployer obligation).

ISO/IEC 42001 alignment substantially supports compliance with Articles 9, 10, and 17 (quality management system requirements); see ISO standards for the standards stack.

Penalties

Three tiers of fines apply under Article 99:[5:1]

Violation	Maximum
Prohibited practices (Article 5)	EUR 35 million or 7% global turnover
High-risk and most other obligations	EUR 15 million or 3% global turnover
Supplying incorrect, incomplete, or misleading information to authorities	EUR 7.5 million or 1% global turnover

SMEs and start-ups face proportionate fines — whichever of the percentage or absolute amount is **lower**, rather than higher.

Enforcement architecture

Three bodies share enforcement:

1. **National competent authorities** in each Member State for high-risk and most other obligations.
2. The **AI Office** within DG CONNECT for GPAI, cross-border cases, and Code of Practice oversight (in force since August 2025).
3. The **European AI Board** for harmonisation and coordination.

Notified bodies, designated under the Act, conduct conformity assessments for high-risk systems requiring third-party evaluation.

Practical compliance checklist (May 2026)

- [ ] Confirm none of your AI systems fall under Article 5 prohibitions.

- [ ] Document AI literacy training for relevant staff.
- [ ] If you provide a GPAI model: maintain training-data summary, technical documentation, copyright compliance evidence; consider signing the Code of Practice.
- [ ] If your model meets systemic-risk thresholds (10^{25} FLOPs or designated by the AI Office): implement adversarial testing, incident reporting, and cybersecurity controls.
- [ ] For any limited-risk system (chatbot, deepfake, generative content): plan for the 2 August 2026 marking obligations and the 2 December 2026 transitional deadline.
- [ ] For any high-risk system: build the Article 9-15 compliance evidence base on the new 2 December 2027 timeline; align with ISO/IEC 42001 to reduce duplicative effort.

---

1. European Commission. Regulatory framework on artificial intelligence. ↩
2. Council of the EU. (2026, May 7). Artificial intelligence: Council and Parliament agree to simplify and streamline rules. ↩ ↩
3. European Commission. (2025, February 4). Guidelines on prohibited artificial intelligence practices. ↩
4. EU GPAI Code of Practice. code-of-practice.ai. ↩ ↩
5. Artificial Intelligence Act EU. Article 99 — Penalties. ↩ ↩

US Federal

US federal AI policy underwent a complete reorientation in 2025. Executive Order 14110 (Biden, October 2023) was rescinded on **20 January 2025**; **Executive Order 14179** ("Removing Barriers to American Leadership in Artificial Intelligence") replaced it three days later. In July 2025, the White House published the America's AI Action Plan, accompanied by three further executive orders covering exports, datacenter permitting, and federal procurement. In December 2025, the administration issued an executive order targeting state-law obstruction of federal AI policy. NIST's AI Safety Institute was renamed the **Center for AI Standards and Innovation (CAISI)** in June 2025, with a mission shift toward standards, national security, and competitiveness.

What did **not** change: the NIST AI Risk Management Framework remains the foundational US voluntary framework, and underlying anti-discrimination law (Title VII, ADA, ADEA, FCRA, ECOA) continues to apply to AI systems — even where the EEOC and OFCCP withdrew their AI-specific guidance in January 2025.

Executive Order 14179 – Removing Barriers to American Leadership in AI (January 2025)

EO 14179, signed **23 January 2025**, is the centrepiece of the current administration's AI policy.[1] Its core elements:

- **Rescission of EO 14110** and related agency guidance issued under it.
- A directive that "AI policy promotes the United States' ability to lead in AI" and removes regulatory barriers to AI development.
- Direction to OMB to revise federal AI-use and procurement guidance (delivered as M-25-21 and M-25-22 in April 2025).
- Direction to develop an AI Action Plan (delivered in July 2025).

EO 14179 is silent on many topics that EO 14110 had addressed in detail (e.g., the Defense Production Act reporting obligations for frontier models). Where prior agency action was taken under EO 14110 authority, it has generally been wound down or repurposed.

OMB M-25-21 and M-25-22 (April 2025)

On **3 April 2025**, OMB issued two memoranda replacing the Biden-era M-24-10 (federal AI use) and M-24-18 (federal AI procurement):

- **M-25-21** — Federal Use of AI. Sets governance, risk management, and transparency requirements for federal agencies using AI, with particular focus on "high-impact" AI (a recast of "rights-impacting" and "safety-impacting" categories under M-24-10).
- **M-25-22** — Federal AI Procurement. Sets procurement standards and supplier requirements for federal AI purchases.

Each cabinet department was required to publish an **AI Compliance Plan** within 180 days; the CFPB published one of the most detailed plans on **26 September 2025**.[2]

America's AI Action Plan (July 2025)

Published **23 July 2025**, America's AI Action Plan organises 90+ federal actions into three pillars: **Innovation**, **Infrastructure**, and **Diplomacy**.[3] Three accompanying executive orders implement specific Action Plan items:

- **AI Technology Stack** export EO — rationalises export controls for AI hardware and software.
- **Datacenter permitting** EO — accelerates federal permitting for AI infrastructure.
- **"Unbiased AI Principles"** procurement EO — defines federal procurement criteria around model "bias," "ideology," and disclosure.

The Action Plan also formalised the **CAISI** mission (see below) and committed the United States to specific milestones in the **International Network of AI Safety Institutes**.

December 2025 preemption executive order

On **11 December 2025**, the President signed an executive order titled "Eliminating State Law Obstruction of National Artificial Intelligence Policy."[4] Its main elements:

- Creation of an **AI Litigation Task Force** at the Department of Justice to challenge state laws considered to obstruct national AI policy.
- Direction to the Department of Commerce to **evaluate state laws** for consistency with federal AI policy.
- Authorisation to **condition certain federal funding** on state cooperation with national AI policy.
- Direction to the FCC to develop a **federal AI disclosure standard**.
- **Carve-outs** preserving state authority over (i) child safety, (ii) datacenter infrastructure decisions, and (iii) state government procurement.

Congress declined to include AI preemption in the FY2026 National Defense Authorization Act. The preemption EO is therefore the operative federal posture; legal challenges to specific state laws under preemption theories are anticipated through 2026.

Center for AI Standards and Innovation (CAISI)

In **June 2025**, Secretary of Commerce Lutnick renamed the NIST AI Safety Institute the **Center for AI Standards and Innovation (CAISI)**.[5] The mission was reframed toward standards, national security, and competitiveness. CAISI has since signed **frontier-model testing agreements** with Google DeepMind, Microsoft, and xAI; Anthropic and OpenAI agreements pre-existed under the prior AISI brand.

CAISI continues to participate in the **International Network of AI Safety Institutes** alongside the UK AI Security Institute, Singapore, Japan, Korea, and others.

NIST AI Risk Management Framework

NIST AI RMF 1.0 (January 2023) remains the foundational US voluntary framework. It organises AI risk management into four functions:

- **Govern.** Establish organisational governance for AI risk — culture, accountability, policies.
- **Map.** Contextualise the AI system; identify what could go wrong and who is affected.
- **Measure.** Analyse, assess, and monitor risks (bias, robustness, drift, security).
- **Manage.** Mitigate and respond — controls, incident response, change management.

Figure: NIST AI RMF core functions — Govern (overarching, cross-cutting culture and policy), with Map, Measure, and Manage operating as a continuous iterative cycle.

NIST profiles and updates

NIST has not published an AI RMF 2.0. Evolution is via **profiles** (use-case overlays) and **crosswalks**:

- **NIST AI 600-1 — Generative AI Profile**, originally issued July 2024, **updated March 2025** to add threat categories for poisoning, evasion, extraction, and model manipulation.[6]
- **NIST IR 8596 — Cybersecurity Framework Profile for AI** — preliminary draft December 2025.
- **AI RMF Profile for Trustworthy AI in Critical Infrastructure** — concept note released 7 April 2026.
- **SP 800-53 AI overlay** — in development for federal cybersecurity controls applied to AI systems.

For organisations subject to federal contracts or working with critical infrastructure, the relevant profile is increasingly important; the GenAI Profile in particular is widely adopted as the de facto reference for generative-AI threat modelling.

TAKE IT DOWN Act (May 2025)

Signed **19 May 2025**, the **TAKE IT DOWN Act** is the first federal statute focused specifically on AI-adjacent harms.[7] It criminalises the knowing publication of non-consensual intimate imagery, including deepfakes, and requires online platforms to remove such content within **48 hours** of receiving a valid notice. Penalties include criminal liability for publication and FTC enforcement against non-compliant platforms.

The Act intersects with AI governance in two ways: it directly addresses one of the most prominent generative-AI harms, and it establishes the FTC as a federal enforcer of an AI-related obligation, complementing the FTC's pre-existing Section 5 unfair-or-deceptive-practices authority.

AI Diffusion Rule rescission (May 2025)

On **13 May 2025**, the Bureau of Industry and Security (BIS) rescinded the **AI Diffusion Rule** (issued at the end of the Biden administration) and replaced it with narrower advanced-IC guidance and a public warning regarding Huawei Ascend chips.[8] The practical effect was to remove the broad export-licensing requirements the Diffusion Rule had introduced for AI hardware destined for many countries, leaving in place targeted controls focused on China and a handful of other jurisdictions.

Anti-discrimination law continues to apply

On **27 January 2025**, the EEOC removed its **2023 AI hiring Technical Assistance documents**, and the OFCCP removed its **AI/EEO guidance**. **The underlying statutes — Title VII, the ADA, the ADEA — remain fully binding.**[9] The removal of AI-specific guidance does not change the legal obligations on employers using AI in hiring, promotion, or termination decisions; it removes the agency's stated interpretation of how those statutes apply to AI. Litigation under the underlying statutes continues, and several state laws (notably New York City Local Law 144 and Illinois HB 0053) impose specific AI-hiring obligations independent of federal guidance.

See US State Laws for state-level AI employment rules and Sectoral for FCRA, ECOA, and other federal regimes that continue to apply to AI used in regulated decisions.

1. The White House. (2025, January 23). Removing Barriers to American Leadership in Artificial Intelligence (EO 14179). ↩
2. Consumer Financial Protection Bureau. (2025, September 26). AI Compliance Plan for OMB M-25-21. ↩
3. The White House. (2025, July 23). America's AI Action Plan. ↩
4. The White House. (2025, December 11). Eliminating State Law Obstruction of National Artificial Intelligence Policy. ↩
5. NIST. Center for AI Standards and Innovation (CAISI). See also FedScoop. (2025, June). Trump administration rebrands AI Safety Institute as CAISI. ↩
6. NIST. AI 600-1: Generative AI Profile. ↩
7. TAKE IT DOWN Act — see overview at Wikipedia. ↩
8. Wiley. BIS Rescinds AI Diffusion Rule. ↩
9. EEOC and OFCCP removed AI-specific employment guidance on 27 January 2025; underlying Title VII / ADA / ADEA obligations remain binding. ↩

US State Laws

US states have moved faster on AI law than the federal government during 2024-2026. By May 2026, **Colorado**, **Texas**, **California**, **Utah**, **Tennessee**, **Illinois**, and **New York City** all have binding AI statutes or ordinances. The December 2025 federal preemption executive order (see US Federal) signals coming conflict, but in the absence of a federal preemption statute, state laws continue to apply.

This chapter is organised state-by-state, with effective dates and enforcement posture as of May 2026.

Colorado — SB 24-205 (Colorado AI Act)

Colorado's SB 24-205, signed in May 2024, was originally to take effect **1 February 2026**. **SB 25B-004** (signed **28 August 2025**) delayed the effective date to **30 June 2026**, leaving the substantive obligations intact.[1]

The Act applies to **developers and deployers** of "high-risk artificial intelligence systems" — AI systems that make, or are a substantial factor in making, **consequential decisions** affecting consumers in employment, education, financial services, essential government services, healthcare, housing, insurance, or legal services.

Core obligations:

- **Developers** must provide deployers with a statement describing intended uses, known risks, training-data summaries, performance metrics, mitigation measures, and information needed for impact assessments.
- **Deployers** must implement a risk-management programme, conduct annual **impact assessments**, notify consumers when a high-risk AI system is used to make a consequential decision, provide an opportunity to correct data and appeal adverse decisions, and disclose certain incidents to the Attorney General.
- A **rebuttable presumption** of compliance is available to deployers using risk-management programmes that reasonably conform to recognised frameworks (NIST AI RMF, ISO/IEC 42001).

Enforcement. The Colorado Attorney General has exclusive enforcement authority. There is no private right of action.

