Securys®

State of AI Regulation

Current and forthcoming rules, compliance guidance and risk mitigation.

—

7 OCTOBER 2026



Contents

Purpose of this Paper 4
Scope 4
Transparency 4
Fairness 5
Safety 5
Recourse 5
Recommendations 6
Definition of "AI" 7
The Securys perspective 10
Current and emerging regulatory landscape 11
GDPR 11
United States ..... Error! Bookmark not defined.
China 23
EU AI Act: current obligations and implementation 12
UK regulatory and legislative position 18
US regulatory developments ..... Error! Bookmark not defined.
International standards and frameworks 26
Practical Considerations 28
Minimum governance and compliance measures 28
Over-estimation of capability risk 31
Case failure 32
Reliance risk 33
Rigidity risk 34
Training failures 35
Query ambiguity, error or bias 39
Cyber risks 40
Transparency risks 41
Oversight and governance 42
Rights limitations and leakage risk 43
Abuse 44
Real world examples 46
Minimum requirements - and next steps 47
Jurisdictional and legislative tracker 48
Primary Sources 49

Purpose of this paper

As the use of artificial intelligence, machine learning and other automated decision-making and decision support systems grows, so likewise do the ethical, cybersecurity and data protection concerns surrounding their deployment. This paper is focused on data protection and considers the positions being adopted by the European Union – which can be deemed to be "leading the charge" in regulating the use of AI –China, the US and elsewhere.

This paper sets out some general guidelines for the ethical and compliant use of AI and works through a number of practical examples to help elucidate the issue and present in each case a compliant approach and appropriate safeguards. This is not a technical paper – please contact Securys if you would like more detailed advice on specific uses of technology or more detail on a particular aspect of the present and proposed regulations.

As this paper is concerned with data protection, it speaks to cases where personal data - information that pertains to or can be connected to an identifiable living person ("the data subject") - are involved, or where outputs from an AI system have legal or material effects on living people.

This paper does not address specific regulatory and draft regulatory provisions or wider considerations regarding the use of AI in a law enforcement context except where this is specifically relevant for the intended commercial audience.

Illustration created using Adobe Firefly AI image generator.



Scope

AI governance must cover the complete system and its intended use, including models, retrieval sources, tools and automated workflows. The EU AI Act uses a functional definition based on autonomy and inference; ordinary software that merely executes rules specified by humans is not automatically within scope. Data protection and other applicable laws can still govern automated processing outside the AI Act.[i]

The EU, UK, China and United States regulate AI through different combinations of AI-specific, data protection, consumer, employment and sectoral rules. Compliance requires jurisdiction-specific analysis; no single regime should be assumed to subsume all others.[i, v, vii, xi, xiii]

The EU AI Act entered into force on 1 August 2024 and is now operational, with prohibitions, AI literacy, general purpose AI and transparency obligations already applicable. The amended high-risk timetable phases in Annex III requirements from 2 December 2027 and Annex I product-related requirements from 2 August 2028. The UK has amended automated-decision rules through the Data (Use and Access) Act 2025. US federal policy has changed, while state legislation remains important. NIST provides voluntary risk-management guidance, not a substitute for law.[i-xi]

The core purposes of all of these present and future regulations can be summarised as seeking to ensure transparency, fairness, safety and recourse.

Transparency

Transparency is an operational compliance requirement, not simply an emerging policy objective. EU AI Act Article 50 distinguishes disclosure of interaction with AI, machine-readable marking of synthetic outputs, and deployer disclosure of deepfakes and specified public-interest text. The main requirements apply from 2 August 2026, with a transition under Article 111(4) to 2 December 2026 for Article 50(2) compliance by providers of relevant systems placed on the market before 2 August 2026.[i, ii]

China has required explicit and implicit labelling of AI-generated synthetic content since 1 September 2025. As recommended operational controls, organisations should preserve provenance metadata and use appropriate labels and watermarking, while recognising that provenance does not establish the truth of the content.[xiii, xiv]

Fairness

This is perhaps the broadest – and most important – aspect of the present and forthcoming regulation. There is both a desire to avoid bias and discrimination in AI systems in terms of their outputs and also a need to ensure that the data used to train heuristic systems is properly sourced and used with the knowledge of the people to whom it relates. The regulations mandate a wide-ranging and demanding data governance and data quality programme to achieve these aims and introduce bans on some potential uses and a distinction between high-risk and lesser-risk AI.

Safety

Both physical and psychological harm are considered in existing EU, Chinese and proposed UK and US regulation. These considerations not only drive the prohibition of certain uses of AI but also a significant extension of requirements for quality management and security provision for machine learning systems.

Taken together the fairness and safety objectives contribute to the need for a structured and consistent application of a code of ethics governing the use of AI within the enterprise. The Chinese in particular have produced a detailed code of ethics for AI which can form the basis of an internal programme.

Recourse

Existing regulation already provides for appeal to human review of automated decisions. This is reinforced in the EU AI Act legislation with a requirement for ongoing and proactive human oversight within organisational use of AI, and by a programme of registration, conformity assessment and market surveillance for machine learning systems.

Recommendations

This paper makes a number of recommendations in the context of known AI risks alongside the various functions and documentation required by regulation. These could be summarised as:

- Provide continuous, role-appropriate AI education. EU AI Act Article 4 has applied since February 2025; as amended, it requires measures supporting staff AI literacy development, rather than a guarantee of a particular literacy level for every individual. Record the measures taken and tailor them to the systems, tasks and affected people.[i, ii]
- Widespread and continuous education for commissioners, developers and users of AI systems to ensure awareness of the limitations of AI capabilities and appropriate application of that understanding.
- A central focus on ensuring that application of AI is of benefit to people and society with the use of existing tools such as the data protection impact assessment to check that the interests of the enterprise are properly balanced against those of individuals both within and outside the organisation.
- Implementation of feedback loops to ensure early detection and correction of errors and bias, alongside mechanisms to ensure that human decision makers can always override AI outputs.
- Policy and procedures making clear that use of AI is a skilled activity and that AI systems – however apparently anodyne or straightforward – should not be introduced without proper review, training and oversight.
- NOTE ON THE TEXT: Throughout the document there are both footnotes (which are numbered using Arabic numerals and are directly relevant to the page on which they occur), and endnotes (numbered in Roman numerals), which refer to the primary sources listed in the section on p. 51, and are referenced from multiple places.

Definition of "AI"

Artificial intelligence is an umbrella term for machine-based capabilities, not a synonym for consciousness or human-like general intelligence. Regulatory definitions focus on the system, its autonomy, inference and effects.[i]

An AI system may generate useful predictions, content, recommendations or decisions without possessing human intention or judgement. Governance should assess demonstrated capabilities and limitations in the intended context, rather than infer human understanding from fluent output.[i, xiv]

At present actual AI in use will either be some form of "machine learning", which uses a variety of techniques to "teach" a system to form outputs from inputs, or a so-called "expert or rule-based system", where rules for making decisions or producing outputs from inputs are algorithmically defined by a human programmer.

While a variety of techniques for teaching machine learning systems exist, this is not especially relevant to this paper. For a detailed explanation of the four most common approaches, see e.g. Yalçın, O.G. *Machine Learning Approaches that Every Data Scientist Should Know*[1]. The important consideration from a data protection perspective is that the actual mechanism by which the trained system reaches its conclusions is often opaque – either very difficult or functionally impossible for a human being to understand or manually replicate. As a consequence, it can both be difficult to determine whether a system is acting fairly and to explain its functioning to those who are affected by the decisions it makes or supports.

One other concern with machine learning is that because the machine has itself built the connections used to derive outputs from inputs, and because as noted above these systems are not sentient and not alert to context, it is unlikely to be drawing its conclusions in the way that a human would. This can lead to non-deterministic outputs, where what to a human appear to be identical input data produce disparate results. For the reasons noted above, it can be very difficult to determine what is causing the variance in response, which again can lead to challenges with ensuring fairness.

Rule-based expert systems can be easier to trace than learned models where their rules and data are documented. Determinism does not itself guarantee a comprehensible explanation or fair outcome; complexity, implementation and the explanation needed by the affected person remain relevant.[xv]

In 2023 we saw a step-change in the capability of the machine-learning approach with the advent of "generative pre-trained transformer" language and image models being

1 https://towardsdatascience.com/4-machine-learning-approaches-that-every-data-scientist-should-know-e3a9350e-c0b9?gi=83842174710b

released for public use. Transformer technology[4] was developed by Google in early 2017; it is in essence a change to the way that neural nets are able to identify and attend to features in training sets of, for example, words or pictures. While powerful it has also proved unreliable when coupled to a "generative" output function designed, not just to find patterns but to create them, and its use was mostly confined to the research departments of large tech companies while they figured out how to use it.

Foundation models are trained on broad data and can support multiple downstream tasks. Their operational capabilities depend on training, evaluation, adaptation and the surrounding application.[xiv]

Models may be adapted through fine-tuning or combined with retrieval-augmented generation, which supplies external information at inference time. Tool integration and multimodal inputs and outputs change the system-level risk profile.[xiv]

General-purpose, multimodal and agentic AI

Recent AI development is no longer adequately described solely as the production of text or images by large language models. General-purpose models increasingly process and generate combinations of text, images, audio, video and software code. They may be connected to retrieval systems, enterprise information, external tools, application programming interfaces and persistent memory. This allows a model to participate in a wider operational system rather than acting only as a standalone conversational interface.

A further development is the emergence of agentic systems. In this context, "agentic" describes systems configured to pursue an objective through a sequence of intermediate steps, which may include planning, retrieving information, selecting and invoking tools, communicating with other systems and taking actions in digital environments. The underlying model should not be treated as possessing human intention or judgement. The material governance distinction is that the surrounding application may now give model outputs operational effect.

This development changes the risk profile. A false statement generated in a draft is an accuracy problem; the same statement used by an automated workflow to update a record, send a communication, approve a transaction or alter access permissions may become a consequential operational event. Risk assessment must therefore cover the complete system, including the model, prompts, retrieval sources, tools, memory, identities and permissions, human approval points, logging, monitoring and connected third parties.

