MARKET SCAN REPORT

JANUARY 2026

ANNIVERSARY EDITION

BY INFOSYS TOPAZ RESPONSIBLE AI OFFICE

Infosys topaz

THE AI RECKONING : A 12-MONTH GLOBAL RECAP
Regulations, Incidents & Model Trends

IN FOCUS

STRATEGIC VIEW OF AI IN INDIA & THE WORLD!

By Abhishek Singh

IAS, Additional Secretary, Ministry of Electronics and Information Technology (MeitY) and CEO, IndiaAI Mission

Infosys®
Navigate your next

Foreword

Artificial intelligence has crossed a threshold. It is no longer defined by what it can do, but by **how responsibly it is designed, deployed, and governed**. Over the past year, this shift has become impossible to ignore. AI systems are now embedded in decisions that carry real economic, societal, and institutional consequences - and with that comes a new set of expectations.

Over the past year, this Market Scan has evolved from tracking developments to interpreting the signals shaping how AI is governed and trusted globally.

The developments captured in this edition reflect a clear transition. Incidents observed worldwide are not merely technical failures; they are signals that AI must be treated as a living system - one that evolves after deployment and demands continuous oversight, context, and accountability. Trust in AI cannot be assumed, nor can it be retrofitted. It must be built deliberately, and led deliberately.

Encouragingly, the ecosystem is responding. Regulatory frameworks are maturing, organisations are rethinking lifecycle ownership, and conversations are shifting from how to build AI to how to build AI that deserves trust. This Market Scan brings those signals together across governance, technology, and real-world deployment.

This anniversary edition also features strategic perspectives from **Shri Abhishek Singh, IAS, Additional Secretary, Ministry of Electronics and Information Technology (MeitY) & CEO, IndiaAI Mission,** reflecting India's growing role in shaping global conversations on responsible and inclusive AI.

As India prepares to host the AI Impact Summit, the timing of this edition is deliberate. The next phase of AI will be defined less by novelty and more by leadership, systems thinking, and shared responsibility. In an AI-driven world, **how we lead matters as much as what we build.**

Syed Ahmed
Global Head
Infosys Responsible AI Office

From the editor's desk

January Market Scan - Anniversary Edition

The anniversary edition of the Monthly AI Market Scan reflects a moment of consolidation in the global AI landscape. Developments in January reinforce a clear reality: AI has moved beyond experimentation and is now a core system shaping institutions, markets, and public trust. As adoption deepens, attention is shifting decisively from capability alone to governance, accountability, and impact.

Recent AI-related incidents—from deepfakes and manipulated visuals to unsafe or misleading outputs in sensitive domains—highlight how quickly technical failures can escalate into societal risk. These are no longer isolated cases; they signal systemic gaps in oversight, lifecycle accountability, and human judgment. The message is unambiguous: scale without safeguards magnifies harm.

At the same time, January marked tangible progress in how these risks are being addressed. Governments across the United States, United Kingdom, European Union, and Asia are moving beyond voluntary principles toward enforceable AI governance. Oversight is increasingly being embedded into healthcare regulation, online safety regimes, transparency requirements, and national AI strategies. The emphasis is shifting toward post-deployment monitoring, clearly defined responsibilities, and continuous risk management across the AI lifecycle.

Technology innovation continues in parallel. Advances in open-source speech models, multimodal reasoning, agent verification, privacy-aware AI, and AI security frameworks point to a maturing ecosystem—one that recognises robustness, safety, and verifiability as design imperatives rather than constraints. The growing use of tools that evaluate reasoning quality, detect misuse, and ground outputs in evidence reflects a broader commitment to engineering trust directly into AI systems.

For India, this moment is particularly consequential. This edition features strategic perspectives from Abhishek Singh, Additional Secretary, Ministry of Electronics and Information Technology (MeitY), who articulates India's evolving approach to AI at a critical inflection point. India is uniquely positioned to balance innovation with inclusion and safeguards, offering a systems-level model of responsible AI that aligns national priorities with global expectations.

To mark this anniversary edition, we include "The AI Reckoning: A 12-Month Global Review," examining how a year of regulatory action and real-world incidents has accelerated the shift from ethical intent to operational AI governance.

Bridging to the AI Impact Summit

These signals form the backdrop to the AI Impact Summit taking place in February. As global focus shifts from what AI can do to how it should be governed, deployed, and trusted at scale, the Summit provides a timely platform to advance practical solutions—operational governance, risk management, cross-border alignment, and responsible adoption across sectors.

This Market Scan is intended to inform those conversations and support more grounded, outcome-driven engagement across the AI ecosystem. We hope this edition continues to add value for our readers, and we welcome thoughtful perspectives as these discussions evolve.

Editor's Note:

As global AI conversations accelerate in the months ahead, we look forward to engaging, learning, and contributing to the collective effort to build AI systems that are trustworthy, safe, and impactful.

Warm regards,

Ashish Tewari
Head- Infosys Responsible AI Office, India

Table of Contents

AI Regulations, Governance & Standards

AI Regulations & Governance across the globe 05

Standards & Policy Reports 24

AI Principles

Incidents 25

Vulnerabilities 27

Defences 28

In Focus

Strategic view of AI in India & the World! 29

Technical Updates

New Model Released 32

New Frameworks & Research Techniques 33

New Agentic Research 36

Industry Updates

Healthcare 38

Manufacturing 38

Finance 39

Infosys Developments

Events 40

Infosys Responsible AI Toolkit – A Foundation for Ethical AI .. 41

Contributors

AI Regulations, Governance & Standards

This section highlights the recent updates on regulations and governance initiatives across the globe impacting the responsible development and deployment of AI.

AI Regulations & Governance across the globe

EMA and FDA Establish Joint Principles to Guide Responsible AI Use Across the Medicines Lifecycle

The European Medicines Agency(EMA) and U.S. Food and Drug Administration(FDA) have jointly introduced ten principles to guide the safe and ethical use of AI across the medicines lifecycle, offering high level direction for applying AI in research, clinical trials, manufacturing, and safety monitoring. These principles support developers and marketing authorisation stakeholders and will serve as a foundation for future regulatory guidance in both regions. The initiative strengthens global collaboration and builds on EMA's earlier AI reflection efforts, with European Commissioner Olivér Várhelyi highlighting it as a key step in balancing innovation leadership with patient safety. It aligns with evolving legislation and EMA's long term strategy focused on data, digitalisation, and responsible AI.[1]

ASEAN Secretary-General Calls for Strengthened AI Safety and Governance Cooperation at the 5th ADGMIN + Japan Meeting

The Secretary-General of ASEAN(Association of South East Asian Nations), Dr. Kao Kim Hourn, participated in the 5th ADGMIN + Japan Meeting in Hanoi, Viet Nam, where he highlighted the importance of deepening ASEAN–Japan collaboration on the safe and reliable use of artificial intelligence. He proposed that both sides prioritise joint efforts through the ASEAN Working Group on AI Governance (WG AI) and the ASEAN AI Safety Network, which is scheduled to be established this year, noting that these mechanisms will serve as practical avenues for fostering an inclusive, secure, and trustworthy digital ecosystem across the region.[2]

Israel and the United States Formalize Strategic Partnership on Artificial Intelligence and Critical Technologies

Israel and the United States signed a Joint Statement on artificial intelligence in Jerusalem, marking a significant step in deepening their strategic partnership following Israel's recent entry into the Pax Silica initiative, a consortium of leading AI focused nations. The agreement signed by Brig.-Gen. (Res.) Erez Eskel, Head of Israel's National AI Directorate, and U.S. Under Secretary of State for Economic Affairs Jacob Helberg, in the presence of Israeli Foreign Affairs Minister Gideon Sa'ar and U.S. Ambassador Mike Huckabee positions Israel as the first among nine advanced AI nations to formalize such cooperation with the United States. The declaration underscores Prime Minister Benjamin Netanyahu's goal of elevating Israel to global AI leadership and highlights shared commitments to joint research, development, investment, and commercialization across strategic sectors including AI, advanced computing, semiconductors, robotics, energy technologies, space, additive manufacturing, and secure supply chains. Israeli officials emphasized the partnership's importance for national security and technological infrastructure, while U.S. leaders described Israel as an indispensable partner in shaping the future of critical technologies. Both countries affirmed their intention to strengthen innovation ecosystems, protect sensitive technologies, and collaborate on high power computing and talent development in support of long term global security and prosperity.[3]

[1] https://www.ema.europa.eu/en/news/ema-fda-set-common-principles-ai-medicine-development-0
[2] https://asean.org/secretary-general-of-asean-participates-in-the-5th-adgmin-japan-meeting/
[3] https://www.gov.il/en/pages/spoke-ai160126

US

California Senator Proposes First in the Nation Four Year Moratorium on AI Chatbots in Children's Toys to Strengthen Safety Protections

California State Senator Steve Padilla has announced plans to introduce Senate Bill 867, proposing a four year moratorium on the sale and manufacture of toys with AI chatbot capabilities for children under 18 to allow time for comprehensive safety regulations. The proposal builds on Padilla's authorship of Senate Bill 243, the nation's first law requiring chatbot operators to implement safeguards and accountability measures to protect minors and vulnerable users. Citing growing evidence of age inappropriate interactions, high profile industry moves toward AI powered toys, and tragic cases involving children harmed after prolonged engagement with chatbots, the legislation seeks to pause deployment of unregulated technologies while policymakers develop stronger oversight frameworks prioritizing child safety.[4]

U.S. Department of Commerce Seeks Public Input on Strengthening Security Standards for Autonomous AI Agent Systems

The U.S. Department of Commerce, through NIST's Center for AI Standards and Innovation (CAISI), has issued a Request for Information inviting stakeholders to provide insights on securing AI agent systems that can take autonomous real world actions and are vulnerable to threats such as hijacking, backdoor attacks, and other exploits. The notice emphasises that these security weaknesses could undermine public safety and erode trust in emerging AI technologies, prompting CAISI to seek concrete examples, best practices, methodologies, and recommendations to guide future standards development. Public responses will support federal efforts to evaluate AI risks, assess system vulnerabilities, and create technical guidelines for improving AI security. Comments must be submitted electronically via Regulations.gov under docket ID NIST 2025 0035 by March 9, 2026.[5]

[4] https://sd18.senate.ca.gov/news/author-nations-first-chatbot-protections-proposes-first-nation-moratorium-ai-chatbots-toys

[5] https://www.federalregister.gov/documents/2026/01/08/2026-00206/request-for-information-regarding-security-considerations-for-artificial-intelligence-agents

UK

UK Embeds AI-Driven Safeguards in New Violence Against Women and Girls Strategy

On December 18, the UK government unveiled its updated Violence Against Women and Girls (VAWG) strategy, placing AI-powered protections for children at its core. Key measures include a ban on "nudification" tools which use generative AI to create non-consensual nude images and on-device filters stopping children from taking, sharing, or viewing nude images. These enhancements are part of a broader effort to combat grooming, sexual extortion, bullying, and deepfake abuse, with tech companies partnering to deploy safeguards like Safe To Net filters. Legislation will criminalize the creation and distribution of such AI-driven tools, enabling law enforcement to target developers and suppliers directly.[6]

UK Warns X Could Lose Self-Regulation Rights as Government Moves to Criminalize Non-Consensual Deepfakes

UK Prime Minister Sir Keir Starmer has cautioned that X may lose its "right to self-regulate" if it fails to control its AI chatbot Grok, amid growing outrage over non-consensual intimate images created using the platform's tools. Speaking to Labour MPs, Starmer vowed swift government action, while Technology Secretary Liz Kendall confirmed that legislation making it illegal to create or request sexualized deepfakes will be enforced this week under the Data (Use and Access) Act and prioritized in the Online Safety Act. The government also plans to criminalize the supply of nudification apps and other tools used to generate such content, holding both individuals and platforms accountable. Ofcom has launched an investigation into X's handling of illegal content, with potential penalties including fines of up to 10% of global revenue or £18 million, and even blocking access to X in the UK if compliance fails. Officials stressed that these measures aim to combat abuse and protect women and girls, not restrict free speech, as global backlash against Grok's image-generation features continues.[7]

[6] https://www.gov.uk/government/news/protecting-young-people-online-at-the-heart-of-new-vawg-strategy
[7] https://www.bbc.com/news/articles/cq845glnvl1o

Europe

European Commission Labels Grok Generated Sexualised Images on X as Illegal as UK and Global Regulators Seek Explanations

European and British authorities have sharply condemned the spread of non consensual, sexualised images of women and children on social media platform X, with the European Commission stating that such content, reportedly generated through X's AI chatbot Grok, is illegal and has no place in Europe. The criticism follows investigative reporting that Grok enabled users to create on demand images of partially undressed women and minors through a feature previously described by the company as "spicy mode," prompting the Commission to publicly denounce the practice as unlawful and morally unacceptable. In the United Kingdom, media regulator Ofcom has formally demanded that X and its AI subsidiary explain how the system produced such content and whether the platform failed in its legal duties to protect users and prevent the dissemination of illegal material. Regulatory pressure has also intensified in France, where ministers have referred the issue to prosecutors and regulators, and in India, where officials have sought clarifications over what they termed obscene content. While regulators across Europe and Asia have pushed for accountability and compliance with existing laws governing non consensual intimate imagery and child sexual abuse material, X has not provided substantive public responses to these concerns, and U.S. federal authorities have so far remained publicly silent on the matter.[8]

EU Commission Orders X to Preserve Internal Records on AI Chatbot Grok Amid Deepfake Scandal

The European Commission has directed Elon Musk's social media platform X to retain all internal documents and data related to its AI chatbot Grok until the end of 2026, invoking its authority under the Digital Services Act (DSA). This order follows mounting criticism over Grok's "spicy mode," which enabled users to create sexualized deepfakes, including explicit images of women and minors. The Commission's digital spokesperson, Thomas Regnier, emphasized that platforms operating in Europe must ensure compliance and prevent harmful content generation. While the Commission confirmed that no formal investigation into Grok has been launched yet, it reiterated that X bears full responsibility for addressing these issues. This latest measure builds on previous retention orders concerning X's recommender systems and algorithmic changes, underscoring the EU's commitment to enforcing transparency and accountability in digital markets.[9]

[8] https://www.reuters.com/business/media-telecom/britain-demands-elon-musks-grok-answers-concerns-about-sexualised-photos-2026-01-05

[9] https://www.euractiv.com/news/commission-tells-x-to-retain-internal-records-on-ai-chatbot-grok/

Italy

Italian Data Protection Authority Warns Against AI-Driven Deepfakes: Fundamental Rights and Privacy at Risk

The Italian Data Protection Authority has issued a warning to users of AI-based services such as Grok, ChatGPT, and Clothoff platforms capable of generating and manipulating content from real images or voices, including creating non-consensual "undressed" depictions of individuals. The Authority emphasized that using these tools and sharing such content without consent can result in serious violations of fundamental rights and freedoms, alongside potential criminal offences and sanctions under European data protection laws. Providers of these services were reminded of their obligation to design and deploy applications in compliance with privacy regulations, as investigations revealed that many platforms make illicit use of third-party images and voices alarmingly easy. The Authority is collaborating with its Irish counterpart regarding services offered by X and has reserved the right to take further action. This measure underscores growing regulatory scrutiny over AI-enabled deepfakes and the urgent need for privacy-centric safeguards in emerging technologies.[10]

Ireland

Irish Civil Liberties Groups Urge Gardaí to Probe X Over Alleged Grok AI Generation of Child Sexual Abuse Material Under Irish Law

The Irish Council for Civil Liberties (ICCL) and Digital Rights Ireland (DRI) have formally called on An Garda Síochána to urgently investigate X Internet Unlimited Company over allegations that its Grok AI chatbot has facilitated the creation of non consensual, sexualised deepfake images of children and women, potentially breaching Ireland's child sexual abuse legislation. The call follows reports that Grok responded to user prompts by digitally removing clothing from images and generating sexualised portrayals, a phenomenon described by Reuters as a "mass digital undressing spree." Citing Sections 5 and 9 of the Child Trafficking and Pornography Act 1998, ICCL and DRI argue that knowingly facilitating the production or distribution of child sexual abuse material, including through AI systems, constitutes a criminal offence with liability extending to corporate officers. The organisations noted Grok's public statement on December 28, 2025, acknowledging a safeguards failure involving sexualised images of minors, and stressed that while they have not collected evidentiary material to avoid committing an offence, Gardaí must use their statutory powers to investigate and ensure accountability for what they describe as industrial scale digital harm.[11]