Texas — HB 149 (Texas Responsible AI Governance Act, TRAIGA)

Signed **22 June 2025**, effective **1 January 2026**.[2] TRAIGA prohibits specific AI uses (e.g., social scoring by government, manipulative AI targeting protected classes), imposes disclosure obligations on AI used in government services, and creates a **regulatory sandbox** for AI development.

Enforcement is by the **Attorney General**, with civil penalties of **$10,000 to $200,000** per violation. TRAIGA also includes a **right-to-cure** provision: violations cured within 60 days of notice are not subject to penalty.

California

California is the most active state on AI legislation. Four laws shape the landscape:

SB 53 – Transparency in Frontier Artificial Intelligence Act

Signed **29 September 2025**, SB 53 is the first US frontier-model safety statute.[3] It applies to "large frontier developers" (computational and revenue thresholds) and requires:

- Publishing a **frontier AI framework** describing how the developer assesses and mitigates catastrophic risks.
- Publishing a **safety case** before deploying a covered model.
- Reporting **critical incidents** to the California Office of Emergency Services.
- Whistleblower protections for employees raising safety concerns.

SB 53 deliberately mirrors the substantive structure of the EU GPAI Code of Practice's Safety & Security chapter, easing dual compliance for developers.

AB 2013 – Training Data Transparency

Effective **1 January 2026**. Requires GenAI developers to publish, on the model's product page, a **summary of the training data** including data sources, types of data, ownership/licensing, whether personal information was included, and whether copyrighted material was used.

SB 942 – California AI Transparency Act

Originally effective 1 January 2026, **postponed by AB 853 to 2 August 2026**. Requires covered GenAI providers to offer free AI-detection tools and to embed both visible and metadata-based provenance disclosures in AI-generated content. Aligned with EU AI Act Article 50(2) deadlines, allowing a single technical implementation to satisfy both regimes.

Other California laws

Additional California statutes effective during 2025-2026 cover deepfakes in elections (AB 2655, AB 2839), digital-replica rights for performers (AB 2602, AB 1836), and required disclosures in AI-generated political advertisements.

Utah – SB 226 (amending the Utah Artificial Intelligence Policy Act)

Effective **7 May 2025**. SB 226 narrows the original 2024 Utah AI Policy Act:

- **GenAI disclosure obligation** is now limited to **"high-risk AI interactions"** rather than all generative AI interactions.
- Adds a **safe harbour** for organisations whose AI usage is governed by sector-specific regulators (e.g., licensed healthcare providers).
- Narrows the definition of generative AI to clearer technical criteria.

Utah's revised framework is widely viewed as a more workable template than the original, and may inform similar laws in other states.

Tennessee – ELVIS Act

The **Ensuring Likeness Voice and Image Security Act** (ELVIS Act) has been in force since **1 July 2024**. It extends Tennessee's right of publicity to cover voice, prohibiting unauthorised AI-generated voice replicas. The Act is most relevant to entertainment and advertising; it has been cited in several voice-cloning enforcement actions during 2025.

Illinois

- **HB 3773** (effective 1 January 2026) prohibits employers from using AI that has the effect of subjecting employees to discrimination on the basis of protected classes; requires notice to applicants and employees when AI is used in employment decisions.
- **Older provisions** of the Artificial Intelligence Video Interview Act continue to apply to AI used in video interviews.

New York City – Local Law 144 (AEDT)

Local Law 144, in effect since 2023, requires employers using **automated employment decision tools (AEDTs)** to (i) conduct an independent bias audit, (ii) publish a summary of audit results, and (iii) notify candidates before use.

Enforcement status (2026): On **2 December 2025**, the **New York State Comptroller** published an audit finding that the Department of Consumer and Worker Protection (DCWP) had identified only 1 violation while auditors identified 17 — concluding DCWP enforcement had been "ineffective."[4] DCWP has committed to proactive 2026 enforcement, and several NYC employer audits are now in process.

Where state law is heading

Two patterns are visible across the 2025-2026 wave of state AI laws:

1. **Convergence on the Colorado / Texas template** — high-risk classification, deployer-and-developer split, AG enforcement, no private right of action, right-to-cure. Several additional states (Connecticut, New York at the state level, Virginia) have considered similar bills during 2025-2026.
2. **Divergent niches** — California's frontier focus (SB 53), Tennessee's voice-replica focus (ELVIS), New York City's hiring focus (LL 144), Utah's "high-risk interaction" disclosure focus — reflecting state-specific industry and political priorities.

Organisations operating across multiple states face a meaningful compliance multiplexing problem. The federal preemption EO (December 2025) signals federal intent to consolidate, but until a federal preemption statute is enacted or courts resolve preemption challenges, multi-state compliance programmes must address each state's instruments individually.

---

1. Clark Hill. Colorado's AI law delayed until June 2026 — what the latest setback means for businesses. ↩
2. Latham & Watkins. Texas Signs Responsible AI Governance Act into Law. ↩
3. Office of the Governor of California. (2025, September 29). Governor Newsom signs SB 53, advancing California's world-leading AI industry. ↩
4. New York State Comptroller. (2025, December 2). Enforcement of Local Law 144: Automated Employment Decision Tools. ↩

International Regulation

Outside the EU and US, the international AI regulatory landscape moved substantially during 2025. **South Korea** enacted a binding AI Basic Act; **Japan** passed a soft-law AI Promotion Act; **China** brought synthetic-content labelling rules into force; **Brazil** continues to advance PL 2338; the **UK** published a Blueprint for AI Regulation with a comprehensive statute signalled for late 2026; and **Canada's** Bill C-27 (which would have created the Artificial Intelligence and Data Act, AIDA) died on the order paper.

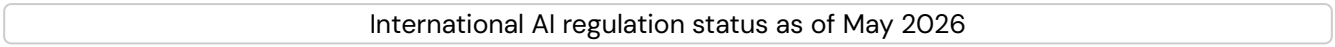

Figure: International AI regulation status as of May 2026. Enforcement maturity, not legitimacy — voluntary regimes (right column) heavily influence binding regimes (left).

Canada – AIDA effectively dead

Canada's **Bill C-27**, which contained the proposed Artificial Intelligence and Data Act (AIDA), **died on the order paper on 6 January 2025** when Parliament was prorogued.[1] Minister of Innovation, Science and Industry Solomon confirmed in **June 2025** that the bill will **not be reintroduced in its previous form**.

Canada therefore has no comprehensive horizontal AI statute. AI-related obligations continue to flow from existing federal law (PIPEDA, the Privacy Act, the Canadian Human Rights Act) and from provincial law (notably Quebec's Law 25). The federal **Directive on Automated Decision-Making** continues to apply to federal government use of AI. A new approach to AI legislation is anticipated but not currently scheduled.

United Kingdom – Blueprint published, statute pending

The UK has not yet enacted a comprehensive AI statute. On **21 October 2025**, the government published the **Blueprint for AI Regulation**, an updated policy document setting out a context-specific, sector-led approach implemented through existing regulators (Ofcom, ICO, FCA, MHRA, CMA).[2] The Blueprint also launched an **AI Growth Lab** regulatory sandbox.

A comprehensive AI Bill is signalled for **H2 2026**. Until then, AI obligations are layered: data-protection obligations under UK GDPR and the Data Protection Act 2018, online safety obligations under the Online Safety Act 2023, sector-specific obligations from individual regulators, and copyright/IP rules applied case-by-case.

The **UK AI Security Institute** (formerly AI Safety Institute) remains the focal point for frontier-model evaluation and is the lead UK participant in the International Network of AI Safety Institutes.

South Korea – AI Basic Act (effective January 2026)

Korea enacted the **Act on Promotion of the AI Industry and Framework for Establishing Trustworthy AI** ("AI Basic Act") in **January 2025**, with an **effective date of 22 January 2026**.[3] Major obligations:

- **Three regulatory tracks**: a transparency track for generative AI interactions, a frontier-safety track for "high-impact" AI, and a governance track applied broadly.
- **Extraterritorial application** to providers serving Korean users.
- A **one-year administrative-fine grace period** following the effective date, during which violations are noted but not fined.

- Disclosure obligations for AI-generated content, similar in structure to EU AI Act Article 50 but with Korean-specific notice requirements.

The Personal Information Protection Commission (PIPC) and the Ministry of Science and ICT share enforcement responsibility.

Japan – AI Promotion Act (effective June 2025)

Japan's **AI Promotion Act** passed the Diet on **28 May 2025**, with provisions effective **4 June 2025** and full operation from **September 2025**.[4] The Act takes a **soft-law approach**:

- **No penalties** for non-compliance.
- Establishes an **AI Strategy Council** at the cabinet level.
- Encourages voluntary risk management and transparency.
- Aligned with the **OECD AI Principles** and Hiroshima AI Process outputs.

Japan's approach reflects a deliberate choice for an industry-friendly framework that emphasises **promotion** over **prohibition**, leaving harder regulation to sectoral regulators and existing law (Personal Information Protection Act, Telecommunications Business Act).

China – synthetic content labelling and ongoing regulation

China's **Measures for the Labeling of AI-Generated Synthetic Content**, issued **14 March 2025**, took effect on **1 September 2025**.[5] Providers of AI services must:

- Apply **explicit labels** (visible to users) to AI-generated text, images, audio, and video.
- Apply **implicit labels** (embedded in metadata) to enable downstream identification.
- Distinguish between **labels for AI-generated content** and **labels for AI-synthesised content** (e.g., deepfakes).

These measures build on China's earlier instruments — the 2022 Internet Information Service Algorithmic Recommendation Management Provisions, the 2023 Provisions on the Administration of Deep Synthesis, and the 2023 Interim Measures for the Management of Generative AI Services — that collectively form one of the most comprehensive AI regulatory regimes globally.

Brazil – PL 2338/2023

Brazil's **PL 2338/2023** ("the Brazilian AI Bill") was **approved by the Senate in December 2024** and is **under review by the Chamber of Deputies**, with a special committee report dated **29 April 2025**.[6] The bill takes a risk-based approach broadly modelled on the EU AI Act, with high-risk categories, fundamental-rights protections, and an enforcement body. Final passage is anticipated during 2026 but not confirmed.

Singapore

Singapore's **AI Verify** framework (Infocomm Media Development Authority and the AI Verify Foundation) continues as a voluntary technical testing toolkit and governance framework. Singapore is a leading participant in the **International Network of AI Safety Institutes** and has signed bilateral AI cooperation agreements with multiple jurisdictions.

The **Model AI Governance Framework for Generative AI** (2024) remains the principal Singaporean guidance for industry.

OECD AI Principles

The **OECD AI Principles**, originally adopted in 2019, were **updated on 9 May 2024** to address generative AI and frontier-model issues.[7] The principles — inclusive growth, human-centred values, transparency, robustness, accountability — remain the lingua franca for high-level AI policy and are incorporated by reference into numerous national instruments (including Japan's AI Promotion Act).

UNESCO Recommendation on AI Ethics

The **UNESCO Recommendation on the Ethics of Artificial Intelligence** (November 2021) remains the principal multilateral ethics instrument. UNESCO continues capacity-building work with member states; the Recommendation is most influential as a reference for national ethics frameworks in jurisdictions without binding statutes.

Paris AI Action Summit and the International Network

The **Paris AI Action Summit** (10-11 February 2025) produced:

- A **Statement on Inclusive and Sustainable AI** signed by **58 countries** (the United States and the United Kingdom did **not** sign).
- An **International AI Safety Report** authored by 96 international experts.
- The launch of **Current AI**, an international public-interest foundation with approximately **$400 million** in initial commitments.
- The **ROOST** (Robust Open Online Safety Tools) initiative.

The **International Network of AI Safety Institutes** continues to coordinate frontier-model evaluation between participating institutes — UK AISI, US CAISI, Singapore, Japan AISI, Korea, France INESIA, and others — despite the AISI-to-CAISI rebrand. **India** is hosting the next major AI summit in the series.