Generative AI also introduces risks that are not fully captured by conventional model-accuracy testing. These include confabulated outputs, prompt injection, leakage of

2 Attention is All You Need

confidential information, insecure tool use, harmful or misleading synthetic content, provenance failures, intellectual-property issues, model or data poisoning and automation bias. The NIST Generative AI Profile was published in July 2024 as a cross-sector companion to the AI Risk Management Framework and is intended to help organisations incorporate generative-AI-specific risks into the design, development, use and evaluation of AI systems.

Controls should consequently be applied at system and workflow level. They should include evaluation against defined use cases, verification of retrieval sources, restriction of tool permissions, separation of development and production environments, human approval before consequential actions, monitoring for anomalous behaviour, incident reporting and the ability to suspend or roll back an AI-enabled process.

The Securys perspective

Deep learning is basically a way of finding patterns in big data, both pre-existing and potential. In a sense it's a form of search, but one that allows you to search extrapolated possibilities in a probability landscape constrained by the training data set, rather than just letting you search what's already there (like a web search does).

It can feel uncannily human, because a lot of what we humans do in culture and conversation is not dissimilar. There are key differences however: humans balance this kind of "fast" system 1 thinking with a slower system 2 set of higher order concepts that help us contextualise the search outputs and see them within wider frameworks of physical, social and cultural implication – what we call "good judgement" or, more colloquially, "common sense". At least at present, AI is not able to do this second piece. It has no common sense because it has no worldview of its own and no sense of itself as an actor in a wider environment. Attempts to build Artificial General Intelligence ("AGI") are attempts to give it that kind of sensibility, something that could ultimately lead to self-consciousness. But at present, all the direction and awareness must come from us, its operators.

This means that AI remains a tool that we are deploying. This doesn't make it neutral – AIs can be dangerous, they can give a false or biased picture of data to their operators, they can give unwarranted access to dangerous knowledge, they can have big environmental impacts in terms of the resources they use, they can distort markets, they can be used for surveillance and control, they can create misinformation, they can enhance weaponry. All these are non-neutral downsides that are latent in the tech. But they are tools, not consciousnesses, and how they perform depends to a great deal on how they are built and deployed, by their users.

AI safety is therefore greatly enhanced by high-quality AI audit and the application of ethics to design principles. At Securys we create and promote such principles and believe that AI built in accordance with and governed by them is very likely to prove a valuable technology that will improve human experience and capabilities for the better.

Current and emerging regulatory landscape

GDPR

The existing GDPR already governs AI in a number of ways:

Personal data must be processed fairly. This is a core principle of the GDPR and one that is reflected in the EU AI Act. Perhaps the primary concern in regulating AI is to ensure fairness – by avoiding bias, by using complete and accurate information and by offering data subjects appropriate rights of appeal and redress.

Personal data must be processed for specific and limited purposes. Many machine learning systems are trained using data originally collected for other purposes and/or once trained are deployed to process data that has been collected for other purposes. Unless the use of data for training or the subsequent processing by machine learning systems was disclosed to the data subject **at the time the data were originally collected** this further use must be disclosed; it must also be separately justified and may be subject to challenge by the data subject. In some cases, this "extension of purpose" can be unlawful.

Processing of personal data must be properly explained to the data subject ("the principle of transparency"). As noted earlier – and expanded upon below – this can be very difficult in practice when using machine learning systems. It is important to note that the explanation must be demonstrably comprehensible to the data subject, which presents additional challenges when processing data on children or on data subjects with limited education or literacy skills.

Rights of access, rectification and erasure apply to personal data across the AI lifecycle, subject to the relevant legal conditions and exceptions. Training datasets, prompts, logs, retrieval stores and trained artefacts must be assessed separately.[xviii]

Deleting source records does not necessarily remove memorised information from trained weights. Equally, not every statistical influence is identifiable personal data. Organisations should assess recoverability, identifiability and effective remedies rather than assume either that all data are embedded permanently or that all models are anonymous.[xiv, xvii]

Personal data should only be retained for so long as is necessary; it can become challenging to demonstrate the necessity of retaining a particular data subject's information in a model if that data subject themself is no longer involved in any way in the processing being performed. Again, this can present a challenging problem given the difficulty in removing specific data from a trained model.

EU GDPR Article 22 restricts solely automated decisions producing legal or similarly significant effects, subject to its exceptions and safeguards. This should not be conflated

with the amended UK regime: UK GDPR Articles 22A-22D now define significant automated decisions and their restrictions and safeguards.

For the UK, Article 22C requires information about significant solely automated decisions, opportunities to make representations, human intervention and contestation. The EU GDPR position remains distinct.[v]

The UK framework is more permissive for decisions outside Article 22B restrictions, but still requires a lawful basis, fairness, transparency and safeguards. Significant solely automated decisions involving special-category data remain restricted, and decisions relying on the recognised-legitimate-interests basis in Article 6(1)(ea) are also restricted.[v]

EU AI Act: current obligations and implementation

Regulation (EU) 2024/1689 was adopted on 13 June 2024, published in the Official Journal on 12 July 2024 and entered into force on 1 August 2024. It establishes a harmonised, risk-based framework for the development, placing on the market, putting into service and use of AI systems in the European Union. Its stated purposes include promoting human-centric and trustworthy AI while protecting health, safety, fundamental rights, democracy, the rule of law and the environment.

The Act applies through a phased implementation timetable rather than taking effect on a single date. Its provisions include prohibitions on specified AI practices, requirements for high-risk AI systems, transparency obligations for certain AI systems and AI-generated content, obligations for providers of general-purpose AI models, market-surveillance arrangements and enforcement powers. The extent of an organisation's obligations depends upon the nature and risk classification of the system and upon whether the organisation acts as a provider, deployer, importer, distributor, product manufacturer or another participant in the AI value chain.

The Act's definition of an AI system focuses on machine-based systems designed to operate with varying levels of autonomy and which may exhibit adaptiveness after deployment. Such systems infer from their inputs how to generate outputs including predictions, content, recommendations or decisions capable of influencing physical or virtual environments. This definition should replace descriptions based principally on whether a system is heuristic, statistical or rules-based. Classification should be based on the enacted definition, the system's intended purpose and the particular context in which it is used.

For organisations, the immediate practical priority is therefore not simply to determine whether they "use AI". They should identify the role they perform under the Act, maintain an inventory of relevant systems and models, determine which systems are prohibited, high-risk or subject to transparency requirements, and establish evidence

showing how applicable obligations are being met. Procurement and supplier-management processes should address the information, documentation and cooperation needed from upstream providers.

Definition, Scope and Prohibited Practices

Article references below refer to Regulation (EU) 2024/1689 as amended, including the July 2026 Digital Omnibus amendments, rather than earlier drafts. The consolidated text is a reference tool without legal effect and should be read alongside the authentic Official Journal acts and the Commission implementation timeline.[i, ii]

An AI system is a machine-based system designed to operate with varying autonomy, potentially adaptive after deployment, that infers from inputs how to generate predictions, content, recommendations or decisions influencing physical or virtual environments [A3(1)]. The old list of techniques is not the enacted definition. Scope extends to relevant providers, deployers, importers, distributors and product manufacturers, including specified third-country actors whose outputs are used in the Union [A2].[i]

Exclusions include specified military, defence and national-security uses, sole-purpose scientific research and development, and pre-market research and testing, but not real-world testing. Personal non-professional deployer activity is excluded. Free and open-source systems are not generally excluded where they are high-risk or fall under Articles 5 or 50. Sectoral product rules interact with the Act; regulated products are not simply outside its scope.[i]

Article 5 prohibits harmful subliminal, manipulative or deceptive techniques; exploitation of vulnerabilities linked to age, disability or social or economic situation; specified social scoring; criminal-risk prediction based solely on profiling or personality characteristics; untargeted facial-image scraping; workplace and educational emotion inference, subject to medical or safety exceptions; specified sensitive biometric categorisation; and real-time remote biometric identification for law enforcement in public spaces, subject to tightly defined exceptions and safeguards. The statutory harm thresholds and exceptions must be applied, not replaced by blanket statements. These original prohibitions have applied since 2 February 2025.[i, ii]

High-Risk Classification

Article 6(1) covers AI that is a safety component of, or itself constitutes, a product covered by Annex I legislation where the product requires third-party conformity assessment. Article 6(2) covers Annex III uses, including specified biometrics, critical infrastructure, education, employment, essential services, law enforcement, migration and justice. The Article 6(3) exception requires the statutory conditions and documentation; Annex III systems profiling natural persons remain high-risk. A general-purpose model is not automatically a high-risk system: its downstream intended use matters.[i]

General-Purpose AI Models

A general-purpose AI (GPAI) model is distinct from the downstream AI system into which it is integrated. Chapter V creates model-provider obligations independently of the high-risk system regime. Article 53 requires technical documentation, information for downstream system providers, an EU copyright-compliance policy and a sufficiently detailed public summary of training content using the AI Office template.[i, iii]

The qualified free and open-source exception in Article 53(2) concerns the documentation and downstream-information duties only. It does not remove copyright-policy or training-summary duties and does not apply to GPAI models with systemic risk. Open weights alone should not be assumed to satisfy every condition.[i]

Articles 51 and 52 govern systemic-risk classification and notification. Article 55 adds model evaluation, adversarial testing, assessment and mitigation of systemic risks, serious-incident tracking and reporting, and adequate cybersecurity protection for the model and physical infrastructure.[i]

The voluntary GPAI Code of Practice was published on 10 July 2025 and endorsed as an adequate compliance tool. Its chapters address transparency, copyright, and safety and security; the last concerns systemic-risk models. The Code supports demonstration of compliance but does not replace the statutory obligations.[iii]

Using a model does not automatically make an enterprise its provider. Significant modification may create GPAI-provider responsibilities under Commission guidance; rebranding, substantial modification or changed intended purpose can also create high-risk system-provider responsibilities under Article 25. Assess model and system roles separately before fine-tuning, integration or release.[i, iii]

Provider Requirements of High-Risk Systems

High-risk system requirements cover lifecycle risk management [A9]; appropriate data governance and dataset quality [A10]; technical documentation [A11]; automatic logging [A12]; instructions and transparency to deployers [A13]; effective human oversight [A14]; and accuracy, robustness and cybersecurity [A15]. These are requirements for the complete system, not simply for a base model.[i]