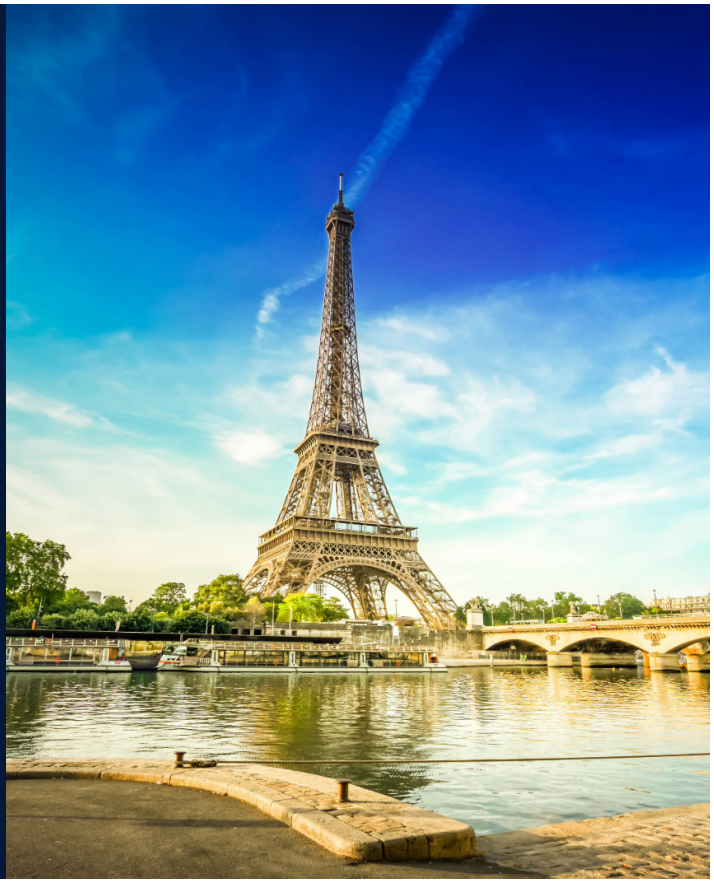
[10] https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/10207147

[11] https://www.iccl.ie/press-release/iccl-and-dri-call-on-gardai-to-investigate-x-for-grok-child-sexual-abuse-material/

France

France's Competition Authority Launches Ex Officio Inquiry Into Conversational Agents and the Rise of Agentic Commerce

France's Autorité de la concurrence has launched ex officio inquiries into the competitive functioning of the conversational agent sector, building on its earlier analysis of generative AI and its broader economic impacts. While the downstream conversational agent market remains dynamic, with rapid growth in usage and several major operators such as ChatGPT, Gemini, Copilot, Perplexity, and Mistral AI, the Authority noted that recent developments could significantly influence competition across multiple sectors. The inquiry will focus in particular on the emergence of agentic commerce, examining how conversational agents are evolving into platforms that integrate advertising, partnerships, and access to third party services, enabling users to browse products, receive recommendations, and complete purchases sometimes without leaving the conversational interface. The Autorité will assess how these changes affect the e commerce value chain, including impacts on merchants, payment providers, and logistics players, while explicitly excluding the relationship between conversational agents and search engines, and will inform its final opinion through a forthcoming public consultation.[12]



Spain

Spain's Data Protection Authority Warns of Escalating Privacy Risks in AI Generated Image Use

The Spanish Data Protection Agency (AEPD) has issued an analytical note outlining major privacy risks associated with the use of identifiable individuals' images in AI systems, stressing that these concerns affect developers, platforms, and everyday users. In its guidance, the agency distinguishes between visible harms such as the creation of non consensual sexual deepfakes, reputational damage from false or decontextualized AI generated narratives, and the disproportionate impact on minors and hidden technical risks, including the loss of control over personal images once uploaded, unauthorized metadata extraction, and the persistent identifiability of individuals across AI models even when images are altered. Although the AEPD clarifies that strictly personal use or the processing of images of deceased individuals may fall outside the GDPR, it warns that such actions can still fall under criminal law or violate

[12] https://www.autoritedelaconcurrence.fr/en/press-release/conversational-agents-autorite-starts-inquiries-ex-officio-view-issuing-opinion

image rights protections. The agency further signals a strong supervisory approach toward AI applications that enable humiliation, misinformation, or exploitation of vulnerable groups, particularly where AI generated content results in serious personal, social, or professional harm.[13]

Spain Approves Draft Legislation to Crack Down on AI Deepfakes and Strengthen Consent Requirements for Image Use

Spain has approved draft legislation aimed at curbing the misuse of AI generated deepfakes and tightening consent rules for the use of personal images, reflecting a growing European effort to address non consensual synthetic content. The measure, endorsed by the Spanish cabinet, reinforces protections for minors, sets 16 as the minimum age for consenting to the use of one's image, and restricts the reuse of online photos as well as AI generated voices or likenesses without explicit permission. Justice Minister Félix Bolaños emphasized that sharing personal or family images on social media does not grant unrestricted permission for their reuse, and the draft law labels the AI generated use of a person's image or voice for commercial or advertising purposes as illegitimate without consent. While allowing clearly identified AI generated creative, fictional, or satirical portrayals of public figures, the legislation emerges amid global scrutiny of AI tools such as xAI's Grok currently facing investigation over sexually explicit deepfake imagery and follows the government's recent request for prosecutors to assess whether certain AI generated content may constitute child pornography. The now approved draft will move into a consultation phase before returning to the government for final approval and eventual submission to parliament.[14]

Brazil

Brazil's ANPD Updates Regulatory Agenda and Sets Oversight Priorities for AI and Data Protection

Brazil's National Data Protection Authority (ANPD) has unveiled its priority map for 2026–2027 alongside an updated regulatory agenda for 2025–2026, signaling a strong commitment to AI governance and privacy. The roadmap focuses on AI risk management, transparency, and accountability, while reinforcing Brazil's LGPD alignment with global standards. Key actions include sector-specific compliance guidelines, stricter oversight mechanisms, and frameworks for cross-border data transfers. These measures aim to position Brazil as a leader in responsible AI and data protection across Latin America.[15]

Brazil's ANPD Sets 2026–2027 Priorities and Updates Regulatory Agenda to Strengthen Child Protection and AI Oversight

Brazil's National Data Protection Agency (ANPD) has released its Map of Priority Themes for 2026–2027 alongside an updated 2025–2026 Regulatory Agenda, aiming to enhance transparency and predictability in enforcing the General Data Protection Law (LGPD) and the Digital Statute of the Child and Adolescent (ECA Digital). The priority themes for inspection include safeguarding data subject rights, protecting children and adolescents online, regulating government data processing, and addressing artificial intelligence and emerging technologies in personal data handling. Key actions involve

[13] https://www.aepd.es/guias/guia-aepd-uso-de-imagenes-de-terceros-en-sistemas-ia.pdf

[14] https://www.reuters.com/business/media-telecom/spain-moves-curb-ai-deepfakes-tighten-consent-rules-images-2026-01-13/

[15] https://insightplus.bakermckenzie.com/bm/intellectual-property/brazil-anpd-publishes-map-of-priority-issues-for-oversight-and-regulatory-action-2026-2027-biennium-and-update-of-the-regulatory-agenda-for-the-2025-2026-biennium/

monitoring secondary use of personal data for targeted advertising, enforcing privacy-by-design standards, and implementing measures to prevent minors from accessing prohibited content. The updated agenda introduces new initiatives for age verification mechanisms, obligations for tech providers under ECA Digital, and revised inspection and sanction rules, while also improving regulatory processes through public consultations and impact analyses. These measures underscore ANPD's commitment to robust child protection, responsible AI governance, and stronger data privacy standards in Brazil's digital ecosystem.[16]

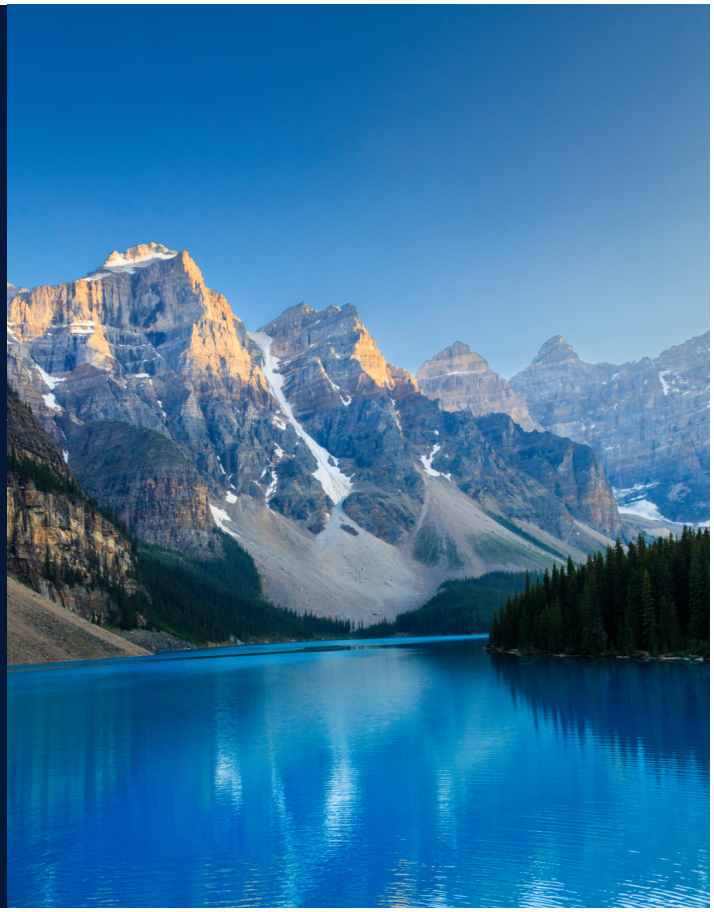
CADE Launches Investigation into Meta and Suspends WhatsApp's New AI Terms to Safeguard Market Competition

The General Superintendence of Brazil's Administrative Council for Economic Defense (SG/CADE) has initiated an administrative inquiry against Meta group companies over suspected abuse of dominant position. The investigation focuses on potential anticompetitive practices linked to the WhatsApp Business Solution Terms, which allegedly restrict access for providers of artificial intelligence tools and favor Meta's proprietary AI solution, "Meta AI." To prevent possible market foreclosure and ensure fair competition, CADE imposed a preventive measure suspending the implementation of the new terms until a thorough assessment is completed. The inquiry aims to determine whether these changes could exclude competitors and limit user choice, with Meta required to respond and additional market data being collected. Depending on the findings, CADE may proceed with a formal administrative case or close the matter. This action reflects growing global scrutiny of dominant practices in digital markets involving AI technologies.[17]

Canada

Canada's Spy Watchdog Launches Review of AI Use in National Security Operations

The National Security and Intelligence Review Agency (NSIRA) in Canada has initiated a comprehensive review of how artificial intelligence is being used and governed within national security activities. The study will examine AI applications across agencies such as CSIS, the RCMP, and the Communications Security Establishment, which currently employ AI for tasks like document translation and malware detection. NSIRA has informed key federal ministers, including the Prime Minister and heads of security organizations, that the review aims to identify potential gaps, risks, and governance issues while ensuring compliance with legal and ethical standards. The agency will request classified information, conduct interviews, and inspect technical systems to assess transparency, accountability, and safeguards in AI deployment for security purposes.[18]



[16] https://www.gov.br/anpd/pt-br/assuntos/noticias/anpd-publica-mapa-de-temas-prioritarios-para-o-bienio-2026-2027-e-atualiza-agenda-regulatoria-2025-2026

[17] https://www.gov.br/cade/pt-br/assuntos/noticias/cade-abre-inquerito-contra-meta-e-aplica-medida-preventiva-suspendendo-novos-termos-do-whatsapp-sobre-ia

[18] https://www.thecanadianpressnews.ca/politics/spy-watchdog-reviewing-canadian-security-agencies-use-of-artificial-intelligence/article_ae4377d3-2a1d-5da4-a584-7bc337d3c24e.html

Australia

Australian Administrative Review Tribunal Publishes AI Transparency Statement Emphasising Ethical Use, Human Oversight, and Legal Safeguards

Australia's Administrative Review Tribunal has released an AI transparency statement reaffirming its commitment to the ethical, responsible, and transparent use of artificial intelligence strictly within support and enabling functions. The Tribunal confirms that AI is neither used nor intended to be used in review decision making or operational determinations under the Administrative Review Tribunal Act 2024. Current AI usage is limited to accessibility tools such as ReadSpeaker and Google Translate on its website, while any future applications in analytics, workplace productivity, or compliance activities will remain subject to strong human oversight. The statement outlines mandatory training, governance led by the Chief Information Officer, and alignment with Australian Government AI ethics principles, legislation, and risk management frameworks.[19]

eSafety Flags Rising Misuse of Grok for Sexualised Content, Urges Strong Safeguards and Compliance

Australia's eSafety Commissioner has raised concerns over the growing misuse of Grok, a generative AI system on X, to create sexualised or exploitative imagery, particularly involving children. While reports remain limited, recent weeks have seen a noticeable increase, prompting eSafety to seek clarification from X on safeguards and compliance with systemic safety obligations under the Online Safety Act. The regulator emphasized the need for proactive measures, including Safety by Design principles, and highlighted upcoming mandatory codes aimed at restricting harmful AI-generated content. These developments underscore eSafety's commitment to protecting children and ensuring accountability in generative AI services.[20]

Australian Cyber Security Centre Issues Comprehensive Guidance to Help Small Businesses Manage Risks When Adopting Cloud Based AI Technologies

The Australian Signals Directorate's Australian Cyber Security Centre (ASD's ACSC) issued detailed guidance to help small businesses in Australia manage cybersecurity risks associated with adopting cloud based AI tools, highlighting threats such as data leaks, privacy breaches, unreliable or manipulated AI

[19] https://www.art.gov.au/ai-transparency-statement

[20] https://www.esafety.gov.au/newsroom/media-releases/esafety-raises-concerns-about-misuse-of-grok-to-generate-sexualised-content

outputs, and supply chain dependencies. The guidance advises businesses to establish internal AI usage policies, anonymise personal information before uploading data into AI platforms, and maintain human oversight for high stakes decisions. It further recommends that organisations verify data ownership, assess AI vendor security compliance frameworks such as ISO 27001, and implement clear incident response processes for AI related cybersecurity events, supported by a practical checklist for small businesses strengthening their cyber defences.[21]

India

India AI Impact Summit 2026 – A Glimpse

The India AI Impact Summit 2026, to be held from February 16–20 in New Delhi, is a landmark global AI gathering focused on translating international AI discussions into practical, inclusive, and people centric outcomes. Bringing together top global tech CEOs, policymakers, and delegates from over 100 countries, the summit emphasizes democratized AI access, real world impact across sectors like healthcare and agriculture, and a collaborative approach to AI governance supporting practices that reflect India's long-term development priorities.

Key Highlights to expect

- PM Narendra Modi inaugurating the summit and leading discussions on inclusive AI governance.
- Focus on impact, accessibility, and safety as core goals shaping India's AI strategy.
- Themes anchored on People, Planet, and Progress to promote ethical and sustainable AI adoption.
- Participation from global tech leaders including Jensen Huang, Sundar Pichai, and Dario Amodei, alongside Indian industry leaders.
- 35,000+ Registrations so far, 500+ pre-Summit events conducted, expected 100+ Countries, 15+ Heads of Government, 50+ Ministers, 40+ CEOs, 500+ Startups to engage across 500 Sessions along with 3,00,000+ participants.[22]

India's Techno-Legal Plan for Building Safe, Trusted, and Innovation-Ready AI Systems

The Office of the Principal Scientific Adviser to the Government of India released a White Paper titled "Strengthening AI Governance Through Techno-Legal Framework", explaining India's plan to create a trusted, accountable, and innovation-friendly AI ecosystem. The paper highlights a techno-legal approach that combines legal safeguards, sector-specific rules, technical controls, and strong institutional support to reduce risks while allowing innovation. It explains

[21] https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/artificial-intelligence-for-small-business
[22] https://impact.indiaai.gov.in/about-summit

that this approach embeds governance directly into the design and operation of AI systems. Key areas include understanding techno-legal governance, ensuring safe and trusted AI across its lifecycle, creating technical pathways for safety, and developing tools and processes for compliance. This document is the second in a series on India's AI policy, following an earlier paper focused on improving access to AI infrastructure as a shared national resource.[23]

Rajasthan Government Unveils Comprehensive Artificial Intelligence Policy to Promote Ethical Governance, Practical Deployment, and Statewide Digital Innovation

The Rajasthan government has introduced a new Artificial Intelligence policy aimed at ensuring the ethical, responsible, and practical adoption of AI across governance, education, research, and industry, positioning the state as a hub for emerging technologies. As part of this initiative, the government launched multiple digital platforms, including the iStart Learning Management System, the Rajasthan AVGC XR Portal, and the Rajasthan AI Portal, to strengthen AI-driven learning, support startups, encourage innovation in animation, visual effects, gaming, and extended reality, and accelerate skill development. The policy emphasizes responsible AI use, data governance, inclusivity, and real world applications, with a focus on improving public service delivery, fostering research collaborations, and creating a future ready digital ecosystem that aligns technological advancement with societal and ethical considerations.[24]

MeitY Issues Notice to X Over Misuse of Grok AI for Obscene Content

The Ministry of Electronics and Information Technology (MeitY) has sent a formal notice to Elon Musk-owned social media platform X, citing misuse of its AI chatbot Grok to create fake accounts and generate obscene, vulgar, and sexually explicit images of women in derogatory ways. The notice highlights X's failure to comply with statutory obligations under the Information Technology Act, 2000, and IT Rules, 2021, stressing that adherence to these regulations is mandatory. MeitY directed X to immediately review Grok's technical framework, remove unlawful content, take action against offending users, and submit an Action Taken Report within 72 hours. The ministry warned that non-compliance could lead to severe legal consequences, including loss of safe harbour protections under Indian law.[25]