Regional posture summary

Jurisdiction	Status (May 2026)	Approach
EU	In force	Horizontal, risk-based, prescriptive (AI Act)
US Federal	Active	Pro-innovation EOs, voluntary frameworks (NIST), sector-specific
US States	Active	Patchwork; CO, TX, CA leading
UK	Pending	Sector-led; comprehensive Bill due H2 2026
Canada	Inactive	AIDA died; existing law applies
Korea	Effective Jan 2026	Risk-based, extraterritorial, three tracks
Japan	In force	Soft-law, promotion-focused, no penalties
China	In force	Comprehensive, prescriptive, multiple instruments
Brazil	Pending	EU-style risk-based; passage anticipated 2026
Singapore	Voluntary	AI Verify, Model Framework for GenAI

1. Fasken. Prorogation's Digital Impact. ↩
2. Osborne Clarke. Regulatory Outlook January 2026: Artificial Intelligence. ↩
3. Cooley. South Korea's AI Basic Act: Overview and Key Takeaways. ↩
4. White & Case. Japan's First AI Legislation Becomes Law. ↩
5. China Law Translate. AI Labeling Measures. ↩
6. Library of Congress. Brazil Senate Advances Discussions on Bill to Regulate AI Use. ↩
7. ANSI. (2024, May 9). OECD Updates AI Principles. ↩

Sectoral Regulation

Beyond horizontal AI law, sector regulators have moved on AI throughout 2025–2026 — particularly in healthcare, financial services, employment, and consumer finance. This chapter surveys the most consequential US sectoral developments. See International for non-US sector regulation and EU AI Act for the European high-risk system regime (which itself functions as sectoral overlay in healthcare, employment, justice, and other domains).

Healthcare — FDA AI/ML medical devices

The Food and Drug Administration's principal AI/ML regulatory instrument is the **Predetermined Change Control Plan (PCCP) Final Guidance**, issued in **December 2024**.[1] The Final Guidance operationalises the FDA's "AI/ML Software as a Medical Device Action Plan" by allowing sponsors to plan, in advance, the modifications an AI/ML-enabled device may undergo without triggering a new premarket submission. A PCCP must include:

- **Description of Modifications** — what changes the sponsor plans to make.
- **Modification Protocol** — how those changes will be validated.
- **Impact Assessment** — expected impact on safety, effectiveness, and overall device performance.

The Final Guidance is significant because it gives AI/ML medical-device sponsors a structured pathway for post-market model updates, addressing one of the longest-standing tensions in AI medical device regulation. **Good Machine Learning Practice** principles, jointly published by FDA, Health Canada, and the UK MHRA, continue to apply.

For non-device clinical AI (e.g., clinical decision-support tools that fall outside FDA's device definitions), HHS Office for Civil Rights guidance and 42 CFR Part 92 nondiscrimination rules apply.

Financial services

Three federal regulators — the **OCC**, **Federal Reserve**, and **FDIC** — jointly oversee bank AI use. Foundational guidance:

- **SR 11-7 / OCC Bulletin 2011-12** — the Federal Reserve and OCC joint guidance on **Model Risk Management** (MRM), originally 2011, remains operative and is the foundational supervisory expectation for AI/ML models used in regulated financial decisions.
- **OCC Bulletin 2025-26** — issued during 2025, provides clarifications for community banks on MRM expectations as AI tools become more accessible to smaller institutions.
- **OCC Bulletin 2026-13** — revised MRM guidance reflecting evolving practice and AI-specific considerations.
- A **joint OCC / Federal Reserve / FDIC RFI on AI/MRM** is in the supervisory pipeline as of mid-2026.

For consumer-facing AI in financial services, the **CFPB published an AI Compliance Plan** on **26 September 2025** detailing its implementation of OMB M-25-21 and its supervisory approach to bank and non-bank use of AI in lending, servicing, and collections.[2]

Fair lending statutes — ECOA, the Fair Housing Act, and Regulation B — continue to apply to AI used in credit decisions. The **Fair Credit Reporting Act** (FCRA) applies to AI used in consumer-report-based decisioning.

Employment

Federal employment law (Title VII, ADA, ADEA, GINA) applies to AI used in hiring, promotion, and termination. Although the EEOC and OFCCP withdrew their AI-specific Technical Assistance documents on **27 January 2025**, the underlying statutes are unchanged.

State and local laws fill the federal guidance gap — see US State Laws for **NYC Local Law 144**, **Illinois HB 3773**, and **Colorado SB 24-205**'s employment coverage.

Practical compliance for AI in employment:

- Conduct **bias audits** under NYC LL 144 if hiring in New York City; consider expanding to all hiring jurisdictions for defensive purposes.
- Document **validation studies** for selection procedures under the Uniform Guidelines on Employee Selection Procedures (UGESP).
- Provide **accommodation pathways** under the ADA for applicants who cannot use AI-mediated screening tools.
- Disclose AI use to candidates where required by state law.

Consumer protection — FTC

The **Federal Trade Commission** continues to enforce Section 5 of the FTC Act against unfair or deceptive AI-related practices. Notable FTC themes during 2024-2026:

- **AI claims** that are false, misleading, or unsubstantiated.
- **Algorithmic discrimination** in violation of the FTC Act or other consumer protection statutes.
- **Privacy violations** via AI-mediated data collection or model training.
- **TAKE IT DOWN Act** enforcement against platforms failing to remove non-consensual intimate imagery within 48 hours (see US Federal).

The FTC's authority is broad and post-hoc; structured compliance with NIST AI RMF and AI-specific Section 5 expectations (truthful claims, evidence base for performance, fairness review where decisions affect consumers) is the most effective preventive posture.

Telecommunications

The **December 2025 preemption executive order** directs the FCC to develop a **federal AI disclosure standard** (see US Federal). The FCC has also enforced existing rules against AI-generated robocalls, including the February 2024 declaratory ruling that AI-generated voices in calls constitute "artificial or prerecorded voices" under the Telephone Consumer Protection Act.

Critical infrastructure

NIST's **AI RMF Profile for Trustworthy AI in Critical Infrastructure** (concept note **7 April 2026**) is the developing reference for AI used in critical infrastructure sectors covered by Presidential Policy Directive 21. Sector-specific cybersecurity rules (NERC CIP for electric grid, TSA security directives for pipelines) increasingly include AI-relevant provisions.

Other sector overlays

- **Education** — Department of Education guidance (2023, updated 2024) on AI in K-12; FERPA continues to apply.

- **Transportation** — NHTSA Standing General Order on AI-equipped vehicles; FMCSA guidance on AI in commercial trucking.
- **Defence** — DoD Directive 3000.09 on Autonomy in Weapon Systems; CDAO governance for DoD AI use.
- **Insurance** — NAIC Model Bulletin on the Use of AI by Insurers; state-by-state adoption.

How to navigate sectoral overlap

Most organisations deploying AI face **multiple overlapping regimes**: a horizontal regime (EU AI Act or US state law), a sector overlay (FDA, OCC, EEOC), and broad consumer protection (FTC, AG). Best practice is to:

1. **Map** each AI system to all applicable regimes during design.
2. **Document** compliance evidence in a single technical file usable across regimes (a 42001-aligned management system enables this).
3. **Track** regulatory developments per sector via a designated owner.
4. **Engage early** with sector regulators when an AI system substantively changes a regulated process — particularly in healthcare and financial services where conformity assessments are slow.

---

1. FDA. Predetermined Change Control Plan for AI-Enabled Device Software Functions (Final Guidance, December 2024). ↩
2. Consumer Financial Protection Bureau. (2025, September 26). AI Compliance Plan for OMB M-25-21. ↩

Copyright & IP

AI training-data copyright is the most economically significant legal question facing AI developers in 2026. The largest US copyright settlement in history — Bartz v. Anthropic for **$1.5 billion** — resolved one major case in September 2025, but the underlying legal questions remain unsettled across jurisdictions. This chapter surveys the cases, the statutory developments (notably California AB 2013 and the EU AI Act's copyright requirements), and the practical implications for AI governance.

Bartz v. Anthropic — the landmark fair-use ruling and settlement

On **23 June 2025**, Judge William Alsup of the Northern District of California issued a summary-judgment ruling in Bartz v. Anthropic that became the most influential US training-data decision to date.[1] Two distinct holdings:

1. **Training on legally acquired books** — including books Anthropic had purchased or licensed — constituted **transformative fair use** under 17 U.S.C. § 107.
2. **Training on pirated copies** — specifically the LibGen and PiLiMi datasets that Anthropic had used — was **not fair use** because the underlying acquisition was infringing.

The class was certified in **August 2025**. On **5 September 2025**, Anthropic announced a **$1.5 billion settlement** — approximately **$3,000 per book** for approximately **482,460 works**.[2] The settlement is the largest in US copyright history.

Practical lessons from Bartz:

- **Acquisition matters.** A "fair use" defence is materially weaker when the underlying acquisition is unlawful. Documentary evidence that training corpora were licensed or otherwise lawfully obtained is now a baseline expectation.
- **Transformativeness still holds for legal acquisition.** The transformative-use analysis remains favourable to AI training; Bartz did not narrow the prior line of cases (e.g., Authors Guild v. Google on book scanning).
- **Settlement, not Supreme Court.** Because Bartz settled, it does not establish binding precedent at the appellate level. Other district courts may follow it; appellate courts have not yet weighed in.

Kadrey v. Meta

On **25 June 2025**, Judge Vince Chhabria ruled in Kadrey v. Meta — a parallel case against Meta's Llama training data.[3] Meta prevailed, but on **procedural grounds** rather than a broad fair-use holding. The court did not endorse Meta's fair-use position; it concluded that the plaintiffs had not produced the evidence needed to defeat Meta's motion. The ruling therefore does **not** establish a broad fair-use precedent for training on pirated copies; readers should not interpret Kadrey as overturning Bartz.

Other ongoing US cases (status May 2026)

- **The New York Times v. OpenAI**, S.D.N.Y. — ongoing; discovery extended through 2026.
- **Andersen v. Stability AI**, N.D. Cal. — image generation case; partial dismissal earlier, claims continue.
- **Getty Images v. Stability AI**, S.D.N.Y. and UK High Court — parallel actions in two jurisdictions; UK trial commenced 2025.
- **Concord Music v. Anthropic**, M.D. Tenn. — music-lyrics case, distinct from Bartz.

- Multiple class actions against **OpenAI**, **Microsoft**, **Google**, and **Meta** at various procedural stages.

EU AI Act copyright obligations

The EU AI Act addresses training-data copyright in two ways:

1. **Article 53(1)©** requires GPAI providers to publish a **sufficiently detailed summary of the content used for training**.
2. **Article 53(1)(d)** requires GPAI providers to put in place a **policy to comply with Union law on copyright and related rights**, in particular to identify and respect text-and-data-mining (TDM) opt-outs reserved under **Article 4(3) of Directive (EU) 2019/790** (the CDSM Directive).

The **EU GPAI Code of Practice** (July 2025) Copyright chapter operationalises these obligations — see Frontier Models for the Code's structure. Practically, the Code requires:

- A documented copyright policy.
- A mechanism for honouring TDM reservations (e.g., robots.txt-based, ai.txt, content provenance signals).
- A point of contact for rightholders.
- Public-facing transparency about copyright compliance.

US state-level training-data laws

- **California AB 2013** (effective **1 January 2026**) — requires GenAI developers to publish a summary of training-data sources, types, ownership, presence of personal information, and presence of copyrighted material.
- Several states have considered similar transparency bills during 2025-2026; AB 2013 is the most concrete model in force.

What's still unsettled

Several core questions remain open as of May 2026:

- **Memorisation and output infringement.** When a model can reproduce significant portions of its training data on demand, courts are wrestling with whether that constitutes direct copying. The Supreme Court has not addressed this; results in district courts have varied.
- **Derivative-works analysis for outputs.** When a generated output resembles a copyrighted work, the analysis (substantial similarity, access) is unsettled in the AI context.
- **DMCA Section 1202(b)** — removal of copyright management information — is increasingly tested in training-data cases.
- **International jurisdiction.** Where models are trained in one jurisdiction and deployed in another with different copyright rules, conflict-of-laws questions arise.
- **AI-generated works' copyrightability.** The US Copyright Office position (most recently the **March 2025 Part 2 Report**) is that purely AI-generated works are not copyrightable; the precise threshold of human contribution required is being litigated.

Practical AI governance steps on copyright

Regardless of how the unsettled questions resolve, the following baseline practices are increasingly expected:

1. **Catalogue training data** with provenance, licensing, and acquisition method documented per dataset.

2. **Avoid known pirated corpora** (LibGen, PiLiMi, Anna's Archive, sci-hub) for any production model. The Bartz settlement makes this unambiguous from an enforcement-cost perspective.
3. **Honour TDM opt-outs** (robots.txt, ai.txt, Spawning) and publish a clear opt-out endpoint.
4. **Publish training-data summaries** to AB 2013 / EU AI Act Article 53 standard.
5. **Maintain a rightholder contact** for copyright concerns.
6. **Document a copyright policy** as part of the model card.
7. **Implement output filters** for known copyrighted content where commercially feasible.
8. **Retain counsel** for jurisdiction-specific analysis of any deployed model trained on contested data.