Article 16 requires providers to ensure conformity, identify themselves, maintain a quality-management system [A17], keep documentation [A18] and controlled logs [A19], perform the relevant conformity assessment [A43], issue an EU declaration [A47], affix CE marking [A48], register where required [A49], take corrective action [A20] and cooperate with authorities [A21]. Accessibility requirements also apply.[i]

Data governance must address collection and origin, preparation, assumptions, representativeness and bias. The amended exceptional permission to process special-category data for bias detection and correction is in Article 4a and requires strict

necessity and safeguards; it is not a general permission to collect sensitive data for AI development.[i]

Conformity assessment is not universally an independent audit: the route depends on Article 43, the Annex III category, applicable standards and product legislation. Post-market monitoring [A72] and serious-incident reporting [A73] remain important after release.[i]

Deployer Obligation and Human Oversight

The Act uses deployer for an organisation using an AI system under its authority. Once the relevant high-risk requirements apply, Article 26 requires use in accordance with instructions, oversight by people with appropriate competence, training and authority, and relevant and sufficiently representative input data where the deployer controls it.[i]

Deployers must monitor operation, inform providers of relevant risks, suspend use where required and escalate serious incidents to the relevant parties and authorities. Automatically generated logs under their control must normally be retained for at least six months, unless other applicable law provides otherwise. This does not authorise unlimited retention of personal data.[i]

Employers must inform affected workers and their representatives before workplace use of high-risk AI. Specified public deployers must comply with registration requirements. Deployers of Annex III systems making or assisting decisions about natural persons must inform affected people they are subject to the system. [A26(7)-(11)] [i]

Human oversight must be operational: reviewers need time, information, competence and authority to disregard or reverse outputs or stop operation. As a recommended control, organisations should test whether review changes outcomes when errors are introduced, rather than treat nominal approval as meaningful involvement.[i, iv]

Fundamental Rights Assessments & AI Literacy

Article 27 requires a fundamental rights impact assessment before first deployment of relevant Annex III high-risk systems by bodies governed by public law, private entities providing public services, and deployers of the specified creditworthiness and life or health insurance systems in Annex III 5(b) and (c). The critical infrastructure category in Annex III point 2 is excepted. This is not a universal obligation for every private-sector deployer.[i]

The assessment covers processes, use periods and frequency, affected people, risks to rights, human oversight, governance and complaint mechanisms. It must be updated when relevant elements change and notified to the market-surveillance authority as required. A GDPR DPIA remains a separate assessment: the amended Article 27 allows relevant parts or cross-references to be reused, not the omission of uncovered fundamental rights risks.[i]

AI literacy is a separate duty, not postponed until the high-risk regime applies. Article 4 has applied since 2 February 2025. As amended, providers and deployers must take measures supporting the development of literacy among staff and others operating or using AI on their behalf, tailored to knowledge, experience, context and affected people. It does not require a guaranteed literacy level for every individual.[i, ii]

Our recommended evidence includes role-specific training, instructions, competence checks, escalation routes and records of refreshers after material system changes. Procurement, technical, oversight and end-user roles should not all receive the same generic training.

Transparency & Synthetic Content [A50]

Providers of systems interacting directly with people must ensure that people are informed they are interacting with AI unless this is obvious to a reasonably well-informed, observant and circumspect person in the circumstances, subject to the statutory exceptions. Providers of systems generating synthetic audio, image, video or text must support machine-readable marking and detection of artificial generation or manipulation, with the qualifications in Article 50(2).[i]

Deployers of emotion-recognition or biometric-categorisation systems must inform exposed individuals. Deployers generating or manipulating deepfake image, audio or video must disclose that fact; evident artistic, creative, satirical or fictional works have a tailored disclosure rule. These requirements do not make an otherwise prohibited use lawful.[i]

AI-generated or manipulated text published to inform the public on matters of public interest requires disclosure, subject to the law-enforcement exception and the qualification for human review or editorial control where a person holds editorial responsibility. The rules do not require every internal AI-assisted draft to carry a public label.[i]

The main transparency duties apply from 2 August 2026. Under Article 111(4), providers of synthetic-content-generating AI systems placed on the market before 2 August 2026 must comply with Article 50(2) by 2 December 2026. This transition does not postpone the other Article 50 duties.

Our recommended controls combine visible disclosures, machine-readable metadata, appropriate watermarking, preservation of provenance through export and publishing, and testing of whether labels survive transformations.[i, ii, xiv]

Implementation Timeline

Position as at 6 October 2026. Entry into force: 1 August 2024. Under Article 113, the delayed high-risk dates concern Chapter III, Sections 1-3, except Article 6(5), rather than every high-risk-related provision. Further amended Article 5 provisions apply from 2 December 2026. Existing-system transitions must be assessed individually.[i-iii]

Obligation	Responsible role	Application	Key qualification
Prohibitions / literacy	Providers and deployers	2 Feb 2025	Definitions and original Article 5 prohibitions; amended Article 4 requires literacy-support measures.
GPAI	Model providers	2 Aug 2025	Pre-existing models: comply by 2 Aug 2027; enforcement powers from 2 Aug 2026.
Transparency	Relevant providers / deployers	2 Aug 2026	Article 50; Article 111(4) gives relevant pre-existing system providers until 2 Dec 2026 for Article 50(2) only.
Regulatory sandboxes	Member States	2 Aug 2027	At least one national AI regulatory sandbox operational.[ii]
Stand-alone high-risk	System providers	2 Dec 2027	Annex III: requirements, documentation, conformity assessment and monitoring. Article 113: Chapter III, Sections 1-3, except Article 6(5).
Stand-alone high-risk	Deployers	2 Dec 2027	Article 26 duties and Article 27 assessments where required; existing-system transitions apply. Article 113: Chapter III, Sections 1-3, except Article 6(5).
Product high-risk	Relevant providers / deployers	2 Aug 2028	Article 113: Chapter III, Sections 1-3, except Article 6(5), for Article 6(1) systems; check Annex I and Article 2 qualifications.

Transitional Arrangements and Enforcement

Article 111 contains specific transitions for existing systems and models. Under Article 111(2), operators of relevant high-risk systems placed on the market or put into service before the applicable Chapter III date are subject to the Regulation only if those systems undergo significant changes in design from that date, without prejudice to Article 5. In any case, providers and deployers of high-risk systems intended for use by public authorities must comply by 2 August 2030. The separate rules for large-scale IT systems in Article 111(1) must also be checked. GPAI models marketed before 2 August 2025 must comply by 2 August 2027.[i, iii]

Authorities can require corrective action, restrict or prohibit supply and require withdrawal or recall where the statutory conditions are met. Under Article 99, the principal maximum penalties are EUR 35 million or 7% of worldwide annual turnover for prohibited practices, EUR 15 million or 3% for specified other breaches, and EUR 7.5 million or 1% for incorrect, incomplete or misleading information. The applicable calculation includes special treatment for SMEs. Article 101 separately provides GPAI-provider fines of up to 3% of worldwide annual turnover.[i]

AI Act compliance does not displace GDPR, consumer protection, employment or product safety obligations. Organisations should map applicable duties by legal role, intended use, jurisdiction and commencement date, and retain evidence of both classification and compliance.[i]

United Kingdom

UK AI and Automated Decision Framework

The UK uses a regulator-led, sectoral approach rather than the EU AI Act model. The 2023 white paper is policy context, not a comprehensive statutory permission to use AI. Existing data-protection, equality, consumer-protection and sectoral duties remain relevant. For public bodies, applicable public-law fairness obligations also need separate assessment.[v, xviii]

The Data (Use and Access) Act 2025 received Royal Assent on 19 June 2025. Section 80 substituted UK GDPR Articles 22A-22D for Article 22. Those amendments commenced on 5 February 2026 under SI 2026/82, regulation 2(j). Regulation 5 preserves the earlier rules for decisions taken before that date to which the specified former safeguards applied. The EU GDPR Article 22 regime has not been amended by this legislation.[v, vi]

Article 22A defines solely automated processing by the absence of meaningful human involvement and defines significant decisions by legal or similarly significant effects. Article 22B retains restrictions for decisions involving special category data and decisions relying on Article 6(1)(ea). Article 22C requires information, representations, human intervention and contestation safeguards. Article 22D enables further regulations.[v]

Organisations should assess the actual decision workflow, not just whether a human appears somewhere in it. Our recommended controls include clear review authority, relevant evidence for reviewers, documented reasons for overrides, accessible challenge routes and checks for automation bias.[v, xiv]

UK Implementation Priorities

For each significant decision, record whether there is meaningful human involvement, whether special-category data or the recognised legitimate-interests basis is used, the applicable lawful basis, and the Article 22C safeguards. Reassess DPIAs, privacy notices, contracts and complaint processes when the decision workflow changes. These are recommended implementation steps, not a replacement for the statutory test.[v, vi]

Parliamentary proposals, including an Artificial Intelligence (Regulation) Bill, must be distinguished from enacted law. They are not presented here as the core UK regime. ICO materials are being updated following the 2025 Act; the March-May 2026 consultation concerned draft guidance, so older Article 22 guidance should not be used to override the amended statute.[v, xix]

The United Kingdom continues to apply a regulator-led and context-sensitive approach to AI governance. The five principles identified in the government's pro-innovation framework remain useful organising concepts: safety, security and robustness; appropriate transparency and explainability; fairness; accountability and governance; and contestability and redress. However, that policy framework must now be read

alongside subsequent legislation, particularly reforms to UK data-protection law.

The Data (Use and Access) Act 2025 received Royal Assent on 19 June 2025. It amended, rather than replaced, the UK GDPR and the Data Protection Act 2018 and introduced substantial changes to the rules governing automated decision-making.

Under the amended framework, a decision is based solely on automated processing where there is no meaningful human involvement in taking it. A decision is significant where it produces a legal effect concerning the individual or has a similarly significant effect. The provisions distinguish decisions involving special-category personal data from other significant automated decisions and specify circumstances in which such processing is restricted.