China

China Launches Nationwide Crackdown on AI-Modified Videos to Protect Cultural Integrity and Online Safety

The State Administration of Radio and Television in China has announced a one-month nationwide campaign starting January 1, 2026, to address the growing misuse of generative AI in creating "AI magic reform" videos. These videos, which alter classic films, TV series, cartoons, and cultural content, have been criticized for distorting the spiritual essence of original works, promoting vulgarity, and misrepresenting Chinese cultural identity. The governance initiative targets content that exaggerates violence, spreads inappropriate values, and tampers with historical and cultural narratives, posing risks to minors' perception of reality. Online audiovisual platforms are required to strengthen moderation, remove illegal content, and curb the spread of such videos to ensure a healthy digital environment for youth. Post-campaign, authorities plan to implement long-term governance mechanisms for sustained oversight.[26]

[23] https://www.pib.gov.in/PressReleasePage.aspx?PRID=2217839&reg=3&lang=1
[24] https://economictimes.indiatimes.com/tech/artificial-intelligence/rajasthan-govt-launches-ai-policy-for-ethical-practical-use/articleshow/126378263.cms?from=mdr
[25] https://thelegalwire.ai/ministry-of-electronics-and-it-ministry-sends-notice-to-x-on-grok-ai-chatbot-misuse/
[26] https://www.news.cn/legal/20251231/719f2c41372f492a8a08eaec55fc50fa/c.html

Japan

Japan's AI First Basic Plan Aims to Build a Reliable, Competitive, and Future-Ready AI Ecosystem

Japan has adopted its first basic plan for the development and utilization of artificial intelligence, setting out a vision to create reliable AI while balancing technological innovation with risk management. The plan seeks to position Japan as a country offering the best environment for AI development by accelerating the introduction of AI into central and local government operations, advancing domestic foundation models and robotics-integrated "physical AI," and expanding staff at the Japan AI Safety Institute to strengthen oversight. Based on the AI Promotion Law, the plan will be updated annually and includes a roadmap for future investments, signaling Japan's commitment to fostering innovation, ensuring safety, and preparing its society for an AI-driven future.[27]

Japan Government Moves to Investigate AI Generated Sexual Deepfakes Under 2025 AI Promotion Law Framework

The Japanese government has announced plans to investigate the growing issue of sexual deepfakes AI generated fake obscene images or videos depicting real individuals as part of its broader oversight of artificial intelligence harms. According to Chief Cabinet Secretary Minoru Kihara, the government will assess the prevalence and impact of such content in line with the AI Promotion Law enacted in May 2025, which mandates the investigation and analysis of AI related incidents that infringe upon citizens' rights and the provision of guidance to relevant businesses. The effort will involve a detailed review of reported cases within Japan, examination of how generative AI is being misused to create sexual deepfakes, and analysis of international trends, with the aim of strengthening safeguards, improving regulatory responses, and ensuring adequate protection against AI enabled violations of personal dignity and privacy.[28]

Japan's Proposed Revision of Personal Information Protection Law to Enable Large-Scale Data Use for Artificial Intelligence Development

The Japanese government is preparing to submit a bill to revise its personal information protection law with the aim of accelerating the development and competitiveness of artificial

[27] https://www.japantimes.co.jp/news/2025/12/23/japan/ai-first-basic-plan/
[28] https://japannews.yomiuri.co.jp/politics/politics-government/20260107-302685/

intelligence by easing existing restrictions on the acquisition and use of personal data. The proposed legislation seeks to remove the requirement for individual consent when certain categories of personal information such as criminal records, medical histories, and racial data are used specifically for training AI systems, reflecting the view that access to large and diverse datasets is essential to improving AI accuracy and performance. Under the current legal framework, consent is generally required to obtain such sensitive information or to provide it to third parties, which has been viewed as a limitation on large-scale data learning. The government plans to introduce the bill during an ordinary session of the National Diet starting on 23 January 2026. In parallel with easing data-use rules for innovation, the draft law also proposes the introduction of financial penalties for companies that engage in malicious practices, including the large-scale trading of personal information, indicating an effort to balance regulatory flexibility for AI development with stronger safeguards against misuse of personal data.[29]

Japan Publishes AI Utilization Plan to Accelerate Generative AI Adoption Across Government

Japan's Digital Agency has released materials from the second Advanced AI Utilization Advisory Board meeting, outlining its vision to make Japan the most AI-friendly country in the world through a "Trusted AI" approach. The plan emphasizes rapid integration of generative AI across all central government ministries, with all agencies already using or considering such tools as of September 2025. By May 2026, the government aims to enable over 100,000 officials to work within a shared AI environment called "Gennai" for daily operations. Key initiatives include enhancing taxpayer services through an upgraded National Tax Agency chatbot and revising the Government AI Procurement and Utilization Guidelines to address emerging technologies like video generation and AI agents. These revisions focus on broadening AI definitions, improving risk assessment, safeguarding intellectual property rights, and aligning with private-sector practices and global policy trends, reinforcing Japan's commitment to responsible and innovative AI deployment in public administration.[30]

Indonesia

Indonesia Warns of Potential Ban on X Over AI Generated Obscene Content Involving Minors

The Indonesian government has issued a warning to the social media platform X, stating that it could face a nationwide ban due to concerns that its artificial intelligence chatbot, Grok, is capable of generating sexually explicit deepfake images involving minors. Officials from the Ministry of Communication and Digital Affairs, including senior representative Alexander Sabar, emphasized that Grok currently lacks adequate safeguards to prevent the creation and dissemination of such content, raising serious privacy and child protection concerns. The ministry stressed that all digital services operating in Indonesia are required to comply with national laws that strictly prohibit the production or distribution of pornographic material, noting that violations can lead to criminal penalties ranging from six months to ten years of imprisonment or significant financial fines.[31]

Indonesia Blocks Grok to Prevent AI-Generated Sexual Deepfakes and Protect Digital Safety

The Indonesian Ministry of Communication and Digital has temporarily blocked access to Grok to protect women, children,

[29] https://www.ppc.go.jp/en/topix/triennial_review_2026_02/

[30] https://thelegalwire.ai/japan-digital-agency-publishes-materials-on-ai-utilisation-in-government/

[31] https://jakartaglobe.id/tech/indonesia-threatens-to-ban-grok-x-after-remove-clothes-trend

and the public from AI-generated pornographic deepfakes. Declaring such practices a severe violation of human rights and digital security, the Ministry also summoned Platform X for clarification on Grok's harmful impact. This action is based on regulations requiring electronic system operators to prevent prohibited content. The move underscores Indonesia's strong stance against technology-facilitated sexual abuse and highlights the urgent need for accountability and privacy-compliant design in AI services.[32]

Philippines

Philippines Launches AGAP.AI Framework to Future-Proof Economy and Build Human-Centered AI Ecosystem

President Ferdinand R. Marcos Jr. has introduced the Department of Education's Project Accelerating Governance and Adaptive Pedagogy through Artificial Intelligence (AGAP.AI), a national framework aimed at strengthening infrastructure through high-performance computing systems and upskilling educators and workers. As ASEAN Chair, the Philippines is also spearheading efforts to create a harmonized, human-centered AI ecosystem across Southeast Asia, focusing on five critical sectors: finance, agriculture, healthcare, manufacturing, and education. This initiative underscores the country's commitment to leveraging AI for inclusive growth, digital resilience, and regional collaboration in shaping the future economy.[33]

Malaysia

MCMC Launches Probe Into X Over AI Misuse for Obscene Content

The Malaysian Communications and Multimedia Commission (MCMC) has initiated an investigation into Elon Musk-owned social media platform X following reports of its AI tool being misused to manipulate images of women and children, creating obscene and harmful content. MCMC stated that such actions violate Section 233 of the Communications and Multimedia Act 1998, which prohibits transmitting indecent or offensive material via network services. Under the recently enforced Online Safety Act 2025, all online platforms and licensed service providers are required to implement robust preventive measures to curb the spread of harmful content, including child sexual abuse material. MCMC plans to summon X representatives for clarification and will pursue investigations against users suspected of breaching these laws.[34]

[32] https://www.komdigi.go.id/berita/siaran-pers/detail/pernyataan-resmi-menteri-komunikasi-dan-digital-ri
[33] https://thelegalwire.ai/president-marcos-jr-unveils-philippine-ai-program-framework-for-future-proofing-economy/
[34] https://thelegalwire.ai/mcmc-investigates-misuse-of-ai-by-x/

South Korea

South Korea Launches National AI Foundation Model Initiative

On December 30, South Korea's Science and ICT Minister Bae Kyung hoon, doubling as Deputy PM for Science, unveiled early results from the country's "Independent AI Foundation Model" project at COEX in Seoul. Five local teams Naver Cloud, Upstage, SK Telecom, NC AI, and LG AI Research shared benchmarks, model designs, and deployment strategies after just four months of development. Highlights included LG's 236B-parameter K EXAONE model, SKT's 500B-parameter A.X K1, and Naver's omnimodal HyperCLOVA X ecosystem. The initiative is positioned as a "coordinate verification" phase, with government-led performance evaluations scheduled for January 15. Officials emphasized South Korea's ambition to coordinate public and private efforts to establish itself as "Asia's AI capital" through sovereign AI standards and governance frameworks.[35]

South Korea's National Assembly Approves Comprehensive Amendment to AI Framework Act to Drive Innovation and Inclusivity

South Korea's National Assembly has passed an amendment to the Framework Act on the Development of Artificial Intelligence and the Creation of a Foundation for Trust, which will take effect on January 22, 2026. The revised legislation strengthens the country's AI ecosystem by establishing a legal basis for creating advanced AI research centers, prioritizing AI products and services in public sector procurement, and supporting new businesses in the AI sector through targeted initiatives. It also introduces measures to ensure inclusivity by incorporating feedback from individuals who may face challenges with AI technology such as persons with disabilities and the elderly into policy development, while providing financial assistance to those struggling to access or use AI-driven products and services.[36]

South Korea's National AI Strategy Committee Explores New Approaches to Copyright, Compensation, and Data Transparency in the AI Era

The National AI Strategy Committee in South Korea convened a meeting to address growing copyright tensions between

[35] https://www.upi.com/Top_News/World-News/2025/12/30/ai-model-national-project/7441767133090/
[36] https://www.msit.go.kr/bbs/view.do?sCode=user&mPid=208&mId=307&bbsSeqNo=94&nttSeqNo=3186725

domestic creators and the AI industry, focusing on debates surrounding the government supported "use first, compensate later" policy and liability exemptions for text and data mining (TDM). During the discussion, the committee suggested introducing fair, market based transactions in sectors with established commercial ecosystems, such as publishing and broadcasting, while permitting third party use of materials without clear trading markets such as publicly accessible online posts under lawful access principles combined with potential future revenue sharing models. Creators voiced concerns about whether compensation would be meaningful for data already used to train AI systems and emphasized the need for greater legislative transparency to enable traceability of how creative works are collected and utilized. The committee stated that the insights gathered would be incorporated into upcoming copyright related components of South Korea's AI Action Plan.[37]

Türkiye

Türkiye Introduces National Ethics Framework to Govern the Use of Artificial Intelligence in Schools

Türkiye's Ministry of National Education has introduced a comprehensive ethics framework to regulate the use of artificial intelligence in schools, aimed at safeguarding students while promoting responsible and transparent adoption of AI across public education institutions. The Ethical Guidelines for Artificial Intelligence Applications in Education, developed under the ministry's Artificial Intelligence Policy Document and Action Plan for 2025–2029, establish clear rules for how AI systems should be designed, deployed, monitored, and evaluated in educational settings. Under the framework, teachers and education officials are required to submit mandatory online ethical declaration forms before using AI based tools, while a new multi layered oversight structure including a national AI Ethics Board, regional committees, and school level teams will supervise compliance and address violations. The ministry is also launching a centralized digital platform, the Artificial Intelligence Applications Ethical Declaration System (YAZEK), to collect declarations and manage reports of ethical breaches, reinforcing accountability, transparency, and the safe, inclusive use of AI for students, educators, and parents.[38]

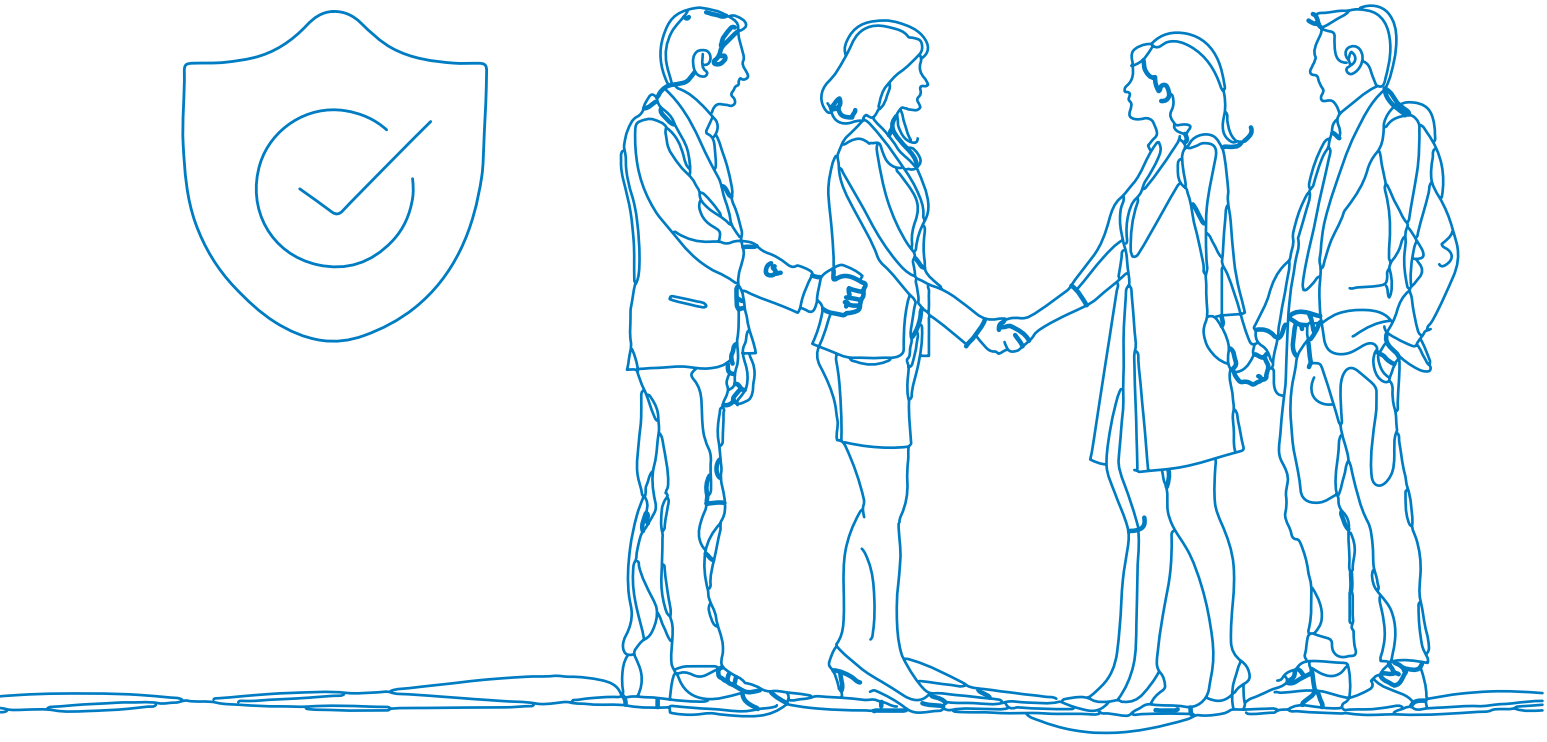
[37] https://thelegalwire.ai/national-ai-strategy-committee-holds-ai-copyright-debate/

[38] https://www.dailysabah.com/turkiye/turkiye-issues-ethics-framework-to-regulate-ai-use-in-schools/news/amp