---

1. ArentFox Schiff. Landmark Ruling on AI Copyright: Fair Use vs. Infringement in Bartz v. Anthropic. ↩
2. Authors Guild. What Authors Need to Know About the Anthropic Settlement. ↩
3. Kadrey et al. v. Meta Platforms, Inc., N.D. Cal. (2025). ↩

Privacy, Data Governance, and Security

Privacy and data governance are critical pillars of AI governance because AI systems consume and generate large volumes of data, often including personal and sensitive information. Compliance with privacy laws — GDPR in the EU, CCPA/CPRA and state-level laws in the US, PIPEDA in Canada, PIPA in Korea — is the starting point, but a mature AI governance programme also addresses data quality, model security, third-party risk, and incident response specifically calibrated to AI.

Foundational privacy law

The **EU General Data Protection Regulation (GDPR)** mandates principles — lawfulness, fairness, transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity and confidentiality, accountability — that apply directly to AI training and inference. **Article 22 GDPR** restricts solely-automated decisions producing legal or similarly significant effects, requiring human review and meaningful information about the logic involved.[1] Fines reach 4% of global turnover.

In the United States, **CCPA/CPRA** gives California residents rights to access, correct, delete, opt-out of sale or sharing, and limit use of sensitive personal information.[2] Other states (Virginia VCDPA, Connecticut CTDPA, Colorado CPA, Utah UCPA, Texas TDPSA, Oregon OCPA, Delaware DPDPA, Iowa ICDPA, Tennessee TIPA, Montana MCDPA, Florida FDBR, New Jersey NJDPA, New Hampshire NHPA) have comparable statutes with varying coverage and enforcement; multi-state operations should track Westin Research or IAPP trackers to keep current.

Sector-specific privacy regimes (HIPAA for healthcare, GLBA for financial services, FERPA for education, COPPA for children) overlay these horizontal laws and apply directly to AI used in those sectors.

International privacy laws relevant to AI include Brazil's LGPD, Japan's APPI, Singapore's PDPA, India's DPDPA (effective 2024-2025 in phases), and Korea's PIPA. Many include provisions for automated decision-making analogous to GDPR Article 22.

Data quality and lineage

AI outcomes are only as good as the data they are trained on. Mature data governance for AI requires:

- **Provenance tracking** — for each dataset, document source, acquisition method, licensing, and any consent obtained.
- **Quality assessment** — measure completeness, accuracy, freshness, representativeness; document known limitations.
- **Datasheets for datasets** — the practice proposed by Gebru et al. (2018) of cataloguing motivation, composition, collection process, preprocessing, uses, distribution, and maintenance per dataset.[3] Increasingly required by the EU AI Act Article 10 data governance obligations.
- **Schema and version control** — treat training data with the same rigour as code.

For models subject to **California AB 2013** or **EU AI Act Article 53**, a published training-data summary is now mandatory — see Copyright & IP.

Data minimisation and access control

AI systems should use the **minimum data necessary** for their purpose. Personal data not needed should not be collected; data that is needed should be pseudonymised, encrypted, and access-controlled. Common patterns:

- **Role-based access controls** restricting training-data access to authorised personnel and processes.
- **Data tokenisation** for training where individual records are not required.
- **Differential privacy** during training to provide mathematical guarantees against record-level disclosure.
- **Output-side restrictions** to prevent models from echoing memorised personal data — particularly important for large language models trained on web-scale data.

Privacy-enhancing technologies (PETs)

PETs are no longer experimental. By 2026 several are production-grade:

- **Differential privacy** — mathematical noise addition during training (DP-SGD) or output (DP-aggregation) that bounds the influence of any individual record.
- **Federated learning** — models trained across distributed datasets without centralising the data; widely deployed in healthcare, mobile keyboards, and cross-institutional research.
- **Homomorphic encryption** — computation on encrypted data; performance has improved but still constrains practical use to specific inference workloads.
- **Secure multi-party computation (MPC)** — joint computation without revealing inputs.
- **Trusted execution environments (TEEs)** — hardware-isolated computation environments (Intel SGX, AMD SEV, Apple Private Compute Cloud); now widely adopted for inference on sensitive data.

PETs increasingly appear in regulatory expectations — the EU AI Act Article 10 references them implicitly, and US sector regulators cite them as appropriate safeguards.

Retention and purpose limitation

Data governance policies should define **retention schedules** and **purpose-limitation controls** for training and inference data. GDPR requires data not be kept longer than necessary; CCPA permits consumer-initiated deletion. Practically:

- **Document retention windows** per dataset; automate deletion at end-of-window.
- **Re-purposing review** — if a dataset is reused for a new model or use case, evaluate consent and purpose limitation before training.
- **Right-to-deletion engineering** — build the ability to remove specific records from training corpora and either retrain or apply targeted machine-unlearning techniques.

Model security

AI models themselves are attack targets. Threat categories include:

- **Model extraction** — adversaries query the model to reconstruct it or its training data.
- **Adversarial examples** — inputs crafted to cause misclassification.
- **Data poisoning** — tampering with training data to embed backdoors or biases.
- **Prompt injection** — manipulating LLM behaviour through crafted input, particularly via untrusted data in retrieval pipelines.
- **Model evasion** — bypassing safety filters or content controls.

The **NIST AI 600-1 GenAI Profile** (updated March 2025) added explicit threat categories for poisoning, evasion, extraction, and model manipulation — this update is the most current US reference for GenAI threat modelling.[4]

Defensive measures:

- **Adversarial training** — train on adversarial examples to harden the model.
- **Input filtering** and **output filtering** — structural controls in the deployment pipeline.
- **Rate limiting and behavioural monitoring** — detect extraction and reconnaissance attempts.
- **Red-teaming** — systematic adversarial evaluation, increasingly required by frontier-model frameworks (see Frontier Models).
- **Provenance verification** of upstream components (base models, weights, fine-tuning datasets).

Data security

AI data pipelines must be secured with general infosec hygiene plus AI-specific considerations:

- **Encryption** in transit and at rest for training data and model weights.
- **Identity and access management** for systems handling training data and models.
- **Integrity verification** of training datasets (cryptographic hashing, version control).
- **Poisoning detection** — statistical anomaly detection in training data.
- **Backup and recovery** for model weights and training pipelines as critical assets.

Third-party and supply chain risks

Most AI systems incorporate third-party components: pre-trained foundation models, open-source libraries, cloud AI services, vector databases, datasets. Governance must extend to these:

- **Vendor assessment** for security, privacy, and AI-specific compliance posture.
- **Model provenance** — document source, version, and license of every model component.
- **License compliance** for open-source models with restrictive terms.
- **Contractual allocation** of responsibility for incident response, data handling, and regulatory cooperation.
- **Cross-border data transfer** mechanisms (Standard Contractual Clauses, adequacy decisions) where data crosses jurisdictions.

EU AI Act Article 25 explicitly addresses provider obligations through the value chain for high-risk systems.

Incident response

AI incident response should be a dedicated discipline within general incident response, with playbooks for:

- **Safety incidents** — AI causes physical or financial harm.
- **Bias incidents** — AI is found to produce discriminatory outcomes.
- **Privacy incidents** — AI reveals or is alleged to reveal personal data.
- **Security incidents** — AI is compromised or used to attack other systems.
- **Misuse incidents** — AI is used by adversaries for harmful purposes.

Mandatory incident reporting now applies in multiple regimes — EU AI Act Article 55 (serious incidents for systemic-risk GPAI), California SB 53 (critical incidents), Korea AI Basic Act, sector regulators (FDA for

medical devices, OCC for banks). Map your reporting obligations early; many regimes have **short windows** (e.g., 15 days for serious incidents under the EU AI Act).

Audits and red-teaming

Independent audits are increasingly expected:

- **Bias audits** — required by NYC LL 144 for hiring AI; emerging as best practice broadly.
- **Privacy audits** — required by sector privacy regulators; increasingly required by procurement.
- **Security audits and red-teaming** — required for frontier models by EU and emerging US frameworks.
- **42001 certification audits** — available since 2025 with the publication of ISO/IEC 42006 (see ISO standards).

Coordination across functions

A robust AI governance programme coordinates **privacy, security, data governance, AI/ML engineering, legal, compliance, and product** functions. Many organisations establish an **AI Governance Council** with representation from each function, supported by an **AI Governance Office** that owns documentation, audits, and regulatory engagement. ISO/IEC 42001 specifies this coordination at the management-system level.

---

1. GDPR Info. Art. 22 GDPR — Automated individual decision-making. ↩
2. Cloudflare. What is the CCPA?. ↩
3. Gebru, T., et al. (2021). Datasheets for Datasets. Communications of the ACM, 64(12). ↩
4. NIST. AI 600-1: Generative AI Profile. ↩

Technical Aspects of AI Safety

AI governance is not just policy. It also requires concrete technical practices to ensure AI systems are safe, reliable, and aligned with intended goals. This chapter covers the engineering disciplines that AI practitioners and risk managers apply as part of governance, with updates for the 2025-2026 state of practice.

Robustness and reliability

Robustness begins with rigorous validation on test data that simulates real-world variability and edge cases. Practical techniques:

- **Stress testing** — computer vision systems tested in varying lighting and noise; text systems tested on out-of-distribution inputs and adversarial paraphrases.
- **Adversarial training** — training on adversarial examples to harden the model.
- **Ensemble methods** — multiple models or sensors compensate for any single component's failure.
- **Operating-domain documentation** — explicitly state what conditions the model was designed for; reject or flag inputs outside the domain.
- **Capability evaluations** — for frontier models, structured pre-deployment evaluations of dangerous capabilities (cyber, CBRN, autonomy). See Frontier Models for the structural pattern.

The **NIST AI 600-1 Generative AI Profile** (updated March 2025) provides the most current US reference for GenAI threat categories and evaluation patterns.

Alignment with human values

Alignment is the engineering discipline of ensuring AI behaviour matches human intentions and values. Foundational techniques in production by 2026:

- **Reinforcement learning from human feedback (RLHF)** — widely deployed for instruction-following and value alignment in large language models.
- **Constitutional AI / RLAIF** — AI-assisted critique and revision against an explicit set of principles, reducing reliance on human raters at scale.
- **Direct preference optimisation (DPO)** and successors — offline preference learning that avoids the reinforcement-learning loop.
- **Specification through constraints** — encoding behavioural rules directly (e.g., refuse-list rules, scoped tool access).
- **Tool-use scoping** — for agentic systems, explicit permissions and human-in-the-loop checkpoints for high-impact actions.

For applications beyond instruction-tuned LLMs, alignment thinking still applies: define objective functions carefully (not just optimise the metric, but the metric subject to fairness and safety constraints), and peer-review model objectives during development.

For agentic AI systems — AI that takes actions in the world — alignment also requires **controllability**: structured ability to interrupt, audit, and roll back AI-initiated changes. This is an active research area; see the International AI Safety Report (Paris, February 2025) for state-of-the-art.

Interpretability and transparency tools

Interpretability sheds light on "black box" models:

- **SHAP and LIME** — per-prediction feature attribution; still standard for tabular and structured models.
- **Saliency maps and integrated gradients** — per-input attribution for vision models.
- **Attention visualisation** — standard for transformer-based models, though attention is not a complete causal account.
- **Probing** — lightweight classifiers attached to model internals to test for specific learned representations.
- **Sparse autoencoders and circuit analysis** — mechanistic interpretability techniques that have advanced substantially during 2024-2026 and now produce human-readable feature dictionaries for parts of large models.
- **Model cards** — documentation in plain language describing intended use, performance, and limitations. [1]

For high-impact AI systems, an **explanation report or model card** is increasingly mandatory:

- EU AI Act Article 11-13 (technical documentation, transparency, instructions for use).
- GDPR Article 22 (meaningful information about logic in automated decisions).
- US sector rules (FCRA adverse action notices, ECOA reason codes).

Bias and fairness mitigation

Measuring and mitigating bias is a discipline with multiple technical entry points:

- **Pre-processing** — ensure training data is balanced or reweighted.
- **In-processing** — add fairness penalty terms or constraints to the objective.
- **Post-processing** — adjust model outputs to reduce disparity.

Open-source toolkits (IBM AI Fairness 360, Microsoft Fairlearn, Google Model Card Toolkit) operationalise these techniques. AI governance can mandate a **fairness check** before deployment for specified system categories, with documented metrics, thresholds, and remediation actions.