The reforms create a more permissive framework for significant decisions based solely on automated processing in some circumstances, but they do not remove the need for safeguards. Where the statutory provisions apply, organisations must provide information about significant decisions, enable affected individuals to make representations and challenge the decision, and provide a route to human intervention.

Meaningful human involvement requires more than nominal approval of an automated output. Organisations should define who is authorised to intervene, ensure that reviewers have relevant competence and sufficient information, and verify that they have the practical ability to reconsider and change an outcome. Records should demonstrate the respective roles of the system and the human reviewer.
The Information Commissioner's Office has been updating its guidance to reflect the 2025 Act. Existing ICO transparency and explainability materials should therefore be used with care where they indicate that they are under review. The ICO has also consulted on updated guidance addressing automated decision-making and profiling following the Act.

For organisations subject to both UK and EU requirements, a single classification of an AI use case may not be sufficient. The same system may fall within UK data-protection rules on automated decision-making, the EU AI Act's role-based and risk-based requirements, sector-specific regulation, equality legislation and contractual or professional duties. Governance should map these overlapping requirements rather than treating compliance with any one regime as conclusive.

AI Safety Summits

In November 2023 the UK government hosted a landmark AI Safety Summit at Bletchley Park. The event brought together international officials, leading AI companies, and experts to discuss the risks and mitigation strategies for advanced AI development. The participants agreed "the Bletchley Declaration", an agreement to pursue an international management framework for cutting-edge artificial intelligence ("frontier AI") that balances safety and development, to be achieved through prudent negotiation and inclusive, mutually beneficial approaches.

Also announced was the establishment of a new UK-based AI Safety Institute, and a proposal for an International Panel on Artificial Intelligence Safety (independent from political interference) that would be able to inform policymakers and the public.

AI summits provide policy context and should be distinguished from binding legislation and its application dates.

United States

Federal Policy

Executive Order 14110 of 30 October 2023 was revoked by Executive Order 14148 of 20 January 2025, section 2(ggg), and is no longer the current federal organising framework. Executive Order 14179 of 23 January 2025 directs review of actions taken under the revoked order and their revision, suspension or rescission where inconsistent with the new policy. Federal executive policy is distinct from legislation and does not remove otherwise applicable sectoral or general law.[vi, vii]

The July 2025 AI Action Plan emphasises innovation, infrastructure and international leadership.[x] Executive Order 14365 of 11 December 2025 seeks a minimally burdensome national framework and directs action concerning restrictive state laws. This federal-state tension is not a basis for assuming that all state AI legislation has been displaced.[ix]

The 2022 Blueprint for an AI Bill of Rights is historical, non-binding policy material, not an enacted federal AI bill. Organisations should assess applicable consumer, civil-rights, employment, privacy and sectoral law separately from executive policy and voluntary NIST guidance.[vii-xii]

Federal enforcement, guidance and legislative direction

The Equal Opportunity Employment Commission has issued **guidance on ensuring that AEDTs do not violate the Americans with Disabilities Act** by screening out disabled applicants. As with much present US federal enforcement the focus is on interpretation of existing law and associated publication of guidelines[3]. More widely but along the same lines, the Federal Trade Commission (FTC) has published guidance[4] on using **AI in compliance** with a variety of existing legislation for which the

3 https://www.jacksonlewis.com/sites/default/files/docs/EEOC-TechnicalAssistanceADA-AI.pdf

4 https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai

FTC has enforcement powers; importantly this includes the FTC Act which is the general basis for privacy enforcement in the US. The key provisions in the guidance documents are that companies should:

- Make sure AI is trained using data sets that are representative, and do not "miss […] information from particular populations."
- Test AI before deployment – and periodically thereafter – to confirm it works as intended and does not create discriminatory or biased outcomes.
- Ensure AI outcomes are explainable, in case AI decisions need to be explained to consumers or regulators.
- Create accountability and governance mechanisms to document fair and responsible development, deployment, and use of AI.

The FTC has been enforcing against misuse of AI; generally, this has been based either on absence of evidence of proper consent or misuse of data for model training even when data subjects have refused consent.

Representative State Developments

California signed SB 53, the Transparency in Frontier Artificial Intelligence Act, on 29 September 2025. It introduces frontier-model transparency and safety oversight. California AB 2013 requires covered generative-AI developers to publish training-data documentation from 1 January 2026, including specified systems or substantial modifications made available to Californians. These are distinct obligations with their own scope and exemptions.[xi, xii]

California also has synthetic-content transparency measures, including the California AI Transparency Act (SB 942). The current SB 942/AB 853 amended text and commencement position were not finalised by the time of this publication. Applicability, thresholds and exemptions must be checked against the current statutory text rather than assumed to match EU Article 50.[xx]

A number of states including New York, Illinois and Maryland have introduced legislation regulating the use of **AI in employment decision-making**. This includes a requirement (in New York) for **annual auditing of so-called Automated Employment Decision Tools (AEDT)** for bias; the audit results must be published. Note that where AEDTs are regulated this is applied to decision-support systems in addition to actual automated decision making.

California, Colorado, Connecticut, Virginia, Utah, Delaware, Iowa, Indiana, Montana, Oregon, Tennessee, Texas and New Hampshire have all passed general state privacy laws. In general, these laws all grant residents some of the rights familiar from GDPR; specifically, they include provisions for **enhanced transparency regarding the use of AI**. The transparency provisions vary; they contain elements of existing GDPR provisions and the new proposals in the EU AI regulation. The consensus is that:

- Data subjects must be informed when AI is being used to process their data
- The algorithm and the work done to eliminate bias and ensure fairness must be

explained

- Where profiling forms part of the processing, it must be justified
- Where material decisions affecting the data subject are made by AI, this must be separately identified. Some states also offer an opt-out from automated decision-making similar to the GDPR Article 22 requirement for human intervention to be made available.

State laws also introduce a requirement for a data privacy impact assessment; the Colorado law goes further in requiring a specific **AI impact assessment** which includes provisions for documentation of:

- Sources and uses of training data
- Logic and statistical methods used to create the AI model
- Evaluations of accuracy and reliability
- Evaluations of fairness and "disparate impact"
- EU compliance can provide useful controls and evidence but is not a legal superset of US requirements. State definitions, notices, employment rules, remedies and exemptions require separate analysis.[xi, xii]

Colorado SB26-189, signed on 14 May 2026, repeals and re-enacts the earlier framework around covered automated decision-making technology materially influencing consequential decisions. From 1 January 2027, its requirements include developer documentation, consumer notices, information following adverse outcomes, correction and meaningful human review, and compliance records retained for at least three years. It does not create a new private right of action.[xxi]

NIST AI Risk Management Framework

The National Institute of Standards and Technology released AI RMF 1.0 on 26 January 2023. It is a final, voluntary framework, not a draft regulation. Its core functions are Govern, Map, Measure and Manage. NIST is revising the framework; work on a successor does not make version 1.0 a draft.[xv]

NIST AI 600-1, the Generative Artificial Intelligence Profile, was published on 26 July 2024. It supplements the framework with generative-AI risks including confabulation, privacy, environmental impacts, harmful bias, human-AI configuration, information integrity and security, intellectual property, and value-chain integration.[xiv]

Our recommended application is to establish ownership and policy under Govern, identify context and affected people under Map, evaluate risks and controls under Measure, and prioritise mitigation, monitoring and withdrawal under Manage. NIST guidance supports evidence-based governance but does not establish legal compliance by itself.[xv]

China

PIPL

The Personal Information Protection Law[5] grants data subjects similar rights to those found in the GDPR, albeit only with regard to the private sector. These include provisions similar to the GDPR's Article 22 governing automated decision-making, notably:

- An obligation to ensure that automated systems operate fairly and transparently, without bias and without unreasonable differential treatment.
- Rights for data subjects significantly affected by AI systems both to an explanation of algorithms and to require human review of decisions.
- Completion of a data protection impact assessment for the use of personal data in automated decision-making.

New Generation AI Ethics

The Ministry of Science and Technology published a *New Generation Artificial Intelligence Code of Ethics*[6] in 2021. The intent of the code is to ensure that AI stays under "meaningful human control" and it introduces 6 fundamental and 18 operational and management ethical practices ("norms") that must be observed by developers and users of AI systems. These are summarised below, based on an English translation from the International Research Center for AI Ethics and Governance[7].

Fundamental ethical norms

1. **Enhancing the wellbeing of humankind.** This covers human rights, common values, national and regional ethical norms; also sustainable development, ecology and the public interest.
2. **Promoting fairness and justice.** Protect the legitimate rights and interests of all relevant stakeholders, promote fair sharing of the benefits of AI in the whole society, and promote social fairness and justice, and equal opportunities.
3. **Protecting privacy and security.** Respect the rights of personal information.
4. **Ensuring controllability and trustworthiness.** Ensure that AI is always under meaningful human control and that humans have the rights to choose whether to accept AI services and decision-making.

5 https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/

6 http://www.most.gov.cn/kjbgz/202109/t20210926_177063.html

7 https://ai-ethics-and-governance.institute/2021/09/27/the-ethical-norms-for-the-new-generation-artificial-intelligence-china/

5. **Strengthening accountability.** Clarify the responsibilities of all relevant stakeholders. Establish an accountability mechanism in AI related activities.
6. **Improving ethical literacy.** Popularize knowledge related to AI ethics, objectively understand ethical issues, and do not underestimate or exaggerate ethical risks.

Norms of management

1. **Promotion of agile governance.** Respect the law for development of AI, fully understand the potential and limitations of AI, continue to optimize the governance mechanisms and methods of AI.
2. **Active practice.** Comply with AI related laws, regulations, policies and standards, actively integrate AI ethics into the entire management process.
3. **Exercise and use power correctly.** Clarify the responsibilities and power boundaries of AI-related management activities and standardize the conditions and procedures of power operations.
4. **Strengthen risk preventions.** Enhance risk awareness, carry out systematic risk monitoring and evaluations in development and operation.
5. **Promote inclusivity and openness.** Pay full attention to the rights and demands of all stakeholders related to AI, encourage the application of diverse AI technologies to solve practical problems.