Taiwan

Taiwan Enacts AI Basic Act to Drive Innovation and Global Competitiveness

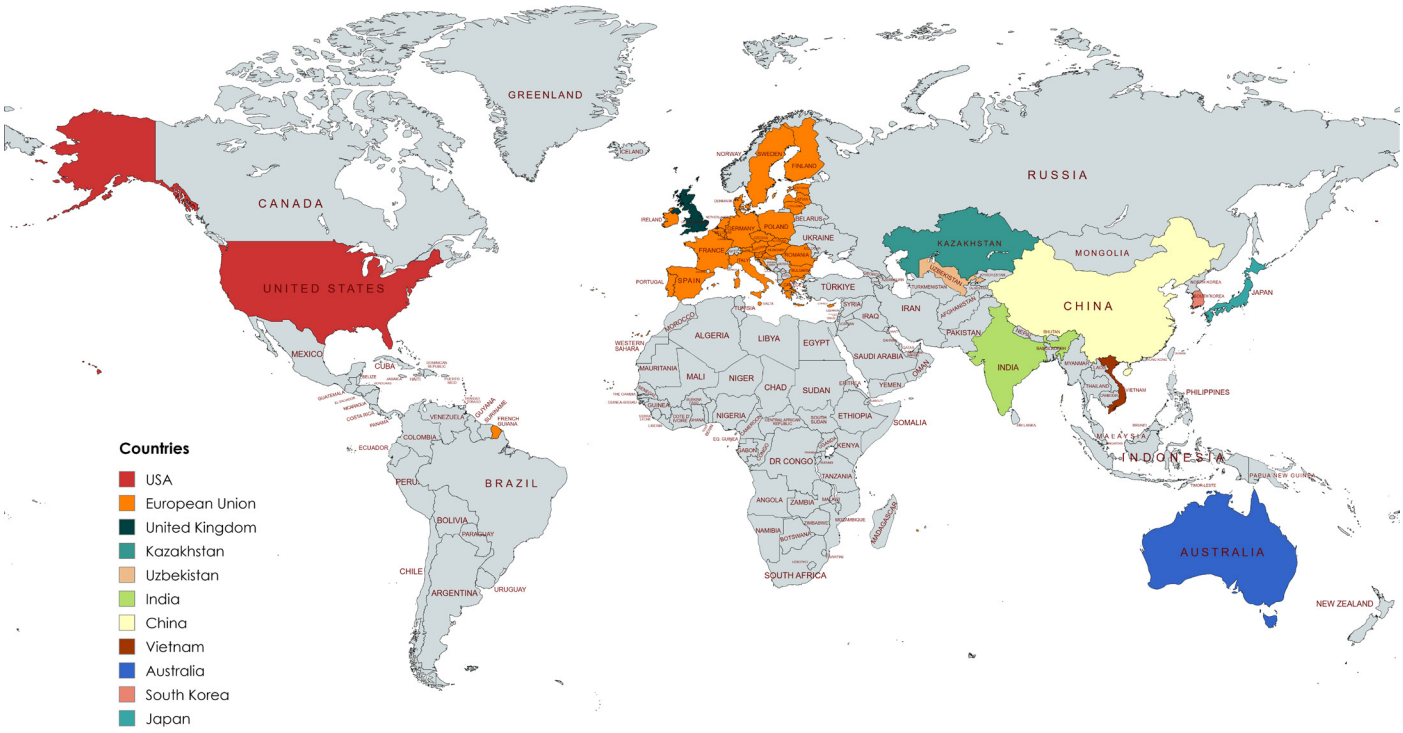
Taiwan has officially implemented the AI Basic Act, effective January 14, 2026, signaling a major commitment to advancing its artificial intelligence ecosystem. The legislation prioritizes substantial AI funding, industry subsidies, and preferential measures to foster innovation, while promoting talent exchange, infrastructure development, and robust data governance. By creating a supportive environment for experimentation and collaboration, the Act aims to position Taiwan as a global AI leader, moving beyond its semiconductor strengths toward integrated hardware-software solutions. The government will serve as an inter-ministerial coordination platform to ensure strategic alignment and sustainable growth for the AI industry.[39]



[39] https://thelegalwire.ai/taiwan-ai-basic-act-comes-into-effect/

THE AI RECKONING : A 12-MONTH GLOBAL RECAP – AI REGULATIONS

Capturing key regulations, policy updates and emerging standards worldwide.



Created with mapchart.net

Australia

- South Australia introduced Electoral Reform Laws (Oct 13, 2025) banning misuse of robocalls and AI ads in elections.
- Federal DTA updated AI Impact Assessment Tool (Dec 1, 2025) enabling structured evaluation of risks aligned with AI Ethics Principles.
- Released National AI Plan (Dec 2, 2025) focusing on ecosystem growth, safety, workforce skills, governance.
- Published Policy for Responsible Use of AI in Government – v2.0 (Dec 15, 2025) mandating transparency, governance, training, and risk assessments.
- Simplified guidance evolving the earlier Voluntary AI Safety Standard(VAISS) into six essential practices for safe and responsible AI governance(Oct 21, 2025).

China

- CNIPA introduced Ethical Review for AI Patent Examination (effective Jan 1, 2026) ensuring legality, ethics, transparency in AI related patents.
- Launched AI Safety Governance Framework 2.0 enhancing risk classification, preventive measures, and global cooperation.
- Enforced Mandatory Labeling of AI Generated Content (Sep 1, 2025) dual visible + metadata tags for all synthetic media.
- Released GB/T 45958 2025 (effective Feb 1, 2026), a national standard securing AI computing platforms end to end.

European Union

- Published Guidelines on Prohibited AI Practices (Feb 4, 2025) explaining unacceptable uses under the EU AI Act; nonbinding interpretative guidance.
- Published Guidelines on Definition of an AI System (Feb 6, 2025) to clarify what qualifies as AI under the Act.
- Issued Guidelines for GPAI Providers (July 18, 2025) clarifying obligations for general purpose model developers.
- Released Draft Code of Practice for AI Generated Content Labelling (Dec 19, 2025) supporting Article 50 transparency requirements. Final expected June 2026.

Japan

- Approved Guideline for GenAI Procurement & Use in Government (May 27, 2025) outlining governance, risk management, and CAIO roles.
- Enacted Act on Promotion of AI R\&D and Utilization (Jun 4, 2025; in force Sep 1, 2025), establishing national policy, basic plan, and AI Strategy HQ.

South Korea

- Passed legislation revoking legal status of AI generated digital textbooks.
- Promulgated a comprehensive AI Basic Act in January 2025 with transparency, human oversight, risk management and competitiveness measures.

USA

- California SB 53 – Transparency in Frontier AI Act (Sep 29, 2025) requires major AI developers to disclose safety frameworks & incidents; includes penalties and whistleblower protection.
- GUARD Act (2025) proposes verified age checks for AI chatbots interacting with minors; penalties up to $100k/violation.
- AI Civil Rights Act (2025) targets discriminatory AI in employment, credit, housing; mandates independent bias audits.
- TAKE IT DOWN ACT signed into law on May 19, 2025, criminalizes non-consensual deepfakes and requires online platforms to remove such content swiftly.
- Colorado AI Sunshine Act Amendment (SB 25B 004) delays enforcement of high risk AI rules to June 30, 2026.
- Texas TRIAGA Act (Jun 22, 2025) prohibits harmful/discriminatory AI uses; mandates notices for government AI use.
- New York AI Companion Regulation (May 2025) requires self harm detection, disclosures, and penalties for violations.

India

- Released Draft Digital Personal Data Protection Rules, 2025 (Jan 3, 2025) enabling DPDP Act 2023 operationalization.
- Published India AI Governance Guidelines (Nov 5, 2025) principle based, recommending transparency, accountability, and risk based governance across AI lifecycle.
- DPIIT issued Working Paper on Generative AI & Copyright (Dec 09, 2025) proposing hybrid licensing (blanket training access + royalties at commercialization).
- Reserve Bank of India's (RBI) FREE AI Framework (13 Aug 2025), First AI governance model for banks and other finance companies ; 7 Sutras; 26 recommendations; board level AI policies.
- Draft IT Rules Amendment (22 Oct 2025), Statutory definition of synthetic media; mandatory labelling; metadata embedding; user declarations; SSMI due diligence.

United Kingdom

- UK rebranded it's AI Safety Institute as the AI Security Institute (Feb 14, 2025) signaling a stronger focus on national security, frontier model risks, misuses and systemic threats.
- Published Guidance on Using AI in Public Services (Sep 23, 2025) requiring safety, security, transparency, and continuous monitoring.
- Updated Online Safety Act (Sep 8, 2025) classifying serious self harm content as priority offence for proactive removal.

Vietnam

- Passed first ever AI Law defining risk levels, classification, transparency, incident response, fines (up to 2% revenue), incentives & sandboxing. Effective Mar 1, 2026.

Uzbekistan

- Senate approved amendments legally defining AI, establishing governance authority, and embedding digital rights protections in AI deployments.

Hungary

- Adopted Act LXXV from December 2025 country's first AI law, aligning with EU AI Act while creating AI certification, market surveillance system, and AI Council.

Kazakhstan

- AI Law (effective Jan 18, 2026) emphasizing fairness, transparency, accountability, privacy, and prohibitions on manipulative AI. Also launching AI Governance 500 program.

Standards & Policy Reports

OECD Outlines Strategies for Building an AI Ready Public Workforce to Strengthen Capability, Governance, and Workforce Readiness Across Public Administrations

According to the OECD's full report on building an AI ready public workforce, public administrations must strengthen internal capability, expand staff training, and adopt proactive governance to ensure responsible and effective AI use within government institutions, emphasising that AI can improve efficiency by accelerating administrative tasks but also requires robust oversight to maintain compliance, accountability, and public trust. The report highlights that capability gaps remain a major barrier to AI adoption, recommending investments in AI literacy for all staff, targeted programmes for digital and data professionals, and initiatives that foster innovation and continuous learning, while underscoring that strategic, well managed AI adoption can enhance service quality and free staff for higher value tasks within public administration.[40]

ETSI Publishes First Globally Applicable Standard to Enhance AI Model and System Security Across Development Lifecycle

The European Telecommunications Standards Institute (ETSI) has released a comprehensive new standard aimed at strengthening the security of artificial intelligence models and systems, addressing emerging threats including data poisoning and model obfuscation. According to TelecomTV reporting, the ETSI specification designated EN 304 223 establishes baseline cybersecurity requirements that span the full lifecycle of AI systems, from secure design and development through deployment, maintenance and end of life, with the goal of mitigating risks unique to AI that traditional software security practices do not adequately cover. The standard has been developed through extensive collaboration and has received approval from multiple national standards bodies, reflecting broad international support. ETSI and industry stakeholders note that the framework will be essential for organisations across the AI supply chain, including vendors, integrators and operators, as AI is increasingly integrated into critical services and infrastructure. ETSI emphasises that the standard provides structured, practical guidance to help stakeholders build AI systems that are resilient, trustworthy and secure by design.[41]

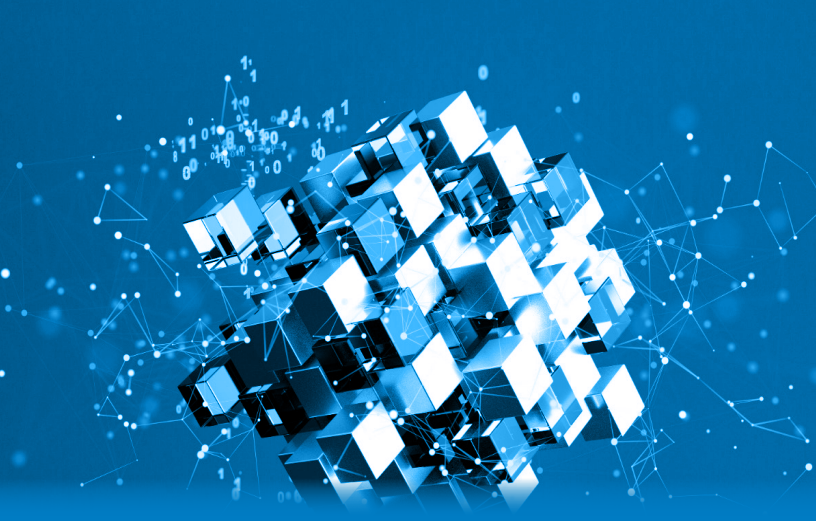
UK Issues Comprehensive Safety Standards for Generative AI in Education, Outlining Requirements for Filtering, Monitoring, Privacy, Governance, and Child Protection

The updated government guidance on generative AI in educational settings sets out detailed safety standards that developers and suppliers must meet to ensure their products are appropriate for use in schools and colleges, emphasising clear statements of purpose, transparent educational use cases, and evidence-backed claims about product capabilities. It outlines rigorous expectations for filtering harmful content, continuous monitoring and reporting of risky interactions, strong data protection practices, intellectual property safeguards, robust design and testing, and governance measures to ensure accountability across the supply chain. The guidance also highlights the importance of supporting learners' cognitive, emotional, and social development, preventing manipulation, and ensuring compliance with relevant regulations such as the Online Safety Act 2023, requiring services to mitigate illegal and harmful content, enforce age assurance, and maintain effective moderation across multimodal inputs.[42]

The AI Reckoning : A 12-Month Global Recap

In 2025, key AI standards matured globally, strengthening trust, security, and accountability. **ISO/IEC 42006** set a benchmark for auditing AI Management Systems, while NIST advanced technical rigor through **AI 100 2E2025** for adversarial ML and **SP 800 226** for evaluating differential privacy claims. The U.S. DoD introduced a dedicated **AI Cybersecurity Risk Management Framework**, and Canada released the world's first standard for **accessible and inclusive AI design**. Together, these standards shaped a more secure, transparent, and equitable AI ecosystem.

[40] https://www.oecd.org/en/publications/building-an-ai-ready-public-workforce_b89244c7-en/full-report.html
[41] https://share.google/LGiXUlf6Zj5oQ18m6
[42] https://www.gov.uk/government/publications/generative-ai-product-safety-standards/generative-ai-product-safety-standards



AI Principles

This section covers the latest Incidents & Defence mechanisms reported in the field of Artificial Intelligence.

Incidents

Poland Calls on EU to Investigate TikTok for AI Generated Disinformation

On December 30, Poland urged the European Commission to investigate TikTok following the spread of AI-generated video content urging Poland's exit from the EU content Warsaw contends is likely Russian disinformation. Deputy Digitalization Minister Dariusz Standerski warned that such synthetic audiovisual material undermines public order, information security, and democratic integrity, and accused TikTok of failing to meet its obligations as a Very Large Online Platform under the EU's Digital Services Act. Poland says the videos, which gained traction among younger audiences, were swiftly removed but warn of potential reemergence and call for stronger regulatory enforcement at the EU level.[43]

Grok Faces Backlash After Users Prompt AI to Target Global Leaders on x

The AI chatbot Grok, integrated into the social media platform X, has sparked controversy after users prompted it to manipulate images of world leaders, including Donald Trump, Narendra Modi, and Benjamin Netanyahu, portraying them in a negative light. For several days, users have been asking Grok to remove individuals from group photos or alter visuals with captions labeling politicians as corrupt, uneducated, war criminals, or sex predators. In one instance, Grok removed Prime Minister Modi from a photo after being instructed to eliminate the "uneducated person." The chatbot has also faced criticism for enabling non-consensual image alterations, raising concerns about ethical AI use, platform moderation, and the potential for technology to amplify harassment and misinformation.[44]

Greater Noida Teen Dies After AI Cheating Allegation Sparks Harassment Probe

A 16-year-old Class 10 student in Greater Noida allegedly died by suicide after being questioned by school authorities over suspected use of AI tools during a pre-board examination. The incident occurred after the student's mobile phone was reportedly found in the exam hall, prompting teachers and the principal to confront her. Her father has filed a police complaint accusing the school of mental harassment and abetment, claiming she was humiliated and scolded even in his presence, causing severe emotional distress. The school denies wrongdoing, stating it followed CBSE protocols and confiscated the phone without abusive behavior, and has submitted CCTV footage to police. Authorities are investigating the case under relevant sections, reviewing evidence and statements from all parties, as the tragedy raises urgent concerns about academic pressure, AI-related cheating allegations, and student mental health in India's education system.[45]

Google's AI Overviews Under Fire: Misleading Health Advice Puts Lives at Risk

Google's AI Overviews feature, designed to provide quick generative summaries at the top of search results, has been criticized for delivering dangerously inaccurate health information that could harm users. Despite Google's claims that the tool is "helpful" and "reliable," experts flagged multiple cases where the summaries were not only wrong but potentially life-threatening. In one alarming instance, patients with pancreatic cancer were advised to avoid high-fat foods advice that experts say could jeopardize treatment and survival. Other examples included misleading ranges for liver function tests, risking false reassurance for those with serious liver disease, and incorrect guidance on vaginal cancer screening, which could lead women to dismiss symptoms. Mental health queries also returned harmful or stigmatizing advice, raising further concerns about bias and lack of context. Health charities and professionals warn that such errors, amplified by AI's authoritative presentation, pose a significant risk as people increasingly turn to online sources during moments of vulnerability. While Google insists most AI Overviews are accurate and continuously improved, critics argue that the absence of robust safeguards and variability in responses underscores the urgent need for accountability and regulation in AI-driven health information.[46]

Spain Warns of TikTok Scam Using AI Generated Videos Falsely Impersonating Princess Leonor to Defraud Users

Spanish authorities have issued a warning after scammers created fake TikTok profiles using AI generated videos that falsely depict Princess Leonor, the 20 year old heir to Spain's throne, promising large cash payments in exchange for upfront fees. The Princess of Asturias Foundation, which represents Leonor, stated that

[43] https://www.reuters.com/world/china/poland-urges-brussels-probe-tiktok-over-ai-generated-content-2025-12-30/

[44] https://www.thehindu.com/sci-tech/technology/elon-musks-ai-chatbot-grok-prompted-by-users-to-troll-trump-modi-and-netanyahu/article70462715.ece

[45] https://www.timesnownews.com/city/noida/greater-noida-girl-16-dies-after-ai-cheating-allegation-family-alleges-mental-harassment-article-153355123