Specific legal regimes now codify fairness expectations:

- **NYC Local Law 144** — bias audits for automated employment decision tools (see US State Laws).
- **EU AI Act Article 10** — data governance requirements for high-risk systems including measures to detect and prevent bias.
- **EU AI Act Article 27** — fundamental rights impact assessments for high-risk deployers.

Engineering teams must produce metrics and evidence; compliance must ensure those satisfy legal thresholds. **Joint development** of fairness definitions between engineering, legal, and product teams is essential because the technical definitions (demographic parity, equalised odds, predictive parity) embed normative choices and cannot all be satisfied simultaneously.

Performance monitoring and drift detection

Deployment is not a one-time event. Governance requires ongoing monitoring:

- **Performance metrics** in production: accuracy, calibration, fairness metrics across subgroups, business metrics impacted by the AI.
- **Drift detection** — distribution shift in inputs (covariate drift), in labels (label drift), or in the input-output relationship (concept drift).

- **Out-of-distribution detection** — identify inputs unlike training data; route to human review or fall back gracefully.
- **Hallucination detection and grounding** for LLM applications — particularly retrieval-augmented systems.
- **Misuse detection** — identify abuse patterns and adversarial attempts.

When monitoring signals a problem, governance triggers a **defined response process** (NIST RMF's "Manage" function): retrain, roll back, restrict, escalate. Document each incident and decision; feed lessons learned back into design.

Safety constraints and testing in simulation

For AI interacting with the physical world (robots, autonomous vehicles, medical devices), pre-deployment safety testing is essential:

- **Simulation** — billions of simulated miles or interactions to explore rare and dangerous scenarios.
- **Formal verification** — mathematical proof that certain safety properties hold under specified assumptions; viable for narrow modules but not yet for full neural-network policies.
- **Redundancy and fail-safes** — engineering patterns from aviation and industrial control extended to AI systems.
- **Guardian systems** — an independent monitor that can override or shut down the main AI if it detects unsafe behaviour.

For non-physical AI, the analogous practice is **defence-in-depth**: pre-deployment red-teaming, runtime guardrails, monitoring, incident response. The frontier-model safety architecture in Frontier Models is the most advanced expression of this pattern.

Agentic systems — the 2026 frontier

The biggest engineering shift between 2024 and 2026 is the rise of **agentic AI** — systems that take multi-step actions in the world via tool use, browsing, code execution, or robotic control. Governance and safety implications:

- **Permission scoping** for tools and APIs the agent can call.
- **Human-in-the-loop checkpoints** for high-impact actions.
- **Audit trails** capturing every tool call and decision rationale.
- **Sandboxing** of code execution and environment access.
- **Rate limiting** and **anomaly detection** at the action layer.
- **Reversibility analysis** — categorise actions by reversibility; require stronger controls for irreversible actions.
- **Identity and authorisation** — the agent acts on behalf of a principal; both internal authorisation systems and emerging external standards (e.g., AI agent identity proposals) increasingly matter.

Best practice for agentic systems is still rapidly evolving. The Frontier Models chapter discusses the structural requirements emerging in the EU GPAI Code, CAISI agreements, and Korea's AI Basic Act.

Verification, validation, and assurance

Technical safety culminates in **assurance** — the structured argument and evidence that a system is acceptably safe. The growing convergence around **safety cases** as the unit of assurance is visible in California SB 53, the EU GPAI Code, and several RSPs. A safety case typically includes:

1. The **claim** (e.g., "this model can be deployed for use case X with acceptable residual risk").
2. The **evidence** — evaluations, audits, monitoring data, third-party assessments.
3. The **argument** linking evidence to claim — explicit reasoning about how the evidence supports the claim.
4. The **counterarguments** — what could undermine the claim and how those risks are addressed.

Engineering and governance functions both contribute to safety cases. Producing them well requires close coordination — another reason ISO/IEC 42001's management-system architecture is foundational rather than optional.

---

1. Mitchell, M., et al. (2019). Model cards for model reporting. ACM FAccT '19. ↩

Frontier Models

The most capable AI systems — "frontier models" — are governed by a distinct set of frameworks that overlap with, but go beyond, general AI governance. The EU AI Act's GPAI obligations, the GPAI Code of Practice, the CAISI testing agreements, California SB 53, Korea's frontier-safety track, and the International Network of AI Safety Institutes all target this category. This chapter consolidates frontier-model governance in one place because the obligations are otherwise scattered across multiple jurisdictional chapters.

Figure: Frontier-model governance is layered — binding regulation at the top, standards and benchmarks operationalising it, voluntary government frameworks bridging to industry, and developer commitments at the base. The AI Safety Institute network coordinates evaluation across all layers.

What counts as a "frontier model"?

There is no universal definition, but converging criteria include:

- **Compute** — training compute exceeding ~10^{25} FLOPs (the EU AI Act systemic-risk threshold under Article 51(2)). This threshold can be adjusted by the AI Office.
- **Capability** — ability to perform a wide range of distinct tasks, particularly tasks plausibly relevant to severe harm (CBRN, cybersecurity, autonomous replication, deception).
- **Designation** — the EU AI Office can designate a model as posing systemic risk regardless of compute.

As of May 2026, models commonly understood as frontier include OpenAI's GPT-5 (released **7 August 2025**), Google's Gemini 3 (released **18 November 2025**), and Anthropic's Claude Opus 4.5 (released **24 November 2025**), along with subsequent Q1-Q2 2026 revisions of each.

EU GPAI obligations (in force August 2025)

Under Articles 53-55 of the EU AI Act, providers of **general-purpose AI models** must:

- Maintain **technical documentation** of the model (training, evaluation, capabilities, limitations).
- Provide **downstream documentation** to providers integrating the model into AI systems.
- Comply with **Union copyright law**, including TDM opt-outs.
- Publish a **summary of training content**.
- For models with **systemic risk** (10^{25} FLOPs or AI Office designation): conduct adversarial testing, track and report serious incidents, ensure adequate cybersecurity.

The **GPAI Code of Practice** (10 July 2025) provides a structured way to demonstrate compliance — see EU AI Act for governance structure. Code signatories as of August 2025: Google, Microsoft, OpenAI, Anthropic. Meta declined.

CAISI testing agreements (US, 2025-2026)

The **Center for AI Standards and Innovation** (formerly NIST AI Safety Institute) has signed bilateral pre-deployment and post-deployment evaluation agreements with frontier developers:[1]

- **Google DeepMind** — 2025 agreement covering Gemini frontier models.
- **Microsoft** — 2025 agreement.
- **xAI** — 2025 agreement covering Grok models.

- **Anthropic** — pre-existing agreement (signed under the prior AISI brand).
- **OpenAI** — pre-existing agreement.

Testing covers cybersecurity, CBRN risk, and a small number of designated dual-use capability categories. The agreements are voluntary; CAISI does not have statutory authority to compel testing of frontier models, but agreements provide structured access for evaluation.

California SB 53 — Transparency in Frontier AI Act

Signed **29 September 2025** and discussed in detail under US State Laws, SB 53 requires large frontier developers to:

- Publish a **frontier AI framework** describing risk-assessment and mitigation methodology.
- Publish a **safety case** before deploying a covered model.
- Report **critical incidents** to the California Office of Emergency Services.
- Provide whistleblower protections for safety-related employee disclosures.

SB 53's substantive structure mirrors the GPAI Code of Practice's Safety & Security chapter, simplifying dual compliance.

Korea AI Basic Act — frontier safety track

Korea's AI Basic Act (effective **22 January 2026**) includes a **frontier-safety track** applicable to "high-impact AI" — broadly aligned with frontier-model concepts. Obligations include safety frameworks, pre-deployment evaluation, and incident reporting to Korean authorities. Extraterritorial application means providers serving Korean users are within scope regardless of where models are developed.

Voluntary frameworks and the AI Safety Institute network

Several voluntary mechanisms supplement statutory obligations:

- **Responsible Scaling Policies (RSPs)** — published by Anthropic and (in different forms) by other developers; commit to specific capability evaluations and risk thresholds.
- **Frontier Model Forum** — industry consortium (Anthropic, Google, Microsoft, OpenAI) focused on frontier-model safety research and best-practice sharing.
- **MLCommons AI Safety v1.0** — benchmark suite for evaluating safety properties of frontier models.
- **International AI Safety Report** (Paris AI Action Summit, February 2025) — 96-expert consensus document on frontier AI risks.

The **International Network of AI Safety Institutes** — including UK AISI, US CAISI, Singapore, Japan AISI, Korea, France INESIA, and others — conducts coordinated evaluations and shares findings through agreed protocols. India is hosting the next major summit in the series.

Common frontier-safety architecture

Across the EU GPAI Code, SB 53, the CAISI agreements, and most RSPs, a common architecture is converging:

1. **Capability evaluations** at defined thresholds — before initial deployment, before scaling, after substantial updates.
2. **Risk identification and mitigation** — with documented thresholds at which mitigation actions trigger.

3. **Safety case** — structured argument that residual risk is acceptable, addressed to a defined audience (internal governance, regulator, public).
4. **Pre-deployment testing** — either internal or in cooperation with AI Safety Institutes.
5. **Post-deployment monitoring** — for misuse, incidents, capability change.
6. **Incident reporting** — to regulators (EU AI Office, Cal OES, Korean MSIT) and, where applicable, to the Frontier Model Forum or peer institutions.
7. **Transparency** — published framework, safety case, model card.

Recommended posture for frontier developers (May 2026)

If you are developing frontier models, the minimum credible posture includes:

- **EU GPAI Code of Practice signature** or equivalent compliance demonstrated in technical documentation.
- **CAISI evaluation agreement** if operating in the US.
- **California SB 53 compliance** if making large frontier models available to California users.
- **Korea AI Basic Act compliance** if serving Korean users (mind the extraterritorial scope).
- **Published frontier AI framework / RSP** with specified capability thresholds and mitigation actions.
- **Safety case** for each major release.
- **ISO/IEC 42001** management system as foundation.
- **Incident reporting** procedures consistent with all applicable regimes.

For deployers and downstream integrators

If you integrate frontier models rather than develop them, the governance focus is:

1. **Verify GPAI compliance** of the upstream model (Code signatory, technical documentation available).
2. **Receive and act on** downstream documentation provided under Article 53(1)(b) EU AI Act.
3. **Implement use-case-specific risk management** — the frontier-safety framework of the upstream developer does not substitute for your own risk assessment of the deployed application.
4. **Maintain provenance** — document which model version is in production, what changed when, and your validation of those changes.
5. **Monitor for incidents** in your deployment and feed back to the upstream developer.

---

1. NIST. Center for AI Standards and Innovation (CAISI). ↩

Audience-Specific Guidance

Different roles have different parts to play in AI governance. This chapter offers tailored guidance for four primary audiences: **AI Practitioners**, **Compliance Officers**, **Executives & Board Members**, and **Policymakers & Regulators**.

AI Practitioners

Data scientists, ML engineers, AI developers

Focus

Practitioners are at the front lines of building and deploying AI. The job is to **operationalise** governance and safety inside the development process: build models that perform well on accuracy metrics and meet criteria for fairness, explainability, robustness, and compliance.

Specific practices

- **Translate principles into code.** Ethics guidelines mean nothing if they don't show up as concrete model-validation steps, bias checks, or model-card sections.
- **Handle data lawfully.** Anonymise where required, obtain proper consent, respect opt-outs, work with privacy reviewers early.
- **Test exhaustively.** Train/test splits are not enough — add stress tests, adversarial tests, fairness checks, and (for GenAI) red-teaming.
- **Maintain a model inventory.** Document each model's purpose, training data, version, evaluations, deployment context. This is now table stakes for EU AI Act and ISO/IEC 42001 compliance.
- **Monitor in production.** Set up performance dashboards and drift alerts. Plan retraining cadence.
- **Partner with compliance early.** When an AI tool is going to fall under EU AI Act high-risk, NYC LL 144, or Colorado SB 24-205, finding out before deployment is much cheaper than after.
- **Cultivate a safety culture.** Ethical AI is everyone's job, like security. Speak up when something looks wrong; build feedback into development culture, not just review gates.

What changed for practitioners in 2025-2026

- **Frontier development frameworks** (RSPs, EU GPAI Code, SB 53) now require pre-deployment evaluations and safety cases for the largest models — see Frontier Models.
- **Training-data transparency** (California AB 2013, EU AI Act Article 53) requires published training-data summaries.

- **Agentic systems** have a much heavier governance burden than traditional ML — permission scoping, audit trails, reversibility analysis.

Compliance Officers

Legal, regulatory, ethics, and risk personnel

Focus

Compliance officers ensure AI systems and processes adhere to law, regulation, and policy. They translate regulatory requirements into controls, guide AI projects, and verify controls are working.