Norms of research and development

1. **Strengthen the awareness of self-discipline.** Actively integrate AI ethics into every phase of technology research and development, do not engage in AI research and development that violates ethics and morality.
2. **Improve data quality.** Strictly abide by data-related laws, standards and norms throughout the data lifecycle.
3. **Improve the completeness, timeliness, consistency, normativeness and accuracy of data.**
4. **Enhance safety, security and transparency.** Improve transparency, control, resilience and defensibility of AI systems and implement verifiable and reliable auditing and supervision.
5. **Avoid bias and discrimination.** Fully consider the diversity of demands, avoid potential data and algorithmic bias, and strive to achieve inclusivity, fairness and non-discrimination.

Norms of supply

1. **Respect market rules.** Abide by the regulations for market access, competition, and trading activities, actively maintain market order, and create a market environment conducive to the development of AI.
2. **Strengthen quality control.** Strengthen the quality monitoring and the evaluations on

the use of AI products and services, avoid infringements on personal safety, property safety, user privacy, etc. caused by product defects.

3. **Protect the rights and interests of users.** Users should be clearly informed that AI technology is used in products and services. Users must be able to choose to use or quit the AI mode.
4. **Strengthen emergency protection.** Emergency mechanisms and loss compensation plans and measures should be investigated and formulated. Implement timely monitoring, address user feedback and make corrective actions swiftly.

Norms of use

1. **Promote good use.** Fully consider the legitimate rights and interests of stakeholders, so as to better promote economic prosperity, social progress and sustainable development.
2. **Avoid misuse and abuse.** Respect the rights of relevant entities not to use AI products or services. Avoid improper use, misuse and abuse of AI products and services.
3. **Forbid malicious use.** AI products and services that do not comply with laws, regulations, ethical norms, and standards may not be used, nor may AI be used for illegal activities or to cause a risk to public safety or the public interest.
4. **Timely and proactive feedback.** Participate in the practice of AI ethics and governance, providing prompt feedback to relevant subjects and assistance for solving problems.
5. **Improve useability.** Improve knowledge and usage skills related to AI so as to ensure the safe and efficient use of the technology.

Deep Synthesis

China's Provisions on the Administration of Deep Synthesis of Internet-based Information Services[8] - their "deep fake" law – is already in force and has detailed provisions for every stage of the digital synthesis lifecycle. Notably their provisions also apply to the creation of text as well as image and sound media. The regulations apply both to service providers and to users.

In addition to strengthening the requirement for compliance with the PIPL, the deep synthesis law expects:

- Improved transparency including a requirement for real-identity information authentication system
- Content management and labelling so that synthetic content can readily be identified
- Technical security – essentially the documentation, data governance, bias-elimination and quality management elements of the EU regulation.

This has particular impact on current trends for the deployment of chatbots in customer service, often with no immediate indication that the customer is corresponding with a

machine or with seamless transition from machine to human interaction under the same apparent agent name, and for the use of generated content in marketing, such as the recently announced Levi's campaign using digital models to enhance diversity[9].

China's Measures for Labelling AI-Generated Synthetic Content took effect on 1 September 2025. They cover text, images, audio, video and virtual scenes and distinguish explicit user-perceptible labels from implicit technical labels. Providers must include specified metadata; dissemination services must check labels and apply appropriate notices. Users publishing generated content must declare it, and malicious removal or alteration of labels is prohibited. Digital watermarking is encouraged, not universally mandated by these Measures.[xiii]

Other Jurisdictions

Other countries around the world are looking to follow the lead set by the EU in particular, as was the case in data protection with GDPR. Singapore has published a proposed governance framework for the use of generative AI[10], and Peru[11] has become the first country in Latin American to enact AI-specific legislation (Law No. 31814).

International standards and frameworks

ISO/IEC 42001:2023

ISO/IEC 42001:2023 specifies requirements for establishing, maintaining and improving an AI management system. It supports structured governance for organisations developing or using AI but does not replace legal obligations.

8 http://www.cac.gov.cn/2022-12/11/c_1672221949318230.htm

9 https://www.businessoffashion.com/articles/technology/levis-will-begin-testing-ai-generated-models/

10 nais2023.pdf (go.gov.sg)

11 Law No. 31814

OECD Framework for the Classification of AI Systems

A tool developed by the Organisation for Economic Co-operation and Development (OECD) to help policymakers, regulators, legislators, and others understand and analyse different types of AI systems in line with the OECD AI Principles, which promote responsible and ethical AI development and deployment. The framework looks at systems along five key dimensions: People & Planet, Economic Context, Data & Input, AI Model, and Task & Output. Each dimension has its own set of properties and attributes that help assess the policy implications of the specific AI system under consideration.

The OECD AI Principles were updated in May 2024 to address developments including general-purpose and generative AI, privacy, intellectual property, safety and information integrity. These frameworks should be mapped to applicable duties, not treated as equivalent to legislation.[xxiii, xxiv]

Assessment List for Trustworthy Artificial Intelligence (ALTAI)

A self-assessment checklist based on the seven key requirements of Trustworthy AI outlined in the Ethics Guidelines for Trustworthy AI developed by the European Commission's High-Level Expert Group on AI (AI HLEG). The key requirements are: Human Agency and Oversight; Technical Robustness and Safety; Privacy and Data Governance; Transparency; Diversity; Environmental and Societal Well-being; Accountability. ALTAI is a self-assessment tool, so its effectiveness depends on the honesty and objectivity of the user.

Practical Considerations

Preamble

This paper is necessarily focused on mitigating risks, especially regulatory risks, arising from the use of machine learning and AI. However, the intention is not to discourage the adoption of machine learning tools where these can add value without risking harm to individuals or groups. Many opportunities exist for positive deployment of AI, especially in safety systems where consistent and rapid responses are a key component of reducing risks of physical harm. For instance, while in the examples below this paper highlights risks arising from prototype self-driving systems on public highways, there is good evidence that the more controlled environments of industrial sites are suitable use cases for automation[10].

Similarly, while there are risks associated with AI generated text and image content – as highlighted by the Chinese regulation of deep synthesis – there are also significant productivity gains to be found provided that users understand the limitations. As a general rule, therefore, perhaps the key mitigation for all AI risks is adequate training and awareness-raising to ensure that both commissioners, developers and users of machine learning systems are properly aware of the risks, regulations and constraints, as well as the benefits and opportunities.

Minimum governance and compliance measures

An organisation developing, procuring, providing or deploying AI should establish a proportionate AI governance programme incorporating, at a minimum, the following measures:

1. **AI inventory and ownership.** Maintain a current record of AI systems, embedded AI functionality and general-purpose models. Record intended purpose, business owner, provider, model or service used, affected individuals, personal-data processing, connected data sources, tools, deployment locations and lifecycle status.
2. **Role and jurisdiction assessment.** Determine whether the organisation acts as provider, deployer, importer, distributor or product manufacturer under the EU AI Act, and record every jurisdiction and sectoral regime that applies.

10 https://www.roboticsbusinessreview.com/opinion/the-levandowski-pardon-and-the-pursuit-of-autonomous-vehicle-technology/

3. **Risk classification.** Identify prohibited practices, high-risk systems, transparency obligations and general-purpose-model obligations. Separately assess data-protection, equality, consumer, employment, cybersecurity, intellectual-property, safety and professional risks.

4. **Impact assessment.** Coordinate the AI assessment with any required data protection impact assessment, fundamental-rights impact assessment, security assessment or conformity-assessment process. Assess the complete sociotechnical system rather than the model in isolation.

5. **Supplier and model due diligence.** Obtain sufficient information about model capabilities and limitations, evaluation results, training-data transparency, copyright policy, security, incident handling, data use, retention, locations, subprocessors and significant model changes.

6. **Data and retrieval governance.** Control the quality, provenance, representativeness, lawful use and permitted reuse of training, testing, fine-tuning and retrieval data. Apply access controls and data-minimisation measures to information made available to prompts, retrieval systems, memory or agents.

7. **Evaluation before deployment.** Test the system against documented requirements and foreseeable misuse. Evaluation should address accuracy, robustness, bias, privacy, security, explainability, accessibility, hallucination, prompt injection, tool misuse and failure under unusual conditions.

8. **Human oversight.** Define decisions that require human approval, provide reviewers with authority and competence to intervene, and avoid oversight arrangements in which the human merely endorses the automated output.

9. **Transparency and redress.** Provide appropriate notice where individuals interact with AI or are affected by AI-supported decisions. Explain the relevant purpose and effect of the processing and provide workable mechanisms for challenge, correction and human review.

10. **AI literacy.** Provide role-appropriate education for personnel who develop, procure, configure, operate or oversee AI. Training should cover capabilities, limitations, applicable policies, escalation routes and the particular risks of the systems they use.

11. **Logging, monitoring and incident management.** Preserve records necessary to reconstruct material decisions and actions. Monitor system performance and significant changes, and establish processes for escalation, containment, investigation, notification and remediation.

12. **Change control and withdrawal.** Reassess systems when models, data sources, prompts, tools, intended purposes or operating environments change. Maintain the ability to suspend, replace or roll back a system that no longer performs safely or lawfully.

    These measures also support the NIST AI Risk Management Framework's four operational functions of Govern, Map, Measure and Manage, and allow generative-AI risks to be integrated into an organisation's wider risk-management system.

Agentic AI and Tool-use Controls

Our recommended agent controls assign a named organisational owner, apply least privilege and separate read, write, approval and execution permissions. Require human approval for high-impact or irreversible actions; protect credentials; test indirect prompt injection and malicious retrieved content; and log prompts, tool calls, data accessed, outputs and actions proportionately to privacy and retention requirements. Provide immediate suspension and permission-revocation mechanisms. Evaluate complete action chains, including downstream effects, not only individual model responses.[xiv, xxii]

Environmental and Infrastructure Proportionality

NIST identifies environmental impacts from the compute used to train and operate generative AI, and recommends comparing expected benefits with energy, water, hardware and infrastructure costs, considering smaller or task-specific models, and recording compute location and supplier governance. Organisations should also assess resilience, exit options and concentration risk where critical workflows depend on a small number of model or cloud providers, and reassess when usage or model scale changes.[xiv]

Over-estimation of capability risk

An overarching danger of using AI is the tendency to exaggerate the capabilities of machine learning models; take for instance the case where a Google engineer claimed sentience for their model[11], or the extensive hype surrounding the capabilities of ChatGPT[12]. This leads to users and other involved persons mistakenly treating machine learning models as sentient general AIs. When this happens, users often fail to account for the limitations of machine learning perception and sensitivity to context.