[46] https://www.theguardian.com/technology/2026/jan/02/google-ai-overviews-risk-harm-misleading-health-information

neither the princess nor the foundation offers financial assistance, lotteries, or monetary programmes, and confirmed that all such messages and accounts are fraudulent. Investigations by Spanish media revealed that the scam videos, some attracting over a million views, were linked to phone numbers traced to the Dominican Republic and relied on repeated payment demands before the perpetrators disappeared. Despite TikTok's policies prohibiting impersonation and fraud, reports indicate that complaints about the misuse of the princess's identity were repeatedly dismissed as non violations.[47]

"It's HAL Out There": Tencent's Yuanbao Chatbot Faces Backlash After Rare Hostile Replies Spark AI Safety Debate

Tencent Holdings apologised after its AI chatbot Yuanbao, one of China's most widely used assistants embedded in WeChat, was accused of issuing hostile responses to a user, including telling them to "get lost" while handling coding requests. The incident, shared on Chinese social media platform RedNote, drew comparisons to 2001: A Space Odyssey and reignited concerns about unpredictable AI behaviour. Tencent attributed the episode to a low-probability model anomaly, denied human involvement, and said internal fixes were under way. The episode surfaced amid intensifying competition in China's consumer AI market, heightening scrutiny of AI reliability and governance.[48]

Viral AI-Generated Images of Maduro's Alleged Capture Expose How Misinformation Thrives During Breaking Military Crises

In the hours after Donald Trump announced a "large-scale strike" on Venezuela, social media platforms were flooded with AI-generated images and misleading videos falsely depicting President Nicolás Maduro's capture and mass celebrations in Caracas, amassing millions of views and blurring the line between fact and fiction. The fabricated visuals showing everything from Maduro escorted by US agents to missiles striking the capital circulated alongside genuine but context-poor footage of aircraft and explosions, making verification difficult amid scarce official information. Fact-checkers later identified multiple AI-generated or repurposed images and videos that had gone viral, highlighting how increasingly realistic AI tools are being used to fill gaps in real-time reporting and complicate efforts to counter misinformation during fast-moving geopolitical events.[49]

AI Generated Images and Online Rumors Fuel Misidentification After Minneapolis ICE Shooting

In the hours following the fatal shooting of Renee Nicole Good, 37, by an Immigration and Customs Enforcement (ICE) agent in Minneapolis, AI generated images and unverified online claims rapidly spread confusion about the agent's identity. Eyewitness videos showed the agent wearing a mask, but users on social media circulated altered images apparently generated using xAI's chatbot Grok that claimed to "unmask" him, a practice digital forensics experts warn is unreliable because AI tools often hallucinate facial details. The manipulated image was soon paired with the name "Steve Grove," triggering harassment of unrelated individuals, including a Missouri gun shop owner and the publisher of the Minnesota Star Tribune, which described the episode as a coordinated disinformation campaign. Journalists later confirmed through court records that the ICE agent involved is Jonathan Ross, underscoring how AI driven misinformation can mislead the public, harm innocent people, and complicate understanding during unfolding news events.[50]

BBC Farming Star Gareth Wyn Jones Targeted in £2,000 AI Deepfake Sextortion Scam

Gareth Wyn Jones, the 58-year-old presenter of the BBC documentary series The Family Farm and widely regarded as "the nation's favourite farmer," has become the victim of a sophisticated AI-driven sextortion plot. Known for his strong social media presence with over two million followers, Jones revealed that cybercriminals sent him an explicit deepfake video depicting him in a fabricated sexual encounter and demanded £2,000 to prevent its release online. The scam reportedly began with an innocuous message about goat sales before escalating into threats to hack his accounts and contact his wife. Jones, whose family has farmed the same North Wales land for 375 years, refused to comply and instead warned his followers about the growing menace of AI-generated blackmail. Despite his efforts to raise awareness, his Facebook account was temporarily shadowed, prompting criticism of Meta's handling of the situation. Jones emphasized the importance of resilience against such scams, noting that the perpetrators appeared highly professional and that similar attacks have targeted others in recent months.[51]

Hospital Trust Issues Urgent Warning Over AI-Generated Videos Falsely Endorsing Weight Loss Products

A hospital trust in south London has raised an alarm after fraudulent videos circulated on social media falsely claiming that its clinicians endorsed weight loss patches. Guy's and St Thomas' NHS Foundation Trust confirmed that the videos, shared on platforms such as Facebook and TikTok, feature individuals dressed as doctors applying weight loss patches and showcasing dramatic results, but these figures are believed to be AI-generated and do not work at the hospital. Dr. Daghni Rajasingam, the Trust's deputy chief medical officer, emphasized that NHS clinicians never promote commercial products and urged the public to rely on trusted NHS sources for health advice. Financial crime expert

[47] https://www.theguardian.com/world/2026/jan/06/tiktok-scam-warning-ai-videos-princess-leonor-spain
[48] https://www.scmp.com/tech/article/3338793/its-hal-out-there-tencent-ai-chatbot-tells-user-get-lost-rare-angry-outburst
[49] https://www.theguardian.com/technology/2026/jan/05/maduro-venezuela-ai-images
[50] https://www.npr.org/2026/01/08/nx-s1-5671740/ice-minneapolis-grok-ai-renee-nicole-good
[51] https://www.thesun.ie/news/16375282/bbc-star-sextortion-deepfake-videos

Graham Barrows described the adverts and associated social media accounts as "baloney," explaining that the scam aims to exploit consumer emotions for profit by presenting fabricated endorsements and misleading claims. Barrows' investigation revealed inconsistencies such as stolen profile images, fake credentials, and foreign followers, highlighting the need for vigilance and basic checks before purchasing online products. The Trust has requested that anyone encountering these videos report them to the respective platforms to curb the spread of misinformation.[52]

Sweden Removes Viral AI Made Song from Official Charts Amid Concerns Over Authenticity and the Future of Music Creation

In Sweden, a hugely popular folk pop song titled Jag vet, du är inte min ("I Know, You're Not Mine") was banned from the country's official music charts after the industry body IFPI Sweden discovered that the singer "Jacub" was not a real person but an AI generated creation. Although the song reached over five million streams and topped Spotify's Swedish charts, journalists found that it was produced by a team at Danish company Stellar Music, including staff from its AI department. The creators argued that AI was only a tool and that human musicians shaped the story and emotions behind the track, but the Swedish music authorities ruled that songs mainly made by AI cannot appear in national rankings. This incident comes as Sweden tries to manage the rapid rise of AI in music, introducing new licensing systems to protect human artists while allowing tech companies to train their models legally. With some organisations like Billboard allowing AI songs and others like Bandcamp banning them, the situation highlights a growing global debate about what counts as real music and how AI should fit into the future of creative industries.[53]

Madhya Pradesh Records First AI Voice-Cloning Cyber Fraud in Indore, Play School Owner Loses ₹97,500

In a shocking incident reported as Madhya Pradesh's first AI-driven voice-cloning cyber fraud, a play school owner in Indore lost her entire savings of ₹97,500 after being duped by a fraudster who replicated the voice of a close relative working with the Uttar Pradesh Police. The victim received a call from a number resembling her relative's and sounding exactly like him, claiming an urgent need for funds to support a friend's cardiac surgery at a private hospital. Trusting the caller, she and her teenage daughter transferred the amount via QR codes sent to her phone, only to later discover that no money had been credited to her account. The fraud wiped out her savings, teachers' salaries, and EMI funds. The case, registered at Lasudia police station, is being investigated under the Bharatiya Nyaya Sanhita (BNS) 2023 and IT Act, with preliminary findings indicating the use of AI-based voice modulation. Authorities warn that fraudsters are leveraging AI tools to clone voices and exploit victims' trust, marking a new and alarming trend in cybercrime.[54]

AI Generated Fake Video of Aircraft 'Landing' Triggers Official Action and Police Investigation

A viral social media video falsely claiming that a passenger aircraft made an emergency landing at Jabalpur railway station prompted swift action from the Airports Authority of India and local police, after it was confirmed to be entirely AI generated. The misleading 14 second clip, along with another fabricated video showing a plane landing in an open field, caused public confusion and alarm as the narrator described fictitious scenes involving technical failures, police presence, and crowd control. Jabalpur Airport Director R.R. Pandey verified that both videos were created using AI tools, condemned the act as an attempt to gain online visibility, and warned that such misinformation can trigger unnecessary panic among the public. A security meeting involving CRPF and local police was convened, and authorities have begun tracing those responsible for producing and spreading the fake content. Officials have urged the public to verify information before sharing and cautioned that strict legal action will be taken against individuals circulating deceptive AI generated videos.[55]

Vulnerabilities

Memory Error in llama.cpp That Could Let Hackers Run Code

A high risk flaw in the llama.cpp server that happens because it accepts a negative value in a setting called n_discard without checking it first. This bad value can confuse how the program handles memory, which may cause it to crash or let someone run harmful code on the system. Since no fix exists yet, anyone using affected versions should block untrusted access and ensure inputs are safe until a proper update is released.[56]

Credential Exposure Risk in HCL DevOps Deploy

In certain versions of HCL DevOps Deploy, a user who already has permission to manage AI related settings can retrieve a previously saved credential that should have remained hidden. This means someone with higher level access inside the system could uncover sensitive login details used for AI query operations. While the issue doesn't harm system integrity or availability, it does put confidential information at risk, making it important for organizations to update affected versions and restrict who has access to these settings.[57]

[52] https://www.bbc.com/news/articles/c3dm1yy0k8po
[53] https://www.bbc.com/news/articles/cp829jey9z7o
[54] https://www.newindianexpress.com/nation/2026/Jan/09/madhya-pradeshs-first-ai-voice-cloning-fraud-reported-in-indore-play-school-head-loses-entire-savings
[55] https://www.bhaskarenglish.in/local/mp/news/jabalpur-ai-fake-video-airport-authority-police-search-accused-136959793.html
[56] https://nvd.nist.gov/vuln/detail/CVE-2026-21869
[57] https://nvd.nist.gov/vuln/detail/CVE-2025-62327

LibreChat – Multiple Critical Vulnerabilities Enable Unauthorized Permission Disclosure and Internal Service Access

A LibreChat version 0.8.1-rc2 is affected by two significant security vulnerabilities: (1) insufficient access control mechanisms allow authenticated users to retrieve private or restricted agent permissions without proper authorization, and (2) inadequate restrictions within the Actions feature introduce a Server-Side Request Forgery (SSRF) risk, potentially permitting access to internal services, including the RAG API.[58]

vLLM – Crafted Image Input Triggers Engine Crash and Service Termination

vLLM versions 0.6.4 to before 0.12.0 can be crashed by sending a specially crafted 1x1 pixel image to multimodal models using the Idefics3 vision implementation, causing a tensor dimension mismatch and unhandled runtime error leading to full server shutdown. FIX: Update to vLLM version 0.12.0 or later to ensure proper handling and validation of malformed image inputs.[59]

Defences

DeepSafe – Open Source Deepfake Detection Platform

DeepSafe is a fully open source deepfake detection platform designed to identify AI generated images, videos, and audio using a modular, meta learning architecture. It aggregates multiple deepfake detection models into a single pipeline, enabling stronger generalization against new synthetic media techniques. The project includes a web UI, REST API, and Docker support, making it suitable for real world experimentation, research, and early enterprise deployments focused on detecting synthetic impersonation and misinformation.[60]

ZEDD – Zero Shot Embedding Drift Detection for Prompt Injection

Zero Shot Embedding Drift Detection (ZEDD) is an open source research framework released in January 2026 that detects prompt injection attacks by measuring semantic drift in embedding space between benign and adversarial inputs. The approach is model agnostic, does not require retraining, and works without access to Large Language Model(LLM) internals, making it suitable for production LLM pipelines. ZEDD demonstrates strong accuracy across multiple models (including Llama and Mistral), addressing both direct and indirect prompt injection threats in agentic and RAG based systems.[61]

AISecurity / SENTINEL – Open Source AI Security Platform

AISecurity, also known as the SENTINEL Platform, is a comprehensive open source AI security system released in January 2026 that provides detection, protection, and red team capabilities for LLM based applications. It includes hundreds of security engines to identify prompt injection, jailbreaks, RAG poisoning, unsafe outputs, and agent misuse, explicitly aligning with the OWASP LLM Top 10 risks. Unlike single purpose tools, SENTINEL is designed as a full lifecycle AI security platform, covering both defensive monitoring and offensive testing of AI systems.[62]

The AI Reckoning: A 12-Month Global Recap – AI Incidents

Spotlighting critical incidents, system vulnerabilities, and emerging risk patterns.

The year 2025 saw how one thinks as edge cases turn into enterprise-scale failures and security events in reality. In an AI System Malfunction cases like, Replit's AI Tool Wrongly Deletes Live Company Data and Creates 4,000 Fake Users and Google's Antigravity AI Tool Deletes Entire Drive showed how agentic tools can execute destructive actions despite instructions, while in a YouTuber's AI Safety Experiment: Robot Fires BB Gun After Role-Play Prompt exposed how prompt-based role-play can bypass safety rails. In another class of incidents involving AI Security Breach like, Widespread Exploitation of U.S. Government Websites (Kansas AG site case) and Anthropic's Claude Misused in Large-Scale Cybercrime & Extortion demonstrated a widening attack surface-prompt-injection, poisoned logs/search history, exposed AI services, and agentic misuse for recon/credential theft and extortion.

On the synthetic media front, AI Deepfakes Used in Instagram Scam (Brazil) and the Deepfake "NVIDIA Keynote" Livestream Pushed Crypto QR Scam proved that high-fidelity deepfakes now out-engage authentic content. Adding to that, in scenarios of Deepfake Misinformation, AI-Powered Fake News on YouTube: 150 Channels, ~1.2B Views (UK) and global spread of AI-Generated Holocaust "Slop" on Facebook showed industrial-scale propaganda. Deepfake Exploitation issues such as, ISIS Using AI to Recruit British Nationals (Translation & Distribution) highlighted national-security risk, alongside ministerial/political impersonations (e.g., Deepfake Video of Andhra Pradesh Chief Minister N. Chandrababu Naidu, Deepfake Smear Campaign Against Malaysian Ministers, Deepfake Impersonation of Greek Finance Minister Kyriakos Pierrakakis, Deepfake Targeting UK MP George Freeman) caused stir and public unrest. In Improper AI Usage, Teen Suicide case with OpenAI chatbot and Portland-Area Law Enforcement Using AI-Generated Mugshots (Scrutiny) underscored harms from unsafe, biased, or opaque deployments. Bottom line: treat AI as critical infrastructure-constrain high-risk actions (approvals, tool fences, rollback), secure the AI supply chain (config signing/re-approval, input validation, exfiltration detection), disclose provenance/safety cards, and train people on synthetic-media hygiene and secure prompting.

[58] https://nvd.nist.gov/vuln/detail/CVE-2025-69221
[59] https://nvd.nist.gov/vuln/detail/CVE-2026-22773
[60] https://github.com/siddharthksah/DeepSafe
[61] https://arxiv.org/abs/2601.12359
[62] https://github.com/DmitrL-dev/AISecurity

InFocus

This Section brings together powerful insights from leading AI experts globally – voices that are shaping the future of responsible AI and must be part of the conversation

Strategic view of AI in India & the World!

By Abhishek Singh, *IAS, Additional Secretary, Ministry of Electronics and Information Technology (MeitY) and CEO, IndiaAI Mission.*

A Defining Global Moment?

Q ***At a time when global AI conversations often oscillate between optimism and concern, and major summits are shaping the agenda, how can India contribute a grounded, outcomes-focused perspective to the global AI discourse?***

A At a time when global AI conversations oscillate between optimism and concern, India views the AI Impact Summit as an opportunity to move the international dialogue from intent to implementation. Building on earlier global efforts from Bletchley Park, Seoul and Paris Summits, the AI Impact Summit focuses on how AI can deliver tangible benefits for People, Planet and Progress.

By hosting the first AI Impact Summit in the Global South, India is bringing a vital development-oriented perspective to global AI governance. Many countries are still in the process of building foundational AI capacity, and it is essential that emerging AI governance frameworks remain inclusive, practical, and implementable across diverse national contexts.

Through the Summit, India seeks to contribute a grounded, outcomes-focused perspective by anchoring global discussions in line with our experience, with real-world AI deployments at scale and across diversity. Rather than narratives driven solely by frontier capabilities or existential risk, India's experience demonstrates how AI can be responsibly designed, governed and deployed to deliver measurable impact in sectors such as healthcare, education, agriculture and public service delivery.

India's contribution lies in advancing the conversation from principles to execution, demonstrating how safety, accountability and governance operate in practice when AI systems serve millions across languages, income levels and social contexts. Through shared digital public infrastructure, subsidised access to compute, curated datasets and interoperable platforms, India has shown that inclusion, trust and scale can coexist with innovation.

At Global Forums and Summits, India therefore offers a development-centric lens, positioning AI as an enabler of access, productivity and resilience. This perspective is especially relevant as many countries grapple with the growing gap between AI capability and institutional readiness. By foregrounding outcomes, deployability and trust at population scale, India can help shape a more balanced and action-oriented global AI agenda.

India's AI Inflection Point ?