Specific practices

- **Track the regulatory landscape.** EU AI Act (and Omnibus rebase), US federal EOs, state laws (CO, TX, CA, UT, IL, NYC), Korea AI Basic Act, Japan AI Promotion Act, China labelling rules, sector regulators. See Legal & Regulatory.
- **Develop internal policies.** AI governance policy, model-development standards, third-party AI usage policy, AI procurement standards.
- **Run training and awareness.** AI ethics training for developers, product managers, executives. Workshops on EU AI Act high-risk classification and Colorado AI Act consequential-decision tests.
- **Review and audit.** DPIA / FRIA (fundamental rights impact assessment), bias audits, model risk assessments, contract review for third-party AI.
- **Incident handling.** Coordinate response to compliance issues; regulator notifications (EU AI Office, Cal OES, state AGs); cross-functional incident triage.
- **Manage the standards stack.** ISO/IEC 42001 (management system), 23894 (risk), 42005 (impact assessment) are no longer optional for large organisations.

Key concerns

Liability, regulatory sanctions, reputational risk. The 2025-2026 enforcement landscape is harsher than the 2024 landscape: EU AI Act penalties (up to EUR 35M / 7% turnover), Texas TRAIGA penalties ($10K-$200K), TAKE IT DOWN Act FTC enforcement, FTC Section 5, state AG actions.

What changed in 2025-2026

- **Federal preemption uncertainty** in the US complicates multi-state compliance — track the December 2025 EO and litigation outcomes.

- **Audit-and-certification market** for ISO/IEC 42001 is now real with 42006 published; certification is increasingly procurement-relevant.
- **Frontier-model rules** add a layer for large developers — see Frontier Models.

Executives & Board Members

C-suite, board directors, AI sponsors

Focus

Executives are responsible for **strategic oversight** and **organisational commitment** to AI governance. The job is to balance innovation with risk and to maintain stakeholder trust.

Specific practices

- **Set strategy.** Decide which AI use cases the organisation will pursue, which are out of bounds, and the corresponding risk appetite.
- **Establish governance structures.** AI Governance Council, model risk management function, AI ethics committee with real authority and budget.
- **Set the tone.** Communicate that responsible AI is a core value; reward responsible behaviour; back compliance teams when they say "not yet."
- **Own accountability.** Boards now routinely ask for AI risk reports. Be ready with metrics: number of AI systems in production, classification by risk tier, recent incidents, audit findings, compliance posture by jurisdiction.
- **Prepare for regulation.** Where regulation is in force, fund compliance programmes proportionate to risk. Where regulation is pending (UK AI Bill, Brazil PL 2338), monitor and plan.
- **Pursue strategic certifications.** ISO/IEC 42001 certification, GPAI Code of Practice signature (for frontier developers), CAISI agreements where applicable.

Considerations for 2025-2026

- **ESG and AI** — trustworthy AI is increasingly part of ESG reporting and investor expectations.
- **AI workforce** — reskilling, AI literacy obligations under EU AI Act Article 4, internal AI usage policies.
- **Geopolitical risk** — export controls, data-localisation rules, jurisdictional fragmentation. The US December 2025 preemption EO and the ongoing state-federal tension affect operational planning.

What changed in 2025-2026

- **AI-related liability** has moved from theoretical to concrete — Bartz v. Anthropic settlement, TAKE IT DOWN Act platform liability, EU AI Act penalties.
- **Frontier developers** face additional executive-level expectations: published frontier framework, safety cases, whistleblower protections under California SB 53.

Policymakers & Regulators

Government officials, regulators, standards body participants

Focus

Policymakers create and enforce the rules. The job is to address public risks while preserving innovation and accommodating sectoral diversity.

Specific focus areas

- **Develop and refine AI regulations.** EU AI Act implementation and Omnibus refinement, state legislative work, sector-specific rules.
- **Harmonise where possible.** OECD, G7 Hiroshima Process, ISO/IEC, IEEE; bilateral cooperation agreements (e.g., the International Network of AI Safety Institutes).
- **Build enforcement capacity.** Stand up AI offices and inspectorates; train staff; develop technical evaluation capability.
- **Address societal impacts.** Workforce displacement, AI literacy, public-sector AI use, election integrity, deepfake harms.

Concerns

Prevent harm; preserve fundamental rights; ensure national security; maintain transparency and accountability; balance with innovation; avoid over-regulation.

What changed in 2025-2026

- **Multilateral fragmentation.** The Paris AI Action Summit produced a 58-nation statement that the US and UK declined to sign — coordination is harder than it was in 2023-2024.
- **Capacity building.** The International Network of AI Safety Institutes is the most concrete operational coordination layer; bilateral CAISI agreements show what national capacity can deliver.

- **Preemption tensions.** The US December 2025 preemption EO has created uncertainty for state policymakers; the EU Omnibus shows how complex even single-jurisdiction harmonisation is in practice.

Cross-audience: the shared language

These four audiences must work together. The common language across them increasingly is:

- **ISO/IEC 42001** as the management-system architecture.
- **NIST AI RMF** as the risk-management vocabulary.
- **EU AI Act risk tiers and high-risk obligations** as the most prescriptive legal baseline.
- **Safety case** as the unit of assurance for high-impact and frontier systems.

Investing in shared vocabulary and shared documentation formats (model cards, datasheets, impact assessments, audit reports) materially reduces the friction between these roles.

AI Maturity Stages Frameworks

To help organisations assess and improve their AI governance and responsible-AI practice, we present six **AI Maturity Stages frameworks**. Each is a **seven-stage** model describing the progression from rudimentary or non-existent capability to highly integrated and continuously improving capability in a specific area:

1. **AI Governance Maturity** — organisational governance capability.
2. **AI Safety Maturity** — technical safety and reliability.
3. **AI Trust & Transparency Maturity** — stakeholder trust and transparency.
4. **Responsible AI Maturity** — ethics and social responsibility.
5. **AI Risk Management Maturity** — holistic risk management.
6. **AI Compliance Maturity** — adherence to external regulations and standards.

Use these to benchmark current state and plan improvements. While details differ per framework, the underlying pattern is consistent: progress from ad-hoc / reactive practice toward proactive, optimised, continuously improving practice. Most organisations are at different stages across the six frameworks; that is expected and often desirable (e.g., a healthcare AI vendor may legitimately have higher Compliance maturity than Trust & Transparency maturity early in its journey).

Figure: The common seven-stage progression. Each of the six frameworks below applies this same arc to a different dimension of AI practice.

1. AI Governance Maturity Stages

Stage 1: Ad Hoc & Chaotic

> No formal AI governance. AI projects in silos with little oversight. Decisions on ethics or risk left to individual teams. No leadership awareness of AI-specific risk.
>
> **Assessment:** No dedicated AI policies or roles exist.
>
> **Challenge:** Lack of coordination — ethical or compliance breaches go unnoticed.
>
> **Best practice:** Begin awareness building — basic AI risk workshop, inventory existing AI projects.

Stage 2: Aware (Initial Awareness & Planning)

> The organisation has recognised the need for AI governance and is planning. Working groups form. Policies in draft. A champion may be advocating internally.
>
> **Assessment:** Initial AI governance framework document; ethics committee formed (even without authority).

Challenge: Moving from talk to action.

Best practice: Roadmap with concrete milestones — publish AI ethics policy, assign roles, pilot procedures on one project.

Stage 3: Fragmented (Basic Policies, Inconsistent Adoption)

Basic policies exist; adoption is spotty. Some teams comply, others don't. Reviews for high-profile projects; many projects slip through.

Assessment: Policies on paper; some training delivered.

Challenge: Enforcement and coverage; viewed as box-ticking.

Best practice: Integrate governance into project lifecycle (sign-off gates); communicate success stories.

Stage 4: Defined & Implemented

Formal AI governance in place and functioning. Central committee or officer. Policies refined and communicated. Most projects follow required steps. AI governance is part of standard operating procedure.

Assessment: High percentage of AI initiatives follow the process; governance artefacts (risk assessments, model cards) exist per project. May target ISO/IEC 42001 alignment.

Challenge: Maintaining quality of execution; avoiding "compliance theatre."

Best practice: Internal audits; investment in tooling that enforces process; named accountable executive.

Stage 5: Managed & Measured

Governance is measured and managed with metrics. KPIs are tracked (e.g., percentage of high-risk systems with completed FRIA, number of incidents, audit findings closed). Process is refined based on data.

Assessment: Operational dashboards; regular reporting to executive leadership; tracked remediation pipelines.

Challenge: Avoiding metric overload; ensuring metrics reflect outcomes rather than activity.

Best practice: Tie incentives to governance outcomes; quarterly governance reviews with the board.

Stage 6: Integrated & Optimised

AI governance is integrated with quality, security, privacy, and risk management. The organisation operates an integrated management system (e.g., ISO 9001 + 27001 + 42001). External certification achieved. Practices continuously improved.

Assessment: Certifications; mature change-management process; cross-functional governance council with real authority.

Challenge: Sustaining maturity through organisational change.

Best practice: Share lessons learned externally; participate in standards bodies.

Stage 7: Transformative & Industry Leader

The organisation is a recognised industry leader on AI governance. Governance practice is a competitive advantage. The organisation shapes external standards and norms.

Assessment: Public thought leadership; participation in standards development; cited by regulators as a model.

Challenge: Avoiding complacency; staying ahead of evolving practice.

Best practice: Open-source governance tooling; publish a transparency report; mentor industry peers.

2. AI Safety Maturity Stages

Stage 1: Negligent to Safety

No deliberate safety practice. Models deployed without testing for adversarial inputs, drift, or failure modes.

Best practice: Establish minimum testing baseline; document known failure modes.

Stage 2: Reactive Safety Fixes

Safety issues addressed after they manifest. No proactive testing.

Best practice: Build incident response playbook; track recurring failure patterns.

Stage 3: Basic Testing & Validation

Standardised validation tests for new models. Some adversarial testing. Documented evaluation suites.

Best practice: Adopt NIST AI 600-1 GenAI Profile threat categories where applicable.

Stage 4: Proactive Risk Assessment & Mitigation

Risk assessment is standard for every AI project. Mitigations documented and tracked. Operating-domain documentation produced.

Best practice: Align with ISO/IEC 23894; produce model cards per system.

Stage 5: Advanced Technical Safeguards

Adversarial training, ensemble methods, guardian systems, formal verification of safety-critical modules.

Best practice: Red-teaming as a standing practice; published evaluation results.

Stage 6: Continuous Safety Management

Continuous monitoring in production; drift detection; automated rollback. Safety incidents trigger root-cause analysis and feedback loops.

Best practice: Integrate safety telemetry into engineering dashboards; quarterly safety reviews.

Stage 7: Safety as a Differentiator

Best-in-class safety record. Safety case methodology operational. Recognised as a safety leader by regulators (EU AI Office, CAISI, sector regulators).

Best practice: Publish safety reports; participate in international AI Safety Institute network.

3. AI Trust & Transparency Maturity Stages

Stage 1: Opaque & Untrusted

AI systems are black boxes; users have no information about how decisions are made.

Best practice: Begin publishing basic system descriptions; identify where transparency is legally required.

Stage 2: Basic Disclosures

Some disclosures — users informed they're interacting with AI; minimal information provided.

Best practice: Comply with EU AI Act Article 50 transparency obligations; provide AI-generated content labels.

Stage 3: Explainability for Internal Use

Engineering teams use interpretability tooling (SHAP, LIME, saliency maps). Internal reviews of model decisions.

Best practice: Produce model cards; document operating domains.

Stage 4: User-Facing Explainability

End-users receive plain-language explanations of consequential AI decisions; appeals process available.

Best practice: Comply with GDPR Article 22; provide actionable reason codes.

Stage 5: Interactive Transparency & Engagement

Users can query AI decisions, provide feedback, and influence outcomes. Public transparency reports published.

Best practice: Adopt content provenance standards (C2PA); publish training-data summaries.

Stage 6: Trusted AI Ecosystem

The organisation's AI is trusted by users, partners, and regulators. Third-party audits regularly published. ISO/IEC 42001 certified.

Best practice: Engage with downstream stakeholders; publish responsible AI use cases.

Stage 7: Industry Transparency Leader

The organisation defines the transparency standard for the industry. Practices are adopted by peers and codified by regulators.

Best practice: Contribute to international standards (ISO, IEEE, C2PA); open-source tooling.

4. Responsible AI Maturity Stages

Stage 1: Unaware / Unprincipled

No articulated ethics or responsibility principles. AI deployed without consideration of societal impact.