> *We need only remember the speed with which Microsoft's incorporation of ChatGPT technology into a function of its Bing search engine devolved into chaos to see an example of this risk*[13].

Recommendation

Ensure that proposals for use of AI accurately reflect real capabilities, and that limitations are properly communicated to all stakeholders and users.

11 https://edition.cnn.com/2022/07/23/business/google-ai-engineer-fired-sentient/index.html#:~:text=Blake%20Lemoine%2C%20a%20software%20engineer,thousands%20of%20messages%20with%20it.

12 What is going on with ChatGPT? | Arwa Mahdawi | The Guardian

13 https://www.popularmechanics.com/technology/robots/a43017405/microsoft-bing-ai-chatbot-problems/

Case failure

Machine learning models are dependent on both training data and nature of inputs. Because they are not actually sentient, they are not "aware" of their limitations. This can lead to case failure, where a condition that is apparent to a human observer is not detected by the machine learning model.

> *Tesla's "self-driving" feature relies on machine vision processing. There have been several cases, one unfortunately fatal*[14]*, where the vision system could not adequately identify a white object as a physical obstruction – it being theorised that the model falsely identified the white area as a cloud or light sky instead – where this would have been obvious to a human.*

Recommendation

Designers should ensure, especially in safety-critical systems, that the compass of machine decision-making is properly understood and identify and mitigate the possible consequences of case failure. See also Zerili, Knott, Maclaurin & Gavaghan, *Algorithmic Decision-making and the Control Problem, Minds and Machines 29, 2019*[15]

14 https://arstechnica.com/cars/2019/05/feds-autopilot-was-active-during-deadly-march-tesla-crash/

15 https://link.springer.com/article/10.1007/s11023-019-09513-7

Reliance risk

Machine learning models can deliver tremendous efficiency gains, both from simple speed of operation and from tireless performance. The danger is that business models are constructed that do not allow for human intervention without impossible reduction in productivity. This in turn can lead to serious consequences for natural people such as e.g. when automated detection of suspicious financial activity leads to transaction or account blocking but the appeal process – operated by humans – is unable to address complaints within a reasonable timeframe.

> *This has been a problem for new-entrant financial services firms in particular, who have used machine learning extensively in detection and prevention of fraud and money-laundering. Failure to support this function with sufficient numbers of trained staff meant that customers who were subject to a false-positive result and subsequent account or transaction block were unable to have the decision rectified in an acceptable time*[16].

Recommendation

Ensure that all processes which can have a material effect on natural people have mechanisms for timely review and intervention by human operators, and that such operators are recruited and trained in sufficient numbers. This is also a key regulatory requirement.

16 https://www.altfi.com/article/8953_why-account-closures-are-a-bigger-problem-for-neobanks-compared-to-incumbents

Rigidity risk

Machine learning models do not exercise judgement or provide for contextual variation from learned rules. Human rule- and law-making has historically relied on a degree of flexibility – whether arising from individual judgment or from limitations in enforcement capability – to compensate for the impossibility of writing defined laws or rules that take account of all circumstances. Machine learning models because of their speed and scalability can offer "perfect" enforcement of rules and laws, but this is unlikely to deliver the best social or organisational outcome.

> *Examples here include automated censorship of social media posts, which recently led to an amusing and much publicised case where the British Trust for Ornithology lost access to its Twitter feed following a post about woodcocks[17]. More seriously, the use of machine learning for e.g. traffic enforcement in place of human officers can lead both to situations where mitigating circumstances are ignored and to a resultant loss of public support for traffic rules.*

Recommendation

Where AI systems are to be used to enforce rules, study the consequences of complete and inflexible enforcement before committing to be sure that potential negative social effects do not outweigh the benefits.

17 https://www.bbc.co.uk/news/uk-england-norfolk-64451977

Training failures

At present this is the single greatest area of concern in machine learning. Where "AI" is based on some form of deep learning such as general adversarial networks, the quality of that training largely conditions the success of the model. A range of potential failures can be identified from historical examples:

Training data sourcing

Where training data involves personal data, the data must be sourced ethically and in compliance with relevant data protection laws. This is often harder to do than it might appear. Notably in GDPR the fact that information is, or appears to be, in the public domain does not in itself provide a justification for processing the data; by the same token just because data are available within an organisation does not mean that their use in training a machine learning model is compatible with the purposes for which the data where originally collected.

Clearview, a vendor of facial recognition software, has been fined substantial sums by several different European data protection authorities including the UK, France and Italy, for making use of facial images obtained without the consent of the data subject[18].

Recommendation

Do not use data from existing systems to train new machine learning systems unless the data subjects concerned have been consulted and either informed or consented depending on the assessment of the lawful basis concerned.

Training data bias (aka proxy risk)

Much training data, whether collected from real sources or synthesised, can incorporate inherent biases. Precisely because machine learning models are not sentient, they will replicate this bias in the model. It can be extremely difficult – and in some cases impossible – to remove this bias from the model once trained in. In many cases, attributes that are highly correlated with demographic features, known as proxies, can contribute to algorithmic discrimination.

18 https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2022/05/ico-fines-facial-recognition-database-company-clearview-ai-inc/

> *Amazon was an early adopter of AEDT, building a CV-screening system that sought to identify the best engineers from submitted CVs. The system was trained on the CVs of existing successful engineers. Unfortunately, the vast majority of those engineers were male and the resulting model was extremely adept at identifying and selecting for male CVs even when obvious gender identifiers were removed. Amazon was unable to remove this bias from the model and was forced to retire the system[19].*

Recommendation

Use rigorous data science to ensure that training data does not contain bias and/or to create synthetic training data that properly reflects the desired diversity of outcome. Test systems for bias, including seeking to identify potential proxies that may have been incorporated into the learned model.

Outcome biases

By the same token, machine learning model training works by having the trainer (directly or by dataset or condition selection) teaching the machine the different between right and wrong answers. The machine is, therefore, seeking to replicate the trainer's definition of success by design. Any biases in the trainer that are not detected and corrected before training will therefore be replicated in the model. This can be exacerbated when a feedback loop is established but the human in the loop continues to exhibit bias in confirming results provided by the model.

> *An exhaustive study of this problem (and others mentioned here) can be found in Garvie, Clare, A Forensic Without the Science: Face Recognition in U.S. Criminal Investigations, Center on Privacy & Technology at Georgetown Law (2022)[20].*

Recommendation

Ensure that operators and trainers receive training to determine their own conscious and unconscious biases and that feedback loops are properly monitored – including by the inclusion of known incorrect outputs in order to test outcome bias and feedback loop quality.

19 https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

20 https://mcusercontent.com/672aa4fbde73b1a49df5cf61f/files/2c2dd6de-d325-335d-5d4e-84066159df71/Forensic_Without_the_Science_Face_Recognition_in_U.S._Criminal_Investigations.pdf

Prompts and hinting

There is an additional risk arising from the need for most current machine learning systems to supplement their direct learning with "prompting" or "hinting" – operator-provided overrides that seek to reinforce desired outputs or suppress undesirable ones. This is particularly common in natural language interfaces such as chatbots (including e.g. ChatGPT) but is also seen when attempts are made to compensate for perceived bias in a trained model. The issue is that the hints themselves can embed operator bias, act to restrict open access to information or prevent a continuous learning model from developing properly in response to new input data. By definition, a prompt that states "the following content is offensive" embeds the operator's personal perception into the model, where it will be treated as fact.

> *The COMPAS algorithm (referred to under Transparency risks below) could be prompted to ignore sensitive characteristics like race and gender. However, separate analysis of offender data sets indicates that female recidivism is less likely, so prompting the system to ignore gender would in fact result in unfair outcomes for women*[21].

Recommendation

If prompts are used to control for bias in a model, establish an independent review process and apply rigorous statistical analysis to ensure that the prompts are actually improving fairness rather than simply introducing alternative bias.

21 https://arxiv.org/abs/1701.08230

Correlative error

Deep-learning machine learning models work by finding matching patterns between inputs and outputs. However, one of the problems with transparency (see later) is that understanding and documenting the specific patterns identified by the machine can be effectively impossible. Consequently, there is a danger of assuming that, simply because a model can reliably produce the desired outputs from a set of inputs, it is doing so in an analogous fashion to a human process. This is not the case and, accordingly, can lead to unpredictable and undesirable outputs when presented with new inputs.

> *A system trained to detect pneumonia using X-rays worked well in a single hospital but failed when sent images from other hospitals. It turned out that the system had included metal marker tokens only used in the original hospital in its input data and correlative matching.*[22]

Recommendation

Make sure that correlations are understood. Remove elements from datasets to ensure that these are not part of the model's logic and test early with data from other sources – noting though always the constraints on testing with live personal data.

22 https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683

Query ambiguity, error or bias

Machine learning systems designed to interrogate data answer the question asked. Allied to the over-estimation of capability risk mentioned earlier is the risk that the operator assumes an understanding of context or of intent by the machine that is not present. This can result in operator bias or assumption driving incorrect outputs; in reinforcement of existing assumptions – a confirmation bias problem – or in actual failure, in the sense that the question answered is not the one the operator thought they were asking.

This is also a general observation in statistics and data analytics. In AI in particular the sudden explosion in use of ChatGPT has led to policies banning the publishing of material generated by the program because so many of the answers are incorrect, often as a result of ambiguity or lack of context in the question.