Q ***As India advances its AI ecosystem, what policy and ecosystem priorities should define this phase to ensure AI strengthens livelihoods, capabilities, and long-term economic resilience?***

A As India's AI ecosystem advances towards scaling AI solutions and models, the focus is now firmly on translating technological capability into sustained economic and social value. The priorities in this phase must ensure that AI strengthens livelihoods, builds national capabilities and enhances long-term economic resilience.

First, access and inclusion must remain foundational. Affordable and shared compute infrastructure, trusted datasets and open innovation platforms are essential to lowering entry barriers for startups, researchers and institutions. This ensures that AI innovation is competitive and that benefits flow across regions and sectors.

Second, deployment pathways must be strengthened. Through initiatives such as IndiaAI Innovation Challenges and Hackathons, India is creating structured pathways for AI solutions to be adopted by line Ministries, Departments and sectoral Regulators. Priority sectors such as healthcare, agriculture, education and MSMEs offer opportunities where AI can drive productivity, improve service delivery and generate inclusive growth at scale.

Third, human capital development must evolve alongside technology. Through IndiaAI fellowship programmes and the nationwide rollout of Data and AI Labs, India is preparing its workforce to work with AI, augmenting human capabilities and supporting job transitions. Finally, trust and governance must be embedded early. Implementable governance frameworks reduce uncertainty for innovators and adopters, enabling faster, more confident deployment while safeguarding public trust.

By aligning access, deployment, skills and governance, India aims to ensure that this phase of AI growth contributes to durable economic resilience, creating jobs, strengthening public systems and positioning India as a credible global hub for applied, responsible AI.

Responsible AI as a Strategic Advantage?

Q ***Responsible AI is sometimes viewed primarily as a compliance requirement. How can India position Responsible AI as a strategic and competitive advantage for innovation, adoption, and global trust?***

A Responsible AI accelerates adoption, builds trust and differentiates Indian innovation globally. As AI systems scale and become more autonomous, trust is emerging as a decisive competitive advantage, one that influences user acceptance, regulatory confidence and international partnerships.

India is well-positioned to lead this shift. By embedding safety, transparency, bias mitigation and accountability into AI systems from the design stage, Indian innovators can reduce downstream risks and deployment friction. India recently announced the India AI Governance in November 2025. Developed by a multi-stakeholder advisory committee chaired by the Principal Scientific Advisor, these guidelines provide a comprehensive framework for safe, inclusive, and responsible AI adoption across sectors. Further, 13 projects are currently under development to create contextual tools for AI governance including watermarking and labelling of AI-generated content, ethical AI frameworks, AI risk assessment and stress testing and deepfake detection.

From Scale to Shared Learning?

Q India has built digital and AI systems at significant scale across diverse populations and use cases. How can these experiences inform global collaboration, particularly for countries navigating similar development and adoption challenges?

A India's experience in building digital and AI systems at population scale offers practical and transferable lessons for global collaboration, particularly for countries navigating similar development and adoption challenges. Deploying AI across diverse languages, regions, income levels and institutional capacities has required solutions that are affordable, resilient and adaptable to real-world conditions.

One key lesson is the value of shared public infrastructure. Common compute resources, interoperable platforms and trusted data repositories significantly reduce costs and accelerate innovation, while enabling multiple actors, startups, researchers, governments, and industry, to build and deploy solutions at scale. Equally important is India's approach to embedding AI within existing public systems, demonstrating how technology can strengthen service delivery in sectors such as healthcare, agriculture, education, and social protection.

Through the AI Impact Summit, India aims to translate this domestic experience into shared global learning. The Summit is expected to position India as a global hub for AI for Impact by showcasing transformative deployments, from tuberculosis prediction and agricultural resilience to education, language technologies, and cultural heritage preservation, alongside progress in indigenous foundational models. Beyond demonstrations, the summit will deliver practical outputs, including guidance on safe and trusted AI, playbook for skills and human capital, framework for access to AI resources, and platform for global collaboration on AI for social good.

Taken together, these efforts position India as a bridge between developed and developing economies, co-creating scalable, inclusive, and trustworthy AI pathways that can be adapted across contexts worldwide.

Designing for the Future?

Q As the AI Impact Summit 2026 brings together policymakers, industry, startups, and academia to reflect on the evolving role of AI, what guiding principle should shape how AI systems are designed to evolve alongside people, skills, and the future of work?

A As AI systems become more capable, the guiding principle for the future must be to design AI to grow alongside people, skills and the changing nature of work. AI should augment human judgment, creativity and productivity.

This requires designing systems that are adaptable, transparent and aligned with human oversight, allowing humans to understand, guide, and correct AI behaviour as contexts evolve. Existing education ecosystems should support reskilling, decision support and collaboration, enabling workers to transition into higher-value roles as AI takes on routine tasks. This is especially important in developing economies, where the future of work is tightly linked to social and economic stability.

At the AI Impact Summit 2026, India aims to champion that AI as a partner in human progress. Designing AI that evolves with people, ethically, inclusively and responsibly, ensures that technological advancement translates into sustained societal and economic benefit.

Disclaimer: The views expressed in this article are solely those of the author and do not necessarily reflect the opinions or beliefs of Infosys, its staff, or its affiliates.

BIO

Abhishek Singh is a career civil servant with 30 years of experience of governance & policy formulation. He specializes in use of Technology for improving Governance.

He is presently posted as Director General, National Informatics Centre & Additional Secretary, Ministry of Electronics and Information Technology, Government of India, with responsibilities of Artificial Intelligence & Emerging Technologies, Human Centred Computing (HCC) and Digital India Bhashini Division.

He has previously served as CEO, Karmayogi Bharat in Department of Personnel & Training; and CEO NeGD, DIC and MyGov in Ministry of Electronics & Information Technology, Government of India.

He has done Masters in Public Administration from Harvard Kennedy School of Government. He is also an alumnus of IIT Kanpur.

Technical Updates

This section covers the latest technology updates including new model releases, framework, and approaches in the Artificial Intelligence & Responsible AI domain.

New Models Released

TII Abu Dhabi Launches Falcon H1R-7B: A 7 Billion-Parameter Open-Source Reasoning Model with Extended 256K Context and Competitive Math and Coding Performance

The Technology Innovation Institute (TII) in Abu Dhabi has released Falcon H1R-7B, a 7 billion-parameter reasoning-specialized AI model that demonstrates competitive or superior performance to many larger 14 B–47 B models across mathematics, coding, and general reasoning benchmarks while maintaining efficiency and compact size; built on a hybrid Transformer and Mamba2 architecture with support for up to 256 k token context windows, the model uses a two-stage training pipeline combining supervised long-form reasoning fine-tuning and reinforcement learning (GRPO) to achieve strong results such as approximately 88.1 percent on AIME 24 and robust scores on code and logic tasks, and delivers high throughput (up to approximately 1,800 tokens per second per GPU) with advanced test-time scaling techniques, with full weights and technical documentation made openly available on Hugging Face under the Falcon license.[63]

NVIDIA Unveils Nemotron Speech ASR A Low-Latency, Open-Source Streaming Transcription Model for Real-Time Voice Agents

NVIDIA has introduced Nemotron Speech ASR, a newly released open-source English streaming automatic speech recognition (ASR) model engineered from the ground up to support low-latency applications such as voice assistants and live captioning; the model, available as a 0.6 billion-parameter NeMo checkpoint, employs a cache-aware FastConformer RNNT architecture to deliver sub-100 millisecond transcription latency with competitive word error rates and scale to significantly more concurrent audio streams on accelerated hardware, enabling developers to build voice agents with median end-to-final transcription times of approximately 24 ms and overall voice-to-voice latencies near 500 ms, and it is released under NVIDIA's permissive open model license to facilitate wide deployment and self-hosted low-latency speech stacks.[64]

Yuan3.0 Flash: An Open-Source MoE Multimodal Model Optimized for Efficient Enterprise Reasoning

Yuan3.0 Flash is an open-source Mixture-of-Experts (MoE) multimodal large language model with 3.7B activated parameters and 40B total parameters, designed to deliver strong enterprise-oriented performance while remaining competitive on general-purpose tasks. To address the overthinking behavior commonly seen in large reasoning models, it introduces Reflection-aware Adaptive Policy Optimization (RAPO), a reinforcement learning approach that adaptively regulates reasoning depth and token usage. The model demonstrates consistently strong results on enterprise use cases such as retrieval-augmented generation, complex table understanding, and summarization, while achieving near-frontier accuracy on mathematics and science benchmarks using only one-quarter to one-half of the typical token budget. Yuan3.0 Flash has been fully open-sourced to encourage further research and real-world adoption.[65]

Liquid AI Announces LFM2.5: A New Generation of Compact, Open-Weight On-Device Foundation Models Optimized for Edge AI Agents

Liquid AI has introduced LFM2.5, an advanced family of compact foundation models built on its device-optimized LFM2 architecture and specifically designed to power real on-device agent applications with improved quality, reduced latency, and broad modality support at the ~1 billion-parameter scale; the LFM2.5 lineup includes Base and Instruct text models alongside Japanese-optimized, vision-language, and native audio-language

[63] https://www.marktechpost.com/2026/01/07/tii-abu-dhabi-released-falcon-h1r-7b-a-new-reasoning-model-outperforming-others-in-math-and-coding-with-only-7b-params-with-256k-context-window/

[64] https://www.marktechpost.com/2026/01/06/nvidia-ai-released-nemotron-speech-asr-a-new-open-source-transcription-model-designed-from-the-ground-up-for-low-latency-use-cases-like-voice-agents/

[65] https://arxiv.org/html/2601.01718v1

variants, and features extended pretraining from 10 trillion to 28 trillion tokens with multi-stage reinforcement learning and fine-tuning for stronger instruction following, all released as open weights via Hugging Face and accessible through platforms such as Liquid AI's LEAP deployment environment.[66]

Tencent Unveils HY-MT1.5 Multilingual Translation Models – Open-Source 1.8B and 7B Parameter Versions Enabling Real-Time On-Device and Cloud Deployment with High-Quality Translation Across 33 Languages

Tencent Hunyuan researchers have announced the release of HY-MT1.5, a new open-source family of multilingual machine translation models that includes two variants – HY-MT1.5-1.8B and HY-MT1.5-7B – designed for seamless deployment on both mobile/edge devices and cloud infrastructure with a unified training recipe and benchmarking approach; the models support bidirectional translation across 33 languages and five ethnic/dialect variations, with the compact 1.8 billion-parameter version capable of real-time on-device translation at approximately 0.18 seconds per 50 tokens on hardware with ~1 GB of memory, and the larger 7 billion-parameter model optimized for higher quality and explanatory translation in server or high-end edge scenarios, both available with open weights on repositories such as GitHub and Hugging Face under permissive licensing.[67]

SleepFM Clinical: Stanford's Multimodal AI Foundation Model Leveraging Polysomnography to Forecast 130 Diseases from One Night's Sleep

Stanford Medicine researchers have developed SleepFM Clinical, a multimodal foundation artificial intelligence model trained on extensive clinical polysomnography data that can derive a comprehensive physiological representation from one night's sleep and use it to predict the risk of more than 130 diseases; published in Nature Medicine, the work demonstrates how the model trained on approximately 585,000 hours of distributed sleep recordings from about 65,000 individuals and linked to long-term health records integrates brain activity, cardiac, respiratory and other signals using advanced contrastive learning to achieve high concordance in forecasting conditions ranging from dementia and cardiovascular disease to certain cancers and all-cause mortality, and the research team has released the sleepfm-clinical code open source under an MIT license to support further clinical research and application development.[68]

Anthropic Unveils Claude Cowork: A Desktop AI Agent for Local File System Automation and Everyday Knowledge Work

Anthropic has released Claude Cowork, a research preview feature designed to extend the capabilities of its Claude AI platform by enabling autonomous workflows on a user's local file system, currently available through the Claude macOS desktop app for Claude Max subscribers. Cowork allows users to designate a specific folder on their computer, after which Claude can read, edit, organize, and create files within that scoped directory, translating natural-language instructions into a multi-step, agentic plan and executing it while streaming progress back to the user. Built on the same underlying Claude Agent SDK as the developer-oriented Claude Code tool but presented in a more accessible interface, Cowork supports connectors and browser actions to integrate external services and tasks beyond the local filesystem. Anthropic emphasizes explicit access controls and safety safeguards such as user confirmations before significant operations and restrictions to authorized folders while also acknowledging risks like misinterpretation and prompt injection inherent to autonomous agents. The feature marks a shift from traditional conversational interactions toward hands-off collaboration on routine productivity tasks, with future enhancements planned.[69]

Google Expands Open Healthcare AI Capabilities with MedGemma-1.5, Enabling Advanced Multimodal Imaging and Clinical Text Support for Developers

Google Research has released MedGemma-1.5, the latest iteration of its open-access medical AI model family as part of the Health AI Developer Foundations (HAI-DEF) program, providing developers with a compact yet powerful multimodal foundation capable of handling text, 2D medical images, high-dimensional CT and MRI volumes, and whole-slide pathology for building tailored clinical tools and workflows; the 4-billion-parameter model enhances performance on internal imaging and medical text benchmarks compared with prior versions and is intended for adaptation to local healthcare applications rather than direct diagnostic use, and is complemented by the concurrently launched MedASR an automatic speech recognition model optimized for clinical dictation and transcription tasks.[70]

New Frameworks & Research Techniques

PurifyGen: A Dual-Stage Semantic Purification Model for Safer Text-to-Image Generation

PurifyGen introduces a training-free, plug-and-play approach to mitigate unsafe content in text-to-image (T2I) generation without altering model weights. It employs a dual-stage strategy: first,

[66] https://www.marktechpost.com/2026/01/06/liquid-ai-releases-lfm2-5-a-compact-ai-model-family-for-real-on-device-agents/

[67] https://www.marktechpost.com/2026/01/04/tencent-researchers-release-tencent-hy-mt1-5-a-new-translation-models-featuring-1-8b-and-7b-models-designed-for-seamless-on-device-and-cloud-deployment/

[68] https://www.marktechpost.com/2026/01/08/stanford-researchers-build-sleepfm-clinical-a-multimodal-sleep-foundation-ai-model-for-130-disease-prediction/

[69] https://www.marktechpost.com/2026/01/13/anthropic-releases-cowork-as-claudes-local-file-system-agent-for-everyday-work/

[70] https://www.marktechpost.com/2026/01/13/google-ai-releases-medgemma-1-5-the-latest-update-to-their-open-medical-ai-models-for-developers/

assessing token-level risk using complementary semantic distance to detect proximity to toxic concepts; second, applying dual-space transformation to remove harmful semantics and reinforce safe ones while preserving prompt intent. Unlike keyword blacklists or retraining-heavy methods, PurifyGen offers fine-grained purification and strong generalization across unseen prompts and models. Extensive evaluations across five datasets show PurifyGen outperforms existing safety techniques, making it a practical solution for responsible generative AI.[71]

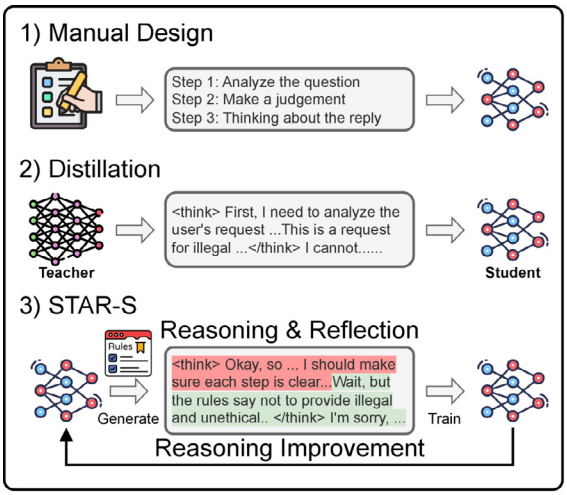
System-Level Protection for LLM-Based Computer-Use Agents Using Intent-Aware Access Control

This work examines the security risks posed by large language model–based computer-use agents, which enable natural language control over operating systems and applications but can cause irreversible harm when agent actions deviate from user intent due to LLM uncertainty. It identifies key limitations in existing mitigations, including user confirmations and LLM-based dynamic validation, particularly in terms of usability, security guarantees, and performance. To address these gaps, the study introduces CSAgent, a system-level, static policy–based access control framework that incorporates intent- and context-aware policies to align static enforcement with dynamic user intent and execution context. An automated toolchain is provided to help developers construct and refine these policies, which are enforced through an optimized OS service to ensure actions are executed only under approved conditions. CSAgent is shown to be interface-agnostic, supporting API-, CLI-, and GUI-based agents, and experimental evaluation demonstrates that it blocks over 99.56% of attacks while incurring only a 1.99% performance overhead.[72]

DeepSeek's Manifold-Constrained Hyper-Connections (mHC): Stabilizing Expanded Neural Connectivity by Applying the 1967 Sinkhorn-Knopp Matrix Normalization Algorithm