Best practice: Begin articulating principles; appoint ethics owner.

Stage 2: Articulated Principles (on Paper)

Ethics principles published; not yet operationalised. Risk of "ethics washing" without enforcement.

Best practice: Move from principles to processes; assign accountability.

Stage 3: Procedures and Training for Ethics

Ethics training rolled out; review procedures established; ethics escalation path defined.

Best practice: Make ethics review a default gate for AI projects.

Stage 4: Integrated Responsible AI Practices

Responsible AI practices integrated into product development. Bias mitigation, fairness checks, FRIA-style assessments standard.

Best practice: Conduct FRIAs under EU AI Act Article 27; align with ISO/IEC 42005.

Stage 5: External Accountability and Audit

External audits of ethics and responsibility. Independent ethics board with real authority. Public ethics commitments.

Best practice: Engage civil society; respond to external concerns.

Stage 6: Culture of Responsibility & Empowerment

Responsible AI is part of organisational culture. Employees feel empowered to raise concerns. Whistleblower protections in place (cf. California SB 53).

Best practice: Reward responsible decisions; protect whistleblowers; act on internal escalations.

Stage 7: Social Stewardship and Advocacy

The organisation actively advocates for responsible AI in the broader ecosystem. Funds research; supports public-interest initiatives (e.g., Current AI, ROOST).

Best practice: Sponsor open-source safety work; contribute to multilateral processes.

5. AI Risk Management Maturity Stages

Stage 1: No AI-specific Risk Management

AI risks not distinguished from general enterprise risks. No AI risk register.

Best practice: Create an AI risk register; inventory AI systems.

Stage 2: Qualitative Acknowledgment of AI Risks

AI risks identified at a high level. Documented but not quantified or mitigated systematically.

Best practice: Adopt NIST AI RMF as a starting framework.

Stage 3: Structured Risk Assessment Process

Standard process for risk assessment per AI project. NIST RMF Map and Measure functions implemented.

Best practice: Use NIST trustworthiness characteristics (privacy, accuracy, safety, fairness, etc.) as risk categories.

Stage 4: Risk Mitigation and Control Implementation

Risks have documented controls. NIST RMF Manage function implemented. Aligned with ISO/IEC 23894.

Best practice: Map controls to ISO/IEC 23894 risk treatment options.

Stage 5: Integrated Risk Management & Monitoring

AI risk integrated with enterprise risk management. Real-time monitoring of risk indicators. Cross-functional risk reviews.

Best practice: Quarterly AI risk reviews at executive level; aggregate dashboards.

Stage 6: Advanced Quantitative Risk Analysis

Quantitative risk models for AI — scenario analysis, sensitivity testing, financial risk modelling for AI-related losses.

Best practice: Apply techniques from financial-services model-risk management (SR 11-7 lineage) to AI broadly.

Stage 7: Adaptive and Resilient Risk Posture

Continuous improvement of risk practice. Resilience tested via tabletop exercises and red-team scenarios. Risk posture adapts to new threats (e.g., novel attacks against frontier models).

Best practice: Industry-leading incident-response drills; contribute to threat-intelligence sharing.

6. AI Compliance Maturity Stages

Stage 1: Non-compliant (Ignorant or Defiant)

Not aware of or not complying with applicable regulations.

Best practice: Audit current AI footprint against applicable regulations (EU AI Act, US state laws, sector rules).

Stage 2: Aware of Regulations

Aware of applicable rules but not yet implementing controls.

Best practice: Map regulations to AI systems; prioritise high-risk gaps.

Stage 3: Implementing Policies and Controls for Compliance

Policies and controls being implemented. Some AI systems compliant; gaps remain.

Best practice: Use ISO/IEC 42001 as architecture; close gaps systematically.

Stage 4: Comprehensive Compliance Management System

End-to-end compliance management. ISO/IEC 42001 aligned. Documented evidence per regulation.

Best practice: Integrate compliance evidence collection into engineering workflow.

Stage 5: Audit Readiness and External Certification

Ready for external audit. ISO/IEC 42001 certification pursued under 42006-accredited bodies. GPAI Code of Practice signed where applicable.

Best practice: Maintain audit evidence continuously; engage external auditors annually.

Stage 6: Compliance as Business Enabler

Compliance posture is a competitive advantage. Certifications and signatures used in procurement. Customers and regulators trust the organisation.

Best practice: Market compliance credentials; use them to enter regulated markets faster.

Stage 7: Thought Leader and Shaper in AI Compliance

The organisation shapes compliance norms. Engages with regulators on rule development. Practices cited as exemplary in regulatory guidance.

Best practice: Participate in standards development; contribute to regulator working groups.

How to use these frameworks

1. **Assess.** Identify your current stage in each of the six frameworks. Honest assessment matters more than aspirational labels.
2. **Prioritise.** Identify the framework where progress matters most for your organisation's strategy and risk — often Compliance for regulated industries, Safety for frontier developers, Responsible AI for consumer-facing deployments.
3. **Plan.** Identify the specific practices needed to move to the next stage. Refer to relevant chapters: Legal & Regulatory, Technical Safety, Privacy, Data & Security, Frontier Models.
4. **Measure.** Track progress with concrete metrics (audits completed, controls implemented, incidents reduced, certifications achieved).

5. **Iterate.** Re-assess annually. Maturity is a journey, not a destination — and the regulatory environment continues to evolve.

Glossary of AI Governance Terms

Definitions are aligned with **ISO/IEC 22989:2022** where the standard provides one; ISO-sourced definitions are explicitly cited.[1] Definitions for newer concepts (agentic AI, GPAI, frontier model) are drawn from current regulatory instruments where available.

Agentic AI

AI system designed to perform multi-step actions in pursuit of a goal, typically via tool use, code execution, or interaction with external systems. Distinguished from non-agentic AI by the ability to take autonomous actions beyond producing output for direct human use.

AI agent

"Entity that senses and responds to its environment and takes actions to achieve goals."

ISO/IEC 22989:2022(E), 3.1.3

AI Act (EU)

Regulation (EU) 2024/1689 establishing harmonised rules on artificial intelligence. The world's first horizontal AI statute, entered into force 1 August 2024 and being enforced in phases. See EU AI Act chapter for current obligations and Omnibus amendments.

AI component

"Functional element that constructs an AI system."

ISO/IEC 22989:2022(E), 3.1.4

AI Office (EU)

The body within the European Commission (DG CONNECT) responsible for enforcing the EU AI Act with respect to general-purpose AI models, cross-border cases, and the GPAI Code of Practice. Established and operational since 2 August 2025.

AI Safety Institute / Center for AI Standards and Innovation (CAISI)

National bodies focused on frontier-model evaluation and AI safety standards. The US AISI was renamed CAISI in June 2025; UK AISI and others continue under the AISI brand. Members of the International Network of AI Safety Institutes coordinate on frontier-model evaluation. See Frontier Models.

Artificial intelligence (AI)

"Capability of an engineered system to acquire, process, and apply knowledge and skills." (ISO/IEC 22989). In regulatory practice, definitions vary; the EU AI Act Article 3(1) definition is widely cited.

ISO/IEC 22989:2022(E), 3.1.2

Artificial intelligence system (AI system)

"Engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives." (ISO/IEC 22989). The EU AI Act and OECD definitions are functionally aligned.

ISO/IEC 22989:2022(E), 3.1.5

Automated decision-making (ADM)

Decision-making by algorithmic or AI systems without human intervention. GDPR Article 22 restricts solely-automated decisions producing legal or similarly significant effects, requiring human review or meaningful information about the logic.

Bias audit

Structured evaluation of an AI system for differential performance or disparate impact across protected groups. Required by NYC Local Law 144 for automated employment decision tools and emerging as best practice broadly.

Conformity assessment

Process required by the EU AI Act for high-risk AI systems to demonstrate compliance before placing on the market. May involve a notified body for third-party assessment or self-assessment depending on the system category (Articles 43, 44).

Consequential decision

Decision affecting a consumer's access to or pricing of employment, education, financial services, essential government services, healthcare, housing, insurance, or legal services. Used in Colorado SB 24-205 to define the scope of high-risk AI systems.

Constitutional AI / RLAIF

Alignment technique using AI-generated feedback against an explicit set of principles to train models toward desired behaviour. Reduces reliance on human raters at scale.

Datasheet for dataset

Standardised documentation describing a dataset's motivation, composition, collection process, preprocessing, recommended uses, distribution, and maintenance. Proposed by Gebru et al. (2018); widely adopted as a transparency tool.

Differential privacy

Mathematical framework providing formal guarantees about the influence of any individual record on a model's output. Implemented through noise addition during training (DP-SGD) or output aggregation.

Explainability

"Property of an AI system to express important factors influencing the AI system results in a way that humans can understand." Distinguished from interpretability (which refers to the model's intrinsic understandability).

ISO/IEC 22989:2022(E), 3.5.4

Federated learning

Training paradigm in which a model is trained across distributed datasets without centralising the data. Widely deployed in healthcare, mobile keyboards, and cross-institutional research where data cannot or should not leave its source.

Foundation model

Large model trained on broad data, intended to be adapted to many downstream tasks. In US discourse, often used interchangeably with "general-purpose AI" or "frontier model" though the terms are technically distinct.

Frontier model

Most capable AI models, typically those exceeding compute thresholds (10^{25} FLOPs in the EU AI Act) or designated by regulators for posing systemic risk. See Frontier Models.

Fundamental Rights Impact Assessment (FRIA)

Assessment required by EU AI Act Article 27 for certain high-risk AI systems used by deployers, evaluating impacts on fundamental rights protected by the Charter. ISO/IEC 42005 provides operational methodology.

General-purpose AI (GPAI) model

AI model "trained with a large amount of data using self-supervision at scale, [that] displays significant generality and is capable of competently performing a wide range of distinct tasks." (EU AI Act Article 3(63)). Subject to Articles 53-55 obligations.

GPAI Code of Practice

Voluntary instrument published 10 July 2025 that operationalises GPAI obligations under the EU AI Act. Three chapters: Transparency, Copyright, Safety & Security. Signatories include Google, Microsoft, OpenAI, Anthropic. See EU AI Act.

Hallucination

Generation of plausible but factually incorrect or fabricated output by a generative AI model. Mitigated through retrieval grounding, output verification, and clearer uncertainty signalling.

High-risk AI system (EU AI Act)

AI system included in Annex III or Annex I of the EU AI Act, subject to risk management, data governance, technical documentation, transparency, human oversight, accuracy, robustness, and conformity assessment obligations. Deadlines rebased by the May 2026 Omnibus.

Impact assessment (AI)

Structured evaluation of an AI system's effects on individuals, groups, and society, covering risks, stakeholder analysis, fundamental-rights impacts, and mitigations. Standardised in ISO/IEC 42005:2025.

Interpretability

Property of a model being inherently understandable to humans. Distinguished from explainability (which refers to producing human-understandable explanations for opaque models).

ISO/IEC 42001

International standard for AI management systems, published December 2023. Certifiable. Anthropic was the first AI developer to achieve certification (January 2025). See ISO Standards.

Machine learning (ML)

"Process using computational techniques to enable systems to learn from data or experience." Subfield of AI.

ISO/IEC 22989:2022(E), 3.2.10

Model card

Documentation describing an AI model's intended use, performance, limitations, training data, evaluation, and ethical considerations. Increasingly required by regulatory regimes (EU AI Act Article 11; California AB 2013 for training-data summary).

NIST AI Risk Management Framework (RMF)

Voluntary US framework published January 2023, organising AI risk management into four functions: Govern, Map, Measure, Manage. Companion profiles include the Generative AI Profile (NIST AI 600-1) and Cybersecurity Framework Profile for AI (NIST IR 8596). See US Federal.

Reinforcement Learning from Human Feedback (RLHF)

Alignment technique training models using human preference signals. Widely deployed for instruction-following and value alignment in large language models.

Responsible Scaling Policy (RSP)

Frontier developer's published commitment to specific capability evaluations and risk thresholds, with mitigation actions triggered at defined thresholds. Originated with Anthropic; adopted in different forms by other frontier developers.

Risk-based regulation

Regulatory approach that calibrates obligations to the risk posed by a system or activity. The EU AI Act, Colorado AI Act, and Korea AI Basic Act are leading examples.

Safety case

Structured argument and evidence demonstrating that a system is acceptably safe for a defined use. Required for frontier models by California SB 53 and emerging as a standard unit of assurance for high-impact AI.