> *A sample from the explanation for such a ban from StackOverflow: "The primary problem is that while the answers which ChatGPT produces have a high rate of being incorrect, they typically look like they might be good and the answers are very easy to produce. There are also many people trying out ChatGPT to create answers, without the expertise or willingness to verify that the answer is correct prior to posting. Because such answers are so easy to produce, a large number of people are posting a lot of answers. The volume of these answers (thousands) and the fact that the answers often require a detailed read by someone with at least some subject matter expertise in order to determine that the answer is actually bad has effectively swamped our volunteer-based quality curation infrastructure."*[23]

Recommendation

Consider use of AI to be a skilled task and ensure that everyone responsible for setting the problems to be solved and questions to be answered is both trained in prompt engineering and has the knowledge to identify incorrect results and positively influence the feedback loop.

23 https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned

Cyber risks

AI security must cover the complete lifecycle and connected application, as well as conventional software, access-control and data-breach risks. NIST AI 100-2e2025 provides a taxonomy of evasion, poisoning, privacy and misuse attacks; NIST AI 600-1 addresses generative-AI information security and component integration.[xiv, xxii]

Direct prompt injection uses attacker-supplied instructions; indirect injection embeds them in external material such as retrieved documents or tool responses. Other relevant threats include poisoned data or models, adversarial examples, malicious fine-tuning, sensitive-data extraction, model inversion, membership inference and model theft. Retrieval sources, vector stores, model repositories, plug-ins and agent frameworks are part of the attack surface.[xxii]

Our recommended controls include least privilege, separation of trusted instructions from untrusted content, validated retrieval sources, isolated execution, credential protection, supplier and dependency checks, and evaluation of tool misuse. Agents should not be able to grant themselves new permissions or expose secrets to external tools. Controls must be tested across text, image, audio and document inputs, not just typed prompts.

Monitor anomalous tool calls, data access and actions; protect logs and model artefacts; define incident escalation and immediate suspension. High-impact or irreversible actions should require independent human approval. Testing should assess chains of actions and realistic adversarial content rather than assume a prompt filter provides complete protection. These are recommended operational safeguards.[xiv, xxii]

> *Both British Airways and Ticketmaster suffered large-scale data breaches – and were fined millions of pounds – as a consequence of insufficiently secure deployment of third-party NLP machine learning models ("chatbots").*[24]

Recommendation

Ensure that information security teams are properly skilled in assessing cyber-risks from machine learning, and avoid the temptation to deploy poorly understood technology

Transparency risks

Deep-learning models can be difficult to interpret, but complexity does not justify abandoning transparency. Organisations should document purpose, data, evaluation, limitations and material decision factors and provide an explanation appropriate to the affected person and use context.[xv, xvii]

Use intrinsically interpretable models where adequate for the task. Where these are unsuitable, our recommended controls include counterfactual testing, feature analysis where meaningful, traceability, system-level evaluation and competent human review. Do not deploy where residual lack of explainability is incompatible with applicable legal or safety requirements. Generated explanations must not be assumed to faithfully reveal a model's internal reasoning. [xiv, xv]

> *Northpointe produces a system, COMPAS, that is widely used in the US to determine parole for criminal offenders. The system has been repeatedly accused of exhibiting systemic racism but the vendor has not provided insight into the operation of the model and the system's users are unable to examine the decision-making logic themselves*[25].

Recommendation

Ensure that vendors provide satisfactory documentation and explanation of algorithmic functions and learning methodologies. For internal development avoid the use of learning techniques and models that become "black boxes" where it is not possible for a human to elucidate the decision-making process.

24 https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

25 https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9685994

Oversight and governance

Not only is emerging regulation clear on the need for enhanced oversight of machine learning, it's also essential as part of demonstrating compliance with existing rules including the GDPR's Article 22 and is central to the conceit of the EU AI regulation, the Chinese AI ethics code and the US blueprint for an AI bill of rights. This includes the need to offer escalation to a human decision-maker – at least for any machine-made decision that has a material effect on the data subject – but more broadly there is an overriding requirement to show fairness towards data subjects. Given the potential issues highlighted in this document it's clear that the risk of acting unfairly – either through undetected bias or by over-reliance on decision making without proper context – is significantly heightened by the use of machine learning. Naturally machine learning can also be used to improve fairness – a model that can be demonstrated not to have bias is likely to act fairly more consistently than humans given the same information.

> *Deliveroo's Italian operation was fined €2.5m in 2021 for implementing an automated scheduling system that was considered to treat delivery riders unfairly and for failing to provide human oversight and an appropriate intervention and appeals process.*[27]

Recommendation

Ensure that processes supported by AI have proper governance and always provide for timely appeal and human intervention. Ensure that all automated decision making is subject to ongoing review to ensure fairness and that the deployment of such systems has provably positive benefits for data subjects.

27 https://www.sciencedirect.com/science/article/abs/pii/S0267364917302091

Rights limitations and leakage risk

Individual-rights compliance requires a data and model deletion architecture, not a blanket assumption that personal data are either permanently embedded or easily erased. Personal data can occur in training records, inference inputs and outputs, or the model itself. The relevant rights and exceptions must be assessed at each stage.[xvii]

Distinguish prompts and interaction logs; source documents and retrieval or vector stores; recoverable memorised training examples; derived statistical influence that is not identifiable personal data; and fine-tuning adapters, checkpoints and backups. Deletion from a retrieval store is different from removing information from base-model weights, and deleting raw training data does not itself establish that trained artefacts no longer contain personal data.[xiv, xvii]

Before training or fine-tuning on personal data, our recommended architecture records provenance, locations, retention, controller and processor responsibilities, deletion propagation and evidence of completion. It should include retrieval indexes, caches, logs, adapters, checkpoints and backups where relevant, with safeguards for legally required retention.

Assess retraining, machine unlearning and output-suppression controls against the particular model and rights request. Do not promise that unlearning is universally available or that suppressing an output is equivalent to erasure. Test residual recoverability and record limitations and legal decisions. Where an effective rights-compliance route cannot be established, reconsider whether the proposed training or deployment is lawful.[xiv, xvii]

Recommendation

There is no easy solution to this problem. Avoiding the use of identifiable personal data for model training, where possible, and focusing on minimising its use in any system that operates on a continuous learning basis will reduce the risk. Some vendors, including IBM[28], claim to have toolkits that allow deletion and correction of personal data in trained models. This paper does not endorse any particular toolset or approach.

28 https://developer.ibm.com/blogs/data-minimization-for-machine-learning/

Abuse

Surveillance

The European regulations in particular highlight the risk of excessive use of machine-recognition of biometrics but the underlying concern is broader. The scalability of automated processing permits extensive and ubiquitous surveillance across a wide range of technologies and circumstances – whether, e.g., monitoring working patterns by analysing keystrokes, software use and network traffic, or combining CCTV, facial recognition and device detection to track individuals' movements.

Regulators in most countries have expressed profound concerns about the growth of surveillance and in particular its application in asymmetric relationships such as that between employer and employee. Again, this is at its root an issue of fairness, but there are also potential clashes with fundamental human rights enshrined in e.g. the European Charter of Fundamental Rights.

> *White Castle, a US fast food chain, was sued by former employees protesting over-collection and processing of biometric data (specifically fingerprints). At one point the potential damages threatened to run to some US$17bn, although the case was finally settled at US$9.4m.*[29] *The recent parliamentary committee discussion of movement tracking by Royal Mail is also instructive.*[30]

Recommendation

Take care to consider the balance of individual rights and productivity (or security) when considering the implementation or extension of surveillance. From a litigation risk perspective especial care is needed regarding remote biometrics (face recognition, centralised fingerprint recognition and so forth) but modern working patterns also see temptations to use location or activity data (such as keystrokes) in automated surveillance models.

29 https://www.natlawreview.com/article/17-billion-slider-illinois-supreme-court-decides-white-castle-bipa-case

30 https://www.telegraph.co.uk/business/2023/02/02/royal-mail-dystopian-tracking-system-tells-bosses-when-postmen/

Misuse of capability including recombination risk

The pattern-matching capabilities of machine learning models are powerful; as noted from some of the examples above they are particularly good at recognising patterns in data that are not apparent to human observers – allowing, for example, the deduction of gender, health conditions or identity from data that has theoretically been stripped of the relevant indicators. Given wide access to datasets the possibilities for abuse of this capability are considerable; such abuse may be intentional and malicious or equally possible a well-intentioned effort to derive useful information without recognition of the privacy, regulatory or broader ethical risks involved.

One extension of this risk, but one not confined to learning models, is that data from multiple sources when combined may reveal information, including sensitive details, about data subjects that were either thought to have been suppressed, or which were never intentionally collected in the first place. However, since many machine learning projects specifically attempt to derive useful insight from the combination of multiple datasets this risk is particularly pronounced.

> *Machine-based pattern recognition has been used to de-anonymise individuals based on supposedly anonymous location data bought in from apps. Notably this featured in two cases where interest groups used machine learning to reveal sexuality – so also breaching rules around processing of sensitive data. In both cases the targets were Catholic priests who were "outed" as being gay and in consequence dismissed or risked dismissal from the* priesthood.[31, 32]

Recommendation

Uses of machine learning systems to derive targeting or categoric information from datasets collected for other purposes must always be subject to ethical review. Ethics committees should not only consider whether the intended use is acceptable but also both whether any unintended negative consequences are possible and whether the dataset, technology or outputs could be abused.