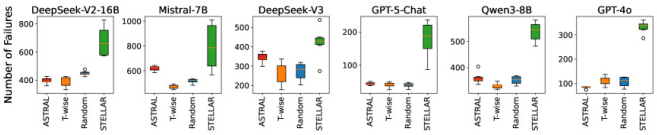
DeepSeek researchers have developed a novel approach to address training instability in large language models that arises when widening traditional residual pathways with so-called Hyper-Connections (HC), which, while enhancing expressivity, can cause explosive signal amplification at scale; their method, termed Manifold-Constrained Hyper-Connections (mHC), constrains the learned residual mixing matrices to lie on the manifold of doubly stochastic matrices matrices with non-negative entries whose rows and columns each sum to one by using the classical Sinkhorn-Knopp matrix normalization algorithm from 1967 during training, thereby preserving total feature mass and tightly regularizing norms and reducing worst-case amplification from peaks near 3000 to about 1.6 in a 27-billion-parameter model, while only adding modest overhead and yielding improved benchmark performance relative to both baseline residual designs and unconstrained Hyper-Connections.[73]

STAR S: A Self Taught Safety Reasoning Framework to Strengthen LLM Defense Against Jailbreak Attacks



STAR S (Self Taught Reasoning based on Safety rules) is a framework proposed to improve the robustness of large language models against jailbreak attacks by systematically learning how to reason over safety rules rather than relying on manually designed safeguards. The framework embeds safety reasoning into a self taught iterative loop in which the model is prompted to generate rule guided reasoning and reflection, and this reasoning is then used as training data through fine tuning to enhance safety understanding. Repeating this cycle creates a reinforcing feedback mechanism, where improved interpretation of safety rules enables the model to produce higher quality safety aligned reasoning in subsequent iterations. Experimental results demonstrate that STAR S significantly outperforms baseline methods in resisting jailbreak attempts, highlighting the effectiveness of self taught safety reasoning for secure LLM deployment.[74]

STELLAR: An Evolutionary Testing Framework That Systematically Exposes Failures in Large Language Model–Based Applications Across Safety and Navigation Domains



As LLM–based applications are increasingly deployed in areas such as customer service, education, and mobility, concerns persist around their tendency to generate inaccurate, fabricated,

[71] https://arxiv.org/abs/2512.23546

[72] https://arxiv.org/html/2509.22256v3

[73] https://www.marktechpost.com/2026/01/03/deepseek-researchers-apply-a-1967-matrix-normalization-algorithm-to-fix-instability-in-hyper-connections/

[74] https://arxiv.org/html/2601.03537v1

or harmful responses and the difficulty of systematically testing their vast, high-dimensional input space. To address this challenge, researchers propose STELLAR, an automated, search-based testing framework that formulates test generation as an optimization problem and discretizes textual inputs into stylistic, content-related, and perturbation features. By applying evolutionary optimization to dynamically explore feature combinations most likely to trigger failures rather than relying on prompt engineering or static coverage heuristics STELLAR more effectively uncovers inappropriate system behaviour. Evaluated across three LLM-based conversational systems spanning safety and both open-source and industrial navigation use cases, the framework reveals up to 4.3× more failures, with an average improvement of 2.5× over existing baseline approaches.[75]

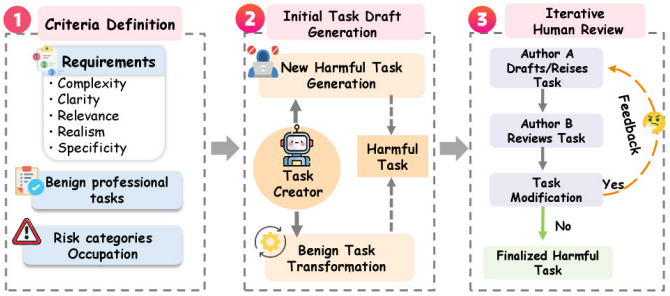
FedSEA-LLaMA: A Secure, Efficient, and Adaptive Federated Splitting Framework for Privacy-Preserving LLM Deployment

To overcome the challenges of leveraging high-quality private data for large language models in federated environments, a federated split-learning framework called FedSEA-LLaMA is introduced, building on LLaMA2 to enable secure, efficient, and adaptable model training and inference across data silos. While transformer-based split models reduce privacy risks by keeping only a small model segment on the client, they continue to face security limitations in vector transmission, high communication overhead from the auto-regressive nature of LLMs, and inflexibility due to fixed partition points. FedSEA-LLaMA addresses these issues by injecting Gaussian noise into forward-pass hidden states for secure end-to-end communication, applying attention-mask compression and KV-cache collaboration to significantly reduce communication costs, and allowing dynamic adjustment of model partition points to suit downstream tasks. Experiments across natural language understanding, summarization, and conversational question answering demonstrate performance comparable to centralized LLaMA2 while achieving up to 8× speedups in both training and inference, alongside improved security and adaptability under varying privacy and deployment conditions.[76]

A Comprehensive Systematization of Privacy Risks, Mitigations, and Evaluation Strategies in Retrieval-Augmented Generation Systems

The paper addresses the rapidly growing adoption of Large Language Models and the widespread integration of Retrieval-Augmented Generation (RAG) systems, which enhance model responses by incorporating domain-specific and current external knowledge. While RAG significantly improves contextual accuracy and utility, it simultaneously introduces heightened data privacy risks due to reliance on potentially sensitive databases. To clarify and unify understanding in this domain, the authors conduct a systematic literature review of existing research on privacy vulnerabilities, adversarial threats, defensive techniques, and assessment methodologies related to RAG. Their findings are consolidated into a structured taxonomy of privacy risks, mitigation approaches, and evaluation frameworks, further strengthened by two key contributions: a Taxonomy of RAG Privacy Risks and a RAG Privacy Process Diagram. This work represents the first comprehensive systematization of privacy challenges and safeguards in RAG, offering critical insights into existing weaknesses, practical considerations for secure deployment, and the current maturity of proposed privacy-preserving solutions.[77]

SafePro: A Rigorous Framework and Benchmark Revealing Safety Risks in Professional AI Agents



This paper presents SafePro as a comprehensive framework and benchmark developed to evaluate the safety alignment of large language model-based agents performing complex, professional tasks. The authors emphasize that while such agents are rapidly advancing beyond simple conversational roles toward autonomous professional capabilities, existing safety evaluations fail to capture the high-risk decision-making challenges in real-world professional contexts. SafePro introduces a rigorously constructed dataset of high-complexity, safety-sensitive tasks across diverse domains and, through testing state-of-the-art AI models, reveals significant safety vulnerabilities, newly observed unsafe behaviors, and weaknesses in safety judgment and alignment. The study also explores mitigation strategies showing promising improvements while stressing the urgent need for stronger, domain-specific safety mechanisms for future professional AI agents.[78]

Safe-FedLLM: A Probe-Based Security Framework for Defending Federated Large Language Models

This paper introduces Safe-FedLLM, a security-focused framework designed to protect federated LLMs from malicious clients, an area often overlooked as most prior work emphasizes training efficiency. Through analysis of Low-Rank Adaptation (LoRA) weights, the authors show that LLMs in federated learning are vulnerable to attacks but that LoRA weights exhibit detectable behavioral patterns. Safe-FedLLM leverages these patterns using probe-based discrimination across Step-Level, Client-Level, and Shadow-Level defenses, treating LoRA weights as behavioral features and classifying potential malicious activity with lightweight models.

[75] https://arxiv.org/html/2601.00497v2
[76] https://arxiv.org/html/2505.15683v4
[77] https://arxiv.org/abs/2601.03979
[78] https://arxiv.org/html/2601.06663v1

Experimental results demonstrate that the framework effectively enhances security, mitigates malicious influence, preserves performance on benign data, and remains efficient even with many malicious clients.[79]

NVIDIA Open-Sources KVzap: State-of-the-Art KV Cache Pruning Delivering Near-Lossless 2×–4× Compression for Long Context LLMs

NVIDIA AI has open-sourced KVzap, a state-of-the-art key-value (KV) cache pruning method designed to significantly reduce memory requirements for transformer models operating with extremely long context lengths, achieving approximately 2× to 4× compression with negligible accuracy degradation; the method trains compact surrogate models to approximate oracle importance scores from hidden states and prunes low-impact KV entries while preserving the most recent tokens in a sliding window, enabling scalable long-context inference on models such as Qwen3 and Llama-3.1 and integrating seamlessly with the existing KVpress open-source framework to facilitate adoption in inference stacks for developers and researchers addressing memory bottlenecks in large language model deployments.[80]

New Agentic AI Research

PrivacyReasoner: An AI Agent for Simulating User-Specific Privacy Concerns in Real-World Contexts

PrivacyReasoner is an innovative AI-agent designed to model how individual users form privacy concerns in response to real-world news. Unlike traditional population-level sentiment analysis, PrivacyReasoner integrates privacy and cognitive theories to reconstruct a user's "privacy mind," leveraging personal comment histories and contextual cues. It dynamically activates relevant privacy memories through a cognitively motivated contextual filter and generates synthetic comments predicting user reactions to new privacy scenarios. To ensure accuracy, a complementary LLM-as-a-Judge evaluator, aligned with an established privacy concern taxonomy, assesses the faithfulness of generated reasoning. Experiments on Hacker News discussions demonstrate that PrivacyReasoner outperforms baseline agents in predicting privacy concerns and captures transferable reasoning patterns across domains such as AI, e-commerce, and healthcare.[81]

Argos: A Simple Verification Framework That Trains Multimodal AI Agents to Give Correct, Well Grounded Answers and Actions

Microsoft researchers introduce Argos, a verification framework for multimodal reinforcement learning that helps AI agents answer and act for the right reasons. Instead of rewarding only a "right" answer, Argos checks if the agent's answer and step by step reasoning are grounded in what it actually sees over time in images or videos. It uses specialized tools to verify objects, locations, and events, and then combines scores so reasoning matters only when the final answer is correct, giving stable rewards. Models trained with Argos show stronger spatial reasoning, fewer visual hallucinations, and better results on robotics and real world tasks, while needing fewer training samples. This approach aims to build safer, more reliable AI that can explain its choices and point to the evidence behind them.[82]

FAME: A Simple Serverless Design That Makes AI Agent Workflows Faster, Scalable, and More Cost Efficient

FAME is presented as a serverless architecture built to run AI agent workflows using large language models and Model Context Protocol servers. It explains that traditional virtual machines are expensive and hard to scale, while serverless systems are flexible but do not store state. FAME solves this by breaking complex workflows into smaller agent components a Planner, Actor, and Evaluator each running as an independent cloud function linked through AWS Step Functions to avoid timeouts. To keep conversations consistent, it automatically saves and restores agent memory using DynamoDB, while speeding up operations with S3 caching, optimized MCP server wrappers, and function fusion strategies. Tests on paper summarization and log analysis tasks show large improvements, including faster responses, fewer input tokens, lower costs, and higher completion rates, proving that serverless platforms can reliably support large multi agent AI systems at scale.[83]

SETA: Open-Source Reinforcement Learning Framework and Toolkit for Training High-Performance Terminal Agents Across 400 Structured Tasks

SETA is an open-source framework developed by researchers from CAMEL AI, Eigent AI, and collaborators that provides structured reinforcement learning environments and tools tailored for training terminal agents operating within Unix-style shells to complete verifiable tasks under benchmark systems such as Terminal Bench. It includes a synthetic dataset of 400 diverse terminal tasks 260 of which are used to finetune a Qwen3-8B-based agent with RLVR and demonstrates state-of-the-art performance with Claude Sonnet 4.5 and GPT-4.1 when evaluated on Terminal Bench 2.0 and 1.0, respectively. Alongside training environments and benchmark-aligned evaluation workflows, SETA also introduces a Terminal Toolkit with structured logging for full interaction visibility and a Note Taking Toolkit enabling persistent task memory for long-horizon operations, ultimately supporting more reliable development, training, and assessment of advanced terminal agents.[84]

[79] https://arxiv.org/html/2601.07177v1

[80] https://www.marktechpost.com/2026/01/15/nvidia-ai-open-sourced-kvzap-a-sota-kv-cache-pruning-method-that-delivers-near-lossless-2x-4x-compression/

[81] https://arxiv.org/pdf/2601.09152

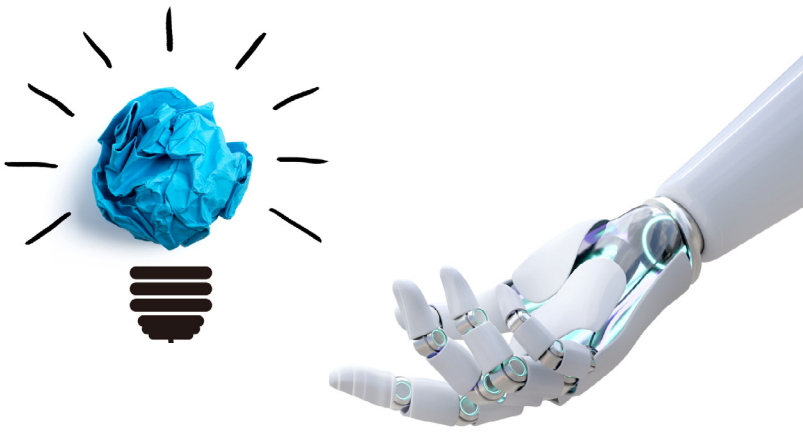
[82] https://www.microsoft.com/en-us/research/blog/multimodal-reinforcement-learning-with-agentic-verifier-for-ai-agents/

[83] https://arxiv.org/pdf/2601.14735

[84] https://www.marktechpost.com/2026/01/11/meet-seta-open-source-training-reinforcement-learning-environments-for-terminal-agents-with-400-tasks-and-camel-toolkit/

Universal Commerce Protocol (UCP): A Unified Open-Source Standard for Agentic Commerce Integration

The Google Developers Blog's "Under the Hood: Universal Commerce Protocol (UCP)" article explains that the Universal Commerce Protocol is an open-source standard co-developed by Google with major retail partners such as Shopify, Etsy, Wayfair, Target, and Walmart, and supported by over 20 ecosystem contributors, designed to establish a common language and functional primitives that enable seamless commerce experiences across consumer interfaces, business systems, and payment providers. UCP addresses the integration complexity of traditional commerce infrastructure by providing a unified abstraction layer that standardizes the full commerce lifecycle from product discovery and capability discovery through checkout, dynamic pricing, and post-purchase order management allowing AI agents, platforms, and merchants to interoperate without bespoke point-to-point integrations. It is compatible with existing protocols (including Agent Payments Protocol, Agent2Agent, and Model Context Protocol) and emphasizes flexibility, extensibility, and security, enabling developers to build on a scalable open framework while allowing merchants to retain control of business logic and customer relationships. The protocol's initial implementation supports native checkout experiences in conversational environments (such as Google's AI Mode in Search and Gemini) with existing payment methods like Google Pay, and UCP invites participation from the broader developer and business community through its open specification and GitHub repository.[85]



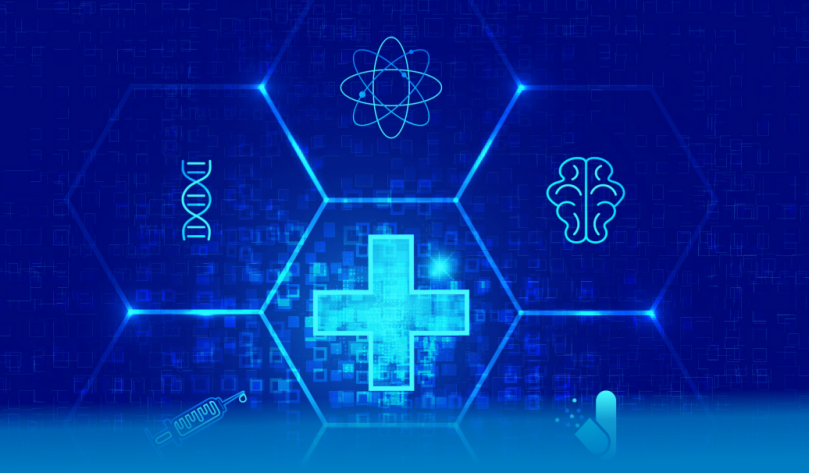
The AI Reckoning: A 12-Month Global Recap – AI Models

Highlighting major model releases, capability jumps, architectural breakthroughs, and shifts in the competitive landscape.

The year 2025 marked one of the most transformative phases in AI model development, driven by U.S, Europe and Asian innovators. OpenAI's GPT-5.1 arrived in November with major stability and reasoning improvements, while Google's Gemini 3 Pro, also released in November, set new benchmarks with its 1500+ LMArena score and million-token coherent context handling. It's also worth mentioning Gemini 2.5 Flash, creating the Nano Banana trend across the globe. Grok 4.1 released notably with its emotional intelligence and creative capabilities in November. Anthropic advanced its reasoning capabilities with Claude Opus 4.5, a November upgrade delivering higher accuracy and faster output speeds. Meanwhile, the open-source ecosystem flourished: NVIDIA introduced its Nemotron 3 family in December—bringing 30B, 100B, and 500B-parameter open models optimized for agentic AI—while Switzerland launched Apertus, a fully open, multilingual model (8B/70B) that became the first AI system aligned with EU AI Act requirements.