Systemic risk (EU AI Act)

Risk specific to high-impact capabilities of GPAI models. Models exceeding 10^{25} FLOPs of training compute are presumed to pose systemic risk; the AI Office may also designate models as posing systemic risk based on capability assessment.

Trustworthiness (in AI)

"Ability to meet stakeholders' expectations in a verifiable way." Cross-cutting property addressed in NIST AI RMF; characteristics include validity, reliability, safety, security, accountability, transparency, explainability, privacy, and fairness.

ISO/IEC 22989:2022(E), 3.5.16

---

1. ISO/IEC 22989:2022, Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. ↩

Additional Resources

Regulators and standards bodies

- European Commission — AI Act
- EU AI Office
- EU GPAI Code of Practice
- NIST — AI Risk Management Framework
- NIST — Center for AI Standards and Innovation (CAISI)
- ISO/IEC JTC 1/SC 42 — Artificial Intelligence
- OECD AI Policy Observatory
- UK AI Security Institute
- Korea Personal Information Protection Commission

Standards and frameworks

- ISO/IEC 42001:2023 — AI Management Systems
- ISO/IEC 23894:2023 — AI Risk Management
- ISO/IEC 22989:2022 — AI Concepts and Terminology
- ISO/IEC 42005:2025 — AI System Impact Assessment
- ISO/IEC 42006:2025 — Requirements for Bodies Providing Audit and Certification of AIMS
- NIST AI 600-1 — Generative AI Profile
- IEEE 7000 series

Open-source tools

- IBM AI Fairness 360
- Microsoft Fairlearn
- Microsoft Responsible AI Toolbox
- Google Model Cards Toolkit
- Hugging Face Evaluate
- ROOST — Robust Open Online Safety Tools

Communities and initiatives

- Partnership on AI
- Frontier Model Forum
- International Network of AI Safety Institutes
- MLCommons AI Safety Working Group
- Current AI

References & Bibliography

Sources cited throughout this handbook. Citations are also embedded as footnotes within each chapter; this page provides a consolidated bibliography organised by topic.

Primary regulatory sources

EU AI Act and related instruments

1. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal.
2. European Commission. Regulatory framework on artificial intelligence.
3. European Commission. (2025, February 4). Guidelines on prohibited artificial intelligence practices defined by the AI Act.
4. EU GPAI Code of Practice. code-of-practice.ai.
5. Council of the EU. (2026, May 7). Artificial intelligence: Council and Parliament agree to simplify and streamline rules.
6. Artificial Intelligence Act EU. Article 99: Penalties.

US federal

1. The White House. (2025, January 23). Executive Order 14179 — Removing Barriers to American Leadership in Artificial Intelligence.
2. The White House. (2025, July 23). America's AI Action Plan.
3. The White House. (2025, December 11). Eliminating State Law Obstruction of National Artificial Intelligence Policy.
4. OMB Memorandum M-25-21 (April 2025). Federal Use of Artificial Intelligence.
5. OMB Memorandum M-25-22 (April 2025). Federal AI Procurement.
6. TAKE IT DOWN Act, Pub. L. 119-12 (2025). Overview.
7. NIST. AI Risk Management Framework.
8. NIST. Center for AI Standards and Innovation (CAISI).
9. NIST. AI 600-1: Generative AI Profile.
10. CFPB. (2025, September 26). AI Compliance Plan for OMB M-25-21.
11. Wiley. BIS Rescinds AI Diffusion Rule.
12. FedScoop. (2025). Trump administration rebrands AI Safety Institute as CAISI.

US state law

1. Colorado General Assembly. SB 24-205 — Colorado AI Act.
2. Clark Hill. Colorado's AI law delayed until June 2026.
3. Latham & Watkins. Texas Signs Responsible AI Governance Act into Law.
4. Office of the Governor of California. (2025, September 29). Governor Newsom signs SB 53.
5. Utah Legislature. SB 226 (2025).
6. New York State Comptroller. (2025, December 2). Enforcement of Local Law 144.

International law

1. Fasken. Prorogation's Digital Impact — Canada Bill C-27.
2. Osborne Clarke. Regulatory Outlook January 2026: Artificial Intelligence (UK).
3. Cooley. South Korea's AI Basic Act: Overview and Key Takeaways.
4. White & Case. Japan's First AI Legislation Becomes Law.
5. China Law Translate. AI Labeling Measures.
6. Library of Congress. Brazil Senate Advances Discussions on Bill to Regulate AI Use.

ISO/IEC standards

1. ISO/IEC. 42001:2023 — AI Management Systems.
2. ISO/IEC. 23894:2023 — AI Risk Management.
3. ISO/IEC. 22989:2022 — AI Concepts and Terminology.
4. ISO/IEC. 42005:2025 — AI System Impact Assessment.
5. ISO/IEC. 42006:2025 — Requirements for Bodies Providing Audit and Certification of AI Management Systems.
6. Osler, Hoskin & Harcourt LLP. The role of ISO/IEC 42001 in AI governance.

Sector-specific

1. FDA. Predetermined Change Control Plan for AI-Enabled Device Software Functions (Final Guidance, December 2024).
2. OCC Bulletin 2025-26. Model Risk Management Clarifications for Community Banks.

Case law

1. Bartz v. Anthropic, N.D. Cal. (2025). Ruling: ArentFox Schiff, Landmark Ruling on AI Copyright.
2. Authors Guild. What Authors Need to Know About the Anthropic Settlement.
3. Kadrey v. Meta Platforms, Inc., N.D. Cal. (2025).

Privacy and data protection

1. GDPR Info. Art. 22 GDPR — Automated individual decision-making.
2. Cloudflare. What is the CCPA?.

Frameworks and principles

1. OECD. AI Principles.
2. ANSI. (2024, May 9). OECD Updates AI Principles.
3. OECD. (2019). Recommendation of the Council on Artificial Intelligence.
4. UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence.
5. AI Action Summit, Paris (2025). Wikipedia overview.

Academic literature

1. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of FAccT '19. https://doi.org/10.1145/3287560.3287596

2. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for Datasets. Communications of the ACM, 64(12). https://doi.org/10.1145/3458723
3. Arrieta, A. B., et al. (2020). Explainable Artificial Intelligence (XAI). Information Fusion, 58. https://doi.org/10.1016/j.inffus.2019.12.012
4. Sánchez, I., et al. (2024). Evolving AI Risk Management: A Maturity Model based on the NIST AI Risk Management Framework. arXiv:2401.15229.

Industry frameworks and reports

1. GSMA. (2024). The GSMA Responsible AI Maturity Roadmap (PDF).
2. Trustible. Everything you need to know about the NIST AI RMF.
3. Thoropass. Understanding the NIST AI Risk Management Framework.

Changelog

This handbook follows Keep a Changelog conventions and Semantic Versioning. Major versions reflect substantive restructures; minor versions add new chapters; patches fix typos and update citations.

v2.0.0 — 2026-05-11

The first major rewrite since v1.0.0. This release brings the handbook up to date with regulatory developments from March 2025 through May 2026, restructures the source from a single 1,917-line HTML file into a multi-page Eleventy site, and adds three new chapters.

Added

- **New chapter: US State Laws** covering Colorado SB 24-205 (delayed to June 30, 2026), Texas TRAIGA, California SB 53 / AB 2013 / SB 942, Utah SB 226, Tennessee ELVIS Act, NYC Local Law 144 (and the December 2025 NYS Comptroller audit), Illinois HB 3773.
- **New chapter: Copyright & IP** covering Bartz v. Anthropic ($1.5B settlement, September 2025), Kadrey v. Meta, ongoing litigation, EU AI Act Article 53 copyright obligations, California AB 2013.
- **New chapter: Frontier Models** consolidating GPAI Code of Practice, CAISI testing agreements, California SB 53, Korea AI Basic Act frontier-safety track, Responsible Scaling Policies, and the International Network of AI Safety Institutes.
- **/changelog/** page (this page).
- **Side-by-side multi-page navigation** with expandable Legal & Regulatory section.
- **ISO/IEC 42005:2025** (AI System Impact Assessment) and **ISO/IEC 42006:2025** (audit/certification body requirements) added to the ISO chapter.
- **Print stylesheet** and **Playwright-based PDF generation** to keep web and PDF outputs visually consistent.
- **Backwards-compatibility shims** — legacy URL hashes (`#legal-frameworks`, `#privacy-security`, etc.) redirect to the new page paths.

Changed

- **EU AI Act chapter** completely rewritten. Past dates (February 2, 2025 prohibitions; August 2, 2025 GPAI obligations) reframed as enforced fact. The May 7, 2026 Digital Omnibus agreement — rebasing the high-risk Annex III deadline from December 2026 to December 2027, and Annex I from August 2027 to August 2028 — is incorporated with a PENDING flag.
- **US Federal chapter** completely rewritten. Rescinded EO 14110 references removed. EO 14179 ("Removing Barriers to American Leadership in AI", January 23, 2025), OMB M-25-21 / M-25-22, the America's AI Action Plan (July 23, 2025), and the December 11, 2025 preemption EO covered. CAISI rebrand documented. TAKE IT DOWN Act covered. EEOC and OFCCP guidance removal noted.
- **International chapter** rewritten to cover Korea AI Basic Act (effective January 22, 2026), Japan AI Promotion Act, China AI labeling measures (effective September 1, 2025), Brazil PL 2338 status, Canada AIDA expiration, UK Blueprint (October 21, 2025), Paris AI Action Summit (February 2025).
- **Sectoral chapter** updated for FDA PCCP Final Guidance (December 2024), OCC Bulletins 2025-26 and 2026-13, CFPB AI Compliance Plan (September 26, 2025).
- **ISO Standards chapter** updated for the publication of 42005:2025 and 42006:2025; certification examples (Anthropic, IBM Granite, UiPath, Changi Airport).

- **Privacy, Data & Security chapter** updated for the NIST GenAI Profile (March 2025 update with poisoning/evasion/extraction/manipulation threat categories), agentic AI considerations.
- **Technical Safety chapter** updated for Constitutional AI, mechanistic interpretability advances, agentic-systems engineering, and safety-case methodology.
- **Audience Guidance chapter** updated to reflect 2025–2026 obligations on each role; frontier-developer additions for executives and practitioners.
- **Frontier Models** model citations refreshed: GPT-5 (Aug 7, 2025), Gemini 3 Pro (Nov 18, 2025), Claude Opus 4.5 (Nov 24, 2025).
- **References** consolidated and updated with current primary sources.

Removed

- Forward-looking framing of February 2, 2025 prohibitions and August 2, 2025 GPAI obligations (now in force).
- References to EO 14110 as current US policy.
- "Emerging" framing for generative AI, EU AI Act transparency, NIST RMF.
- ChatGPT as the primary LLM example.
- Outdated "Bill C-27 will create AIDA" framing (Bill C-27 prorogued).

Infrastructure

- Migrated from single hand-authored `index.html` (1,917 lines) to Eleventy multi-page source.
- Source files now Markdown with YAML front matter; footnotes via `markdown-it-footnote`.
- GitHub Actions workflow rewritten to build the site, generate PDF via Playwright, and deploy to GitHub Pages.
- Removed duplicate `governance.yml` workflow (byte-identical to `static.yml`).
- Repository now follows standard Node project structure: `src/`, `_includes/`, `_data/`, `assets/`, `scripts/`, `package.json`.

v1.0.0 — 2025-03

Initial public release. Single-page HTML; PDF distributed via Microsoft Forms gate. Covered:

- ISO/IEC 42001, 23894, 22989.
- EU AI Act risk classification and phased timeline (as understood in March 2025).
- NIST AI RMF 1.0.
- GDPR, CCPA basics.
- Six seven-stage AI maturity models.
- Audience-specific guidance for practitioners, compliance, executives, policymakers.
- Glossary.

Versioning policy

- **MAJOR** — significant restructure, deletion of chapters, or breaking changes to page URLs.
- **MINOR** — new chapter or section; substantive new material.
- **PATCH** — typo fixes, citation updates, link maintenance, small clarifications.

About the Author

Khullani M. Abdullahi, J.D.

Founder of Techne AI

Khullani M. Abdullahi specialises in AI governance, compliance, and ethics. With a background in law and technology, Khullani helps organisations navigate the complex landscape of AI regulation and responsible implementation.

LinkedIn Website Newsletter

About this handbook

The handbook is research-supported rather than legal advice. Where it references specific regulations, statutes, or court decisions, citations point to primary sources or recognised secondary commentary — see References. Organisations should consult qualified counsel for compliance decisions in their specific jurisdictions and sectors.

All content has been reviewed and edited by the author. The bibliography in References is the authoritative list of sources consulted.