31 https://arstechnica.com/tech-policy/2021/07/catholic-priest-quits-after-anonymized-data-revealed-alleged-use-of-grindr/

32 https://www.theguardian.com/world/2023/mar/10/colorado-catholic-group-identify-priests-gay-apps

Real world examples

Recommendation algorithms on social media platforms like Instagram, X and YouTube have been shown to prioritise extreme and emotionally compelling content, regardless of accuracy. In 2015-8, Facebook's tendency to promote fake news articles designed to maximise clicks based on sentiment overwhelmed legitimate news and information sources in Myanmar and helped foment racial and political tensions in the country. In 2018, a United Nations investigation determined that the violence against the Rohingya constituted a genocide and that Facebook had played a "determining role" in the atrocities.[33]

AI models tend to be trained on biased data sets, which can lead to biased outputs. Such bias can reinforce existing social inequities, as was the case with a machine learning algorithm deployed by the city of Rotterdam in 2017, that was intended to help detect welfare fraud by assigning people risk scores. While the data used to train the model did not explicitly include race, ethnicity, or place of birth among the 315 variables, it used to calculate the risk score, some second order characteristics (for example, migrant status) were used by the model as inadvertent proxies for ethnicity and gender, and it began discriminating unfairly against people on these bases. Such discrimination is illegal under Dutch law, and the tool had to be withdrawn from service in 2021.[34]

In 2019 the US National Institute of Standards and Technology (NIST) tested 200 commercially available facial recognition algorithms on their accuracy for "one-to- one" matching (matching a photo of someone to another photo of the same person in a database, used to unlock smartphones or check passports), and "one-to-many" searching (determining whether a photo of someone has any match in a database,
used by police departments to identify suspects in an investigation). All the algorithms developed in the US proved less accurate at matching Asian, African American and Native American faces than they were at matching white faces.[35]

A recruitment engine developed by Amazon in 2014 had to be shut down after three years because it was not rating potential applicants in a gender-neutral manner. The company had trained the algorithm on resumés from its previous 10 years of hiring, but as it had predominantly hired men in that period, the model had learned that women were not suitable for the job. Despite multiple attempts to fix the bias the project eventually had to be abandoned. [36]

33 https://www.technologyreview.com/2021/11/20/1039076/facebook-google-disinformation-clickbait/

34 https://www.wired.com/story/welfare-state-algorithms/

35 https://www.technologyreview.com/2019/12/20/79/ai-face-recognition-racist-us-government-nist-study

36 https://www.reuters.com/article/idUSKCN1MK0AG/

Minimum requirements – and next steps

AI regulation is now operational, with significant obligations already applicable and further high-risk requirements being phased in. Organisations need an implementation programme tied to the relevant jurisdiction, legal role, intended use and commencement date, rather than a plan merely to prepare for future legislation.[i-vi, ix-xiii, xxi]

As part of this they should maintain a dated legal and system inventory covering providers, deployers, models, retrieval sources, agents and connected services. Distinguish mandatory duties from recommended controls and voluntary standards.[i, xv]

Priorities are to screen prohibited uses, implement AI-literacy measures and transparency duties, establish competent oversight and challenge routes, prepare applicable high-risk evidence, and test security and deletion controls. On top of this environmental proportionality, supplier resilience and a workable suspension or withdrawal route can be added to the governance programme.[i, xiv]

One certainty is that AI regulations will follow on from and build upon existing data protection legislation and practices, as these provide the best existing exemplar for audit and assessment. As with most compliance activities, internal transparency and good record keeping are crucial. Assessing your AI infrastructure as you develop and deploy it in line with the main risk assessment principles being used by international legislators for their own guidance will not only save you unnecessary time, expense and effort later but will equally guide your own exploration of this exciting new technological space and make sure that your efforts produce outcomes that are genuinely beneficial to your employees and customers and a credit to your business.

Jurisdictional and legislative tracker

Baseline: 6 October 2026.

EU | Enacted, progressively applicable | Regulation (EU) 2024/1689 as amended | Track Article 111(4) marking transition (2 December 2026), existing GPAI models and national sandboxes (2 August 2027), Annex III (2 December 2027), Annex I (2 August 2028), and Article 111(2) public-authority-use deadline (2 August 2030) | Sources i-iii.

UK | Enacted and commenced, with savings | DUAA section 80 / UK GDPR 22A-22D | Automated-decision amendments commenced 5 February 2026; track further Article 22D regulations and ICO guidance | Sources v-vii, xix.

US federal | Executive policy, not a comprehensive AI statute | EO 14179 / EO 14365 | Track agency implementation, litigation and federal-state interaction separately from applicable law | Sources vii-xix.

California | Representative regimes | SB 53, AB 2013 and SB 942 | SB 942/AB 853 amended commencement position remains unverified; do not treat it as publication-cleared. Track thresholds, amendments and dates separately | Sources xi-xii, xx.

Colorado | Replacement enacted, key duties forthcoming | SB26-189 | Covered ADMT duties from 1 January 2027 | Source xxi.

China | Applicable | AI-generated-content labelling Measures | Effective 1 September 2025; track standards and implementation guidance | Source xiii.

Primary Sources

Primary sources checked on 7 October 2026 for the regulatory position as at 6 October 2026. References identify the source of legal and technical statements; recommendations are expressly labelled. The earlier paper's historical references are retained where relevant. Official legislation prevails over guidance and summaries.

i. EU AI Act, consolidated text (27 July 2026), including Regulation (EU) 2026/1744. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02024R1689-20260727

ii. European Commission AI Act implementation timeline. https://ai-act-service-desk.ec.europa.eu/en/ai-act/timeline/timeline-implementation-eu-ai-act

iii. European Commission GPAI Code and provider guidance. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

iv. European Commission guidelines for GPAI providers. https://digital-strategy.ec.europa.eu/en/policies/guidelines-gpai-providers

v. Data (Use and Access) Act 2025, section 80. https://www.legislation.gov.uk/ukpga/2025/18/section/80

vi. Commencement No. 6 Regulations 2026, including savings. https://www.legislation.gov.uk/uksi/2026/82/made

vii. Executive Order 14179, 23 January 2025. https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/

viii. Executive Order 14148 of 20 January 2025, section 2(ggg): revocation of Executive Order 14110.

ix. Executive Order 14365, 11 December 2025. https://www.whitehouse.gov/presidential-actions/2025/12/eliminating-state-law-obstruction-of-national-artificial-intelligence-policy/

x. America's AI Action Plan, July 2025. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

xi. California Governor: SB 53 signed 29 September 2025. https://www.gov.ca.gov/2025/09/29/governor-newsom-signs-sb-53-advancing-californias-world-leading-artificial-intelligence-industry/

xii. California AB 2013: training-data transparency. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB2013

xiii. CAC: Measures for Labelling AI-Generated Synthetic Content (Chinese official text). https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm

xiv. NIST AI 600-1, Generative Artificial Intelligence Profile, July 2024. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

xv. NIST AI RMF: final framework and current revision status. https://www.nist.gov/itl/ai-risk-management-framework

xvi. NIST AI RMF core: Govern, Map, Measure and Manage. https://airc.nist.gov/airmf-resources/airmf/

xvii. ICO: Individual rights in AI systems (guidance under review following DUAA). https://cy.ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-individual-rights-in-our-ai-systems/?q=100

xviii. UK AI regulation white paper: policy context. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

xix. ICO consultation on updated ADM guidance, March-May 2026. https://cy.ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/2026/03/ico-consultation-on-the-draft-guidance-about-automated-decision-making-including-profiling/

xx. California SB 942 official legislative record. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB942

xxi. Colorado SB26-189: enacted ADMT regime. https://leg.colorado.gov/bills/sb26-189

xxii. NIST AI 100-2e2025: adversarial machine-learning taxonomy. https://csrc.nist.gov/pubs/ai/100/2/e2025/final

xxiii. ISO/IEC 42001 overview.

xxiv. OECD updated AI Principles.

Securys Your AI compliance partner

In a world of fast evolving AI regulations, it is difficult for companies to understand how they should be using artificial intelligence within their businesses. aiEthix from Securys, the global privacy and data protection consultancy, provides practical AI implementation and governance solutions that take the guesswork out of managing AI responsibly and allow firms to explore machine learning opportunities that will withstand future developments in technology and legislation.

What we offer

Our team has many years of experience helping organisations navigate many kinds of data protection and privacy challenges. Now, we are applying our advanced knowledge and unique insights into managing risk in these areas to the burgeoning and exciting field of machine learning.

Our approach is refreshingly practical and our methods leave no stone unturned. Our start point is our clients' aspirations to ensure their AI infrastructure aligns with wider organisational goal

AI discovery
Mapping your existing AI footprint

An AI discovery draws upon our experience of data privacy discovery and audit for organisations of all types to conduct low impact high insight analysis of your workflows and data flows involving internally built and externally sourced AI/ML tools as well as unsanctioned "shadow AI", which we investigate in a non-confrontational, non-judgemental way.

AI Governance
Making light work of your regulatory requirements

Data knows no border. Even a modest AI/ML deployment exposes you to regulatory and legislative enforcement across multiple jurisdictions. Let our experts work alongside your risk and compliance teams to identify, monitor and mitigate all the risk that groundbreaking technologies run, without sacrificing your ability to profit from the opportunities that AI/ML bring.

AI by Design
Put data protection to work

Our consultants will work with you at every stage of the AI product cycle, from inception to delivery, advising on data preparation and system architecture to ensure that your models source and handle data responsibly, conform to relevant regulation, and are adequately risk-assessed and proactively risk-mitigate

About Securys



Global knowledge, locally delivered

Our teams of international experts, operating across dozens of jurisdictions, deliver relevant advice and see the world through your lens.



A deep understanding of data

We bring decades of experience delivering privacy-by-design programmes to clients of all scopes and scales, across multiple industries and jurisdictions.



Practical approach

We will tackle any bespoke AI project in a constructive fashion, breaking down barriers within teams to enable better compliance and a united approach to implementation.



About Securys

Securys is a specialist data privacy consultancy with a difference. We're not a law firm, but we employ lawyers. We're not a cybersecurity business but our staff qualifications include CISSP and CISA. We're not selling a one-size-fits-all tech product, but we've built proprietary tools and techniques that work with the class-leading GRC products to simplify and streamline the hardest tasks in assuring privacy.

We're corporate members of the IAPP, and all our staff are required to obtain one or more IAPP certifications. We are ISO 27001 and ISO 27701 certified with a comprehensive set of policies and frameworks to help our clients achieve and maintain certification.

Above all, our relentless focus is on practical operational delivery of effective data privacy for all your stakeholders.

We're not just a consultancy. We're your privacy engine room. We can stand in your boardroom and do strategy with the top team, and work with your compliance teams to solve knotty problems. We can audit your compliance and deliver drillable risk dashboards across the organisation. But above all we can get involved at ground level and help your frontline teams get the job done. That's Privacy Made Practical®.



Securys®
Privacy Made Practical®

4th floor
91 Goswell Road
London
EC1V 7EX

T 0800 193 8700
E info@securys.co.uk
W www.securys.co.uk

securys-limited
@SecurysUK

© Copyright 2026 Securys Ltd. All Rights Reserved.

UK0029/0124