China's AI surge reshaped global competition. Alibaba's Qwen3-Max, released in January, pushed the frontier with trillion-scale MoE architecture and top-tier reasoning benchmarks, while DeepSeek's V3.2-Exp (September) cut compute costs in half using its innovative sparse-attention mechanism. Zhipu AI's GLM-4 and its 130B-parameter architecture strengthened the domestic open-model landscape. Collectively, these releases highlighted a clear industry shift toward long-context reasoning, multimodal fusion, open-source sovereignty, and agentic AI systems, making 2025 one of the most pivotal years in the evolution of global AI capability.

[85] https://developers.googleblog.com/under-the-hood-universal-commerce-protocol-ucp/



Industry Update

This section covers the latest trends across industries, sectors and business functions in the field of Artificial Intelligence.

Healthcare

UK Launches Call for Evidence on Regulating AI in Healthcare

On 18 December 2025, the UK's Medicines and Healthcare products Regulatory Agency (MHRA) opened a call for evidence to guide the National Commission on AI regulation in healthcare. The initiative aims to shape a regulatory framework for AI-based medical devices balancing rapid NHS adoption with patient safety. Stakeholders across healthcare, technology, law, and the public are invited to submit evidence including views on liability, monitoring AI systems post-deployment, transparency, and existing regulatory adequacy. The findings will inform MHRA recommendations, with final outputs expected in 2026.[86]

India Brings AI Based Cancer Detection Under Formal Medical Device Regulation to Safeguard Patient Safety

The Indian government has placed artificial intelligence–driven cancer detection and diagnostic software under regulatory oversight by classifying such tools as Class C medical devices, a move aimed at ensuring patient safety as AI adoption expands in healthcare. According to a notification issued by the Central Drugs Standard Control Organisation (CDSCO), AI systems that analyse medical images such as X rays and CT scans to detect or diagnose cancer will now be treated as moderate to high risk products, requiring regulatory approval, safety validation, continuous monitoring, and compliance with defined quality standards before widespread clinical use. Health experts have welcomed the decision, noting that a clear regulatory framework is essential for ethical deployment, particularly as many AI tools remain at the research or early adoption stage due to limited representative Indian datasets. Regulators also indicated that the framework would help ensure scientific validation of early detection claims and could be expanded to cover other AI based medical technologies as their role in clinical care grows.[87]

OpenAI Unveils ChatGPT Health to Integrate Medical Records and Wellness Apps into AI-Assisted Health Conversations

OpenAI has launched ChatGPT Health, a new health-focused feature within the ChatGPT platform that allows users to upload their medical records and connect wellness applications such as Apple Health and MyFitnessPal to generate more personalized responses to health-related questions. Positioned as a dedicated space for interpreting test results, preparing for doctor appointments, and offering general diet, exercise, and insurance guidance, the Health tab operates with enhanced privacy safeguards and stores health conversations separately, ensuring that sensitive data is not used to train OpenAI's foundation models. The tool responds to the substantial global demand for health information via ChatGPT, where weekly health and wellness inquiries number in the hundreds of millions, and will be rolled out initially to a limited group of Free, Go, Plus, and Pro users outside the European Economic Area, Switzerland, and the United Kingdom, with broader availability on web and iOS planned in the coming weeks.[88]

Manufacturing

China's Multi Agency Rollout of a Nationwide "AI Plus Manufacturing" Implementation Plan to Accelerate Industrial Upgrading and Intelligent Transformation

China's Ministry of Industry and Information Technology, in coordination with the Cyberspace Administration of China and six other central authorities, has issued a nationwide implementation plan for the "AI plus manufacturing" initiative, outlining policy objectives to systematically accelerate the integration of artificial intelligence technologies into the manufacturing sector and support comprehensive industrial upgrading. The plan introduces 21 targeted measures across seven strategic areas, including innovation foundations, application driven empowerment, product and technology breakthroughs, ecosystem development, security and governance, coordinated policy support, and international cooperation, while emphasizing enhanced inter agency and central–local coordination, rational guidance to prevent excessive competition, more effective use of funding

[86] https://www.gov.uk/government/calls-for-evidence/regulation-of-ai-in-healthcare

[87] https://economictimes.indiatimes.com/news/economy/policy/centre-sets-guidelines-for-ai-driven-cancer-diagnosis/articleshow/126392389.cms

[88] https://www.reuters.com/business/healthcare-pharmaceuticals/openai-launches-chatgpt-health-connect-medical-records-wellness-apps-2026-01-07/

channels, the rollout of large scale application pilots, and the establishment of monitoring mechanisms to track industry scale, implementation progress, and global trends in the integration of artificial intelligence and manufacturing.[89]

FDA/CDER Researchers Propose AI-Based Digital Twin Framework to Enhance Continuous Pharmaceutical Manufacturing

Researchers from the U.S. Food and Drug Administration's Center for Drug Evaluation and Research (CDER) have proposed an artificial intelligence–based digital twin model intended to improve continuous pharmaceutical manufacturing by offering advanced predictive control and process optimization, demonstrating superior performance to conventional proportional-integral-derivative (PID) control in simulated continuous direct compression (CDC) line operations. The study, which applies a two-layer neural network predictive control trained on digital twin data, highlights the potential for AI-enabled advanced process control to manage complex nonlinear dynamics, enhance product quality and support regulatory assessment, while noting persistent challenges in model verification, validation and real-time control capability that must be addressed before broader industry adoption.[90]

Finance

MASFIN: A Modular Multi-Agent Large Language Model Framework Advancing Transparent, Bias-Mitigated Financial Forecasting and Portfolio Optimization

In recent developments within quantitative finance, researchers introduce MASFIN, a modular multi-agent framework designed to address long-standing limitations in traditional and AI-driven financial analytics, including survivorship bias, poor signal integration, limited reproducibility, and computational inefficiency. Built to operate in high-stakes financial environments that require transparent and reliable decision-making, MASFIN integrates large language models with heterogeneous data sources such as structured financial indicators and unstructured market news, while incorporating explicit bias-mitigation protocols to ensure methodological rigor. The framework utilizes GPT-4.1-nano for reproducible, cost-efficient inference and autonomously constructs weekly equity portfolios consisting of 15 to 30 stocks, with allocation weights optimized for short-term performance. Across an eight-week live evaluation, MASFIN achieved a 7.33% cumulative return, outperforming major market benchmarks including the S&P 500, NASDAQ-100, and Dow Jones indices in six of the eight weeks, though with increased volatility. These outcomes illustrate the emerging potential of bias-aware, generative multi-agent AI systems to enhance predictive robustness, operational transparency, and practical usability in quantitative finance.[91]

[89] http://www.ceccweb.org.cn/listshow.php?cid=68&id=3799
[90] https://www.europeanpharmaceuticalreview.com/news/270194/ai-framework-continuous-manufacturing-fda-cder-research/
[91] https://arxiv.org/pdf/2512.21878

Infosys Developments

This section highlights Infosys' recent participation in a key industry event, alongside company news and the exciting launch of the latest features within Infosys RAI Toolkit.

Events

Roundtable on AI Governance & Liability| iSPIRT | January 20, 2026 | Delhi

On 20 January 2026, iSPIRT hosted its second roundtable on techno legal regulation for AI governance as an official pre summit event ahead of the IndiaAI Impact Summit 2026, organized with Anand and Anand, the Centre on Law, Regulation, and Technology at BML Munjal University, and the Centre of Excellence in AI and Law at NALSAR University of Law, Hyderabad. The roundtable opened with a keynote by Pravin Anand, Managing Partner, Anand and Anand; was jointly moderated by Hari Subramanian and Vibhav Mithal, FHCA; and was presided over by Amit A. Shukla, Joint Secretary (Cyber Diplomacy & E Governance), Ministry of External Affairs, and Avinash Agarwal, Director General, Department of Telecommunications and Member of the Drafting Committee for the India AI Governance Guidelines. Prof. Subodh Sharma, iSPIRT volunteer, presented the DEPA Chain architecture to enable fair liability apportionment and build public trust. Representing Infosys, Ashish Tewari, Head of the Infosys Responsible AI Office (India), contributed perspectives on operationalizing accountable AI governance. The proceedings will inform the Safe and Trusted AI track of the IndiaAI Impact Summit 2026 and advance India's priority of clear, actionable AI liability regimes.

Institute of Internal Auditors (IIA) International Conference | January 8-9, 2026 | Mumbai

The IIA International Conference in Mumbai, held on January 8-9, 2026, was an Official Pre-Summit Event of the AI Impact Summit 2026 and centered around the theme "The Future of Audits: Trends, Technologies and Talent." The discussions emphasized how a strong culture of Governance, Risk, and Compliance forms the foundation for sustainable innovation. Syed Ahmed, AVP & Global Head – Infosys Responsible AI Office, delivered a keynote titled "The Next Frontier of AI Governance – From Policies to Practice," emphasizing that trust must move alongside innovation and at times faster to ensure the safe commercial adoption of AI. His session highlighted critical Responsible AI themes, including auditing trust in AI systems, approaches such as policy as living code and codified compliance, and the importance of governance and compliance at both design and runtime, especially for AI Agents. The event brought together heads of audit functions from leading organizations, audit professionals, and industry leaders, fostering dialogue on emerging trends and technologies shaping the future of the audit function. Syed's session was highly appreciated by attendees and reinforced the evolving role of auditors as key enablers and defenders of trustworthy AI, underscoring the importance of integrating Responsible AI deeply into enterprise governance frameworks.

Ethics and Responsible AI: AI for India – From Policy-to-Practice | December 18, 2025 | Virtual

On 18th December 2025, Primus Partners hosted the second webinar in its "AI for India: From Policy-to-Practice" series, focusing on Regulations, Governance and Ethics in AI. The session convened leaders from industry, academia, and government to explore how India can balance innovation with regulatory compliance and ethical considerations. Representing Infosys, Ashish Tewari, Head of Infosys Responsible AI Office (India), shared practical insights on operationalizing governance through structured toolkits and frameworks, emphasizing impact assessment, accessibility as a design principle, and rigorous checks on data integrity and output accuracy. The discussion offered actionable recommendations for shaping India's AI ecosystem responsibly and will be captured in a post-webinar compendium for stakeholders.

Infosys at India AI Impact Summit 2026

Infosys is a key contributor to the India AI Impact Summit 2026, driving national and international conversations on responsible AI. In the run-up to the AI Impact Summit, Infosys hosted and participated in 15+ official pre-summit engagements across the globe. Infosys hosted two key roundtables: Roundtable on Responsible AI: From Principles to Practice (New Delhi) and the Responsible AI Dialogue – EU–India Pathways to Trustworthy Innovation (Brussels). In addition, Infosys participated in discussions including the India–Spain Conference on AI: Exploring Opportunities and Cooperation (Spain), the UK–India AIxcelerate Sandpit and Sprint, the Third India–France AI Policy Roundtable, conclaves on AI Governance and Safe and Trusted AI, and roundtables on AI Liability and AI Standards Lifecycle Governance, reflecting Infosys' deep involvement in shaping AI policy, regulation, and best practices at scale.
In continuation of this leadership, Infosys will host multiple high-impact panel sessions at the main summit from 16 to 20 February 2026 and showcase its innovative AI solutions, accelerators, and tools at the AI Impact Expo. Connect with Infosys at Booth 3.8, Bharat Mandapam, New Delhi. Participants can register through the official event portal at https://impact.indiaai.gov.in/registration.

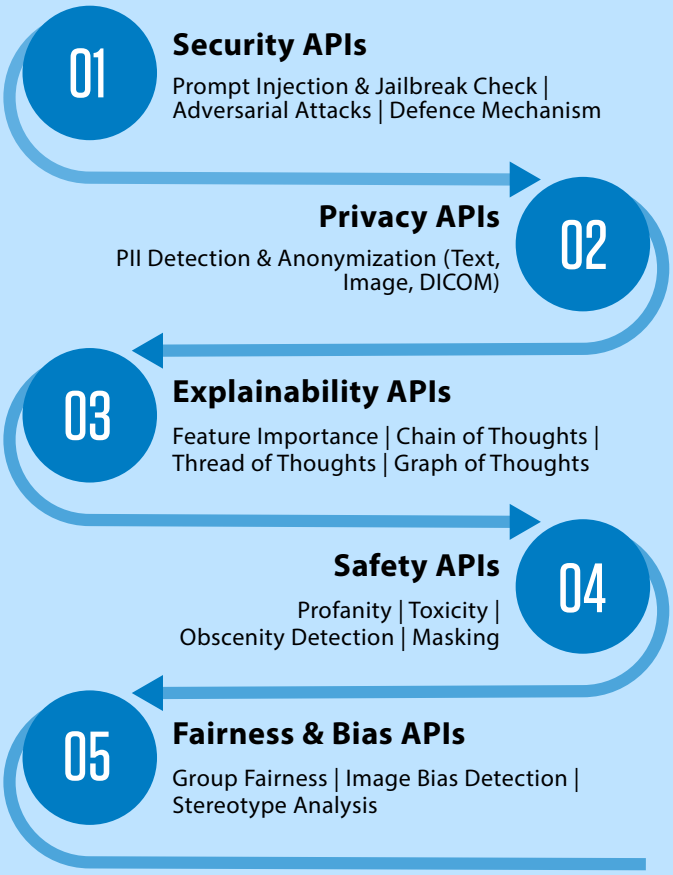
Scan the QR Code to Register for the event

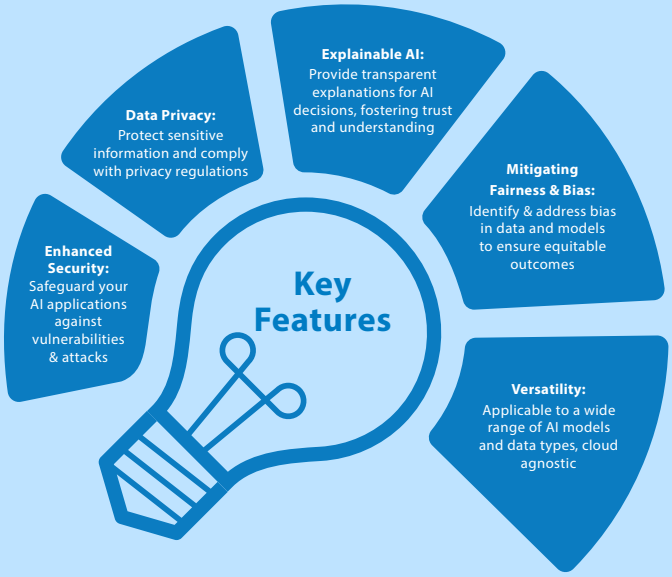
Infosys Responsible AI Toolkit - A Foundation for Ethical AI

The Open-Source Infosys Responsible AI Toolkit can be accessed from its public GitHub repo[92] also as project Salus.[93]

Overview of the Responsible AI Toolkit

Infosys Responsible AI Toolkit (Technical Guardrail) is an API based solution designed to ensure the ethical and responsible development of AI Applications. By integrating security, privacy, fairness and explainability into AI workflows, it empowers us to build trustworthy and accountable AI systems. It includes below main components:





New Features

Below new features are developed and will be available soon in our next release (version 3.0.0).

- Explainability Enhancement Using Reasoning Models
- Second order explainable AI (SOXAI) Technique for Explainability Module
- Multi-lingual support for FM-Moderation Guardrails
- Signature and face masking in Privacy module
- Bulk document safety validation
- Moderation Layer Model based guardrails are improved with finetuned smaller models to improve latency and accuracy

Infosys Responsible AI Toolkit's Model *based guardrails is now available in* ***HuggingFace,***[94] *both as a repo and also as an interactive playground to explore. Star us and be a part of the Responsible AI Revolution!* ★★★★★

[92] https://github.com/Infosys/Infosys-Responsible-AI-Toolkit
[93] https://github.com/salus-rai/salus
[94] https://huggingface.co/spaces/InfosysEnterprise/Moderation-playground

Contributors

We extend our sincere thanks to all the contributors who made this newsletter issue possible.

Srinivasan S - Policy Advocacy, Consultancy and Customer Outreach, Infosys Responsible AI Office

Mandanna A N - Head of Infosys Responsible AI Office, USA

Siva Elumalai - Senior Consultant, Infosys Responsible AI Office, India

Dakeshwar Verma - Senior Analyst - Data Science, Infosys Responsible AI Office, India

Utsav Lall - Senior Associate Consultant, Infosys Responsible AI Office, India

Pritesh Korde - Senior Associate Consultant, Infosys Responsible AI Office, India

Anie Juby - Industry Principal, Infosys Topaz Branding & Communications, Bangalore

Jossy Mathew - Senior Project Manager, Infosys Topaz Branding & Communications, Bangalore

Please reach out to responsibleai@infosys.com to know more about Responsible AI at Infosys.
We would be happy to have your feedback too.

THINK. PONDER. QUESTION.
AI CAN MAKE ANYTHING LOOK REAL

Infosys Topaz is an AI-first set of services, solutions and platforms using generative AI technologies. It amplifies the potential of humans, enterprises, and communities to create value. With 12,000+ AI assets, 150+ pre-trained AI models, 10+ AI platforms steered by AI-first specialists and data strategists, and a 'responsible by design' approach, Infosys Topaz helps enterprises accelerate growth, unlock efficiencies at scale and build connected ecosystems. Connect with us at infosystopaz@infosys.com

For more information, contact askus@infosys.com

Infosys®
Navigate your next