MARKET SCAN REPORT

DECEMBER 2025

BY INFOSYS TOPAZ RESPONSIBLE AI OFFICE

Infosys topaz

IN FOCUS

HOW TO CONCEIVE OF AND IMPLEMENT A THIRD WAY FOR AI?

By Prof. Francis Rousseaux

Infosys®
Navigate your next

Foreword

Trustworthy AI does not emerge by default; it is a leadership choice. Over the course of 2025, this reality became increasingly visible across governments, enterprises, and technology ecosystems worldwide. As AI moved from experimentation to scaled deployment, the focus shifted decisively from possibility to consequence forcing leaders to confront not only what AI can do, but how responsibly it is being built and used.

Throughout the year, several global patterns stood out. Governments accelerated AI governance through new laws, conventions, and standards, signaling a transition from principle-setting to operational oversight. Enterprises moved beyond pilots toward enterprise-wide adoption, while courts and regulators began testing the boundaries of accountability, intellectual property, and market power in the age of generative AI. At the same time, rapid advances in reasoning, multimodal systems, and real-time interaction expanded AI's capabilities, even as incidents revealed how fragile systems can be when safeguards and oversight lag behind innovation.

Across these developments, one theme remained consistent: **capability is advancing faster than institutional readiness**.

These signals reinforce two leadership imperatives I have reflected on in recent months. First, **trustworthiness must become a leadership mandate** shaping strategy, culture, and decision-making at the highest levels. Second, while **AI has not broken agile ways of working, it has changed the rules**, requiring organizations to evolve toward **Responsible Agile** practices that embed governance, accountability, and human oversight into everyday delivery.

The December signals captured in this edition bring these themes into sharp focus. As governance frameworks become more actionable and real-world deployments continue to test guardrails, responsibility must move from aspiration to execution.

As we enter the new year, I hope this market scan supports leaders in reflecting the lessons of 2025 and strengthening their ability to deploy AI systems that are not only powerful, but worthy of trust. Wishing you clarity in direction, confidence in execution, and continued commitment to responsible innovation.

Happy New Year!

Syed Ahmed
Global Head
Infosys Responsible AI Office

From the editor's desk

December Signals: What to Carry Forward

December is less about closure and more about calibration. The developments of the past month did not redefine the AI landscape, but they clarified where momentum is building, where pressure is rising, and what leaders should prepare for as we enter the new year.

This month's market signals point to a clear pattern: AI governance is becoming operational, guardrails are being stress-tested in real systems, and innovation continues now under sharper expectations of trust.

Governance Moves from Intent to Execution: December reinforced that AI governance is no longer a policy-only exercise. Multilateral and national actions continued to translate principles into enforceable frameworks. Bosnia and Herzegovina's accession to the Council of Europe's AI Convention strengthened alignment around human rights and democratic safeguards. The UK–Israel Abraham Accords AI framework highlighted how shared AI principles are beginning to support diplomatic and strategic cooperation.

Across regions, regulatory focus sharpened. The U.S. advanced legislation addressing AI civil rights and talent readiness. Vietnam passed its first AI law, Japan signaled a push toward widespread enterprise adoption, and China introduced ethical review requirements for AI patents. Canada set an important precedent with the world's first accessibility standard for AI, while NIST's draft Cyber AI Profile provided concrete guidance for securing AI systems in practice.

The December signal is unmistakable: **governance is moving closer to deployment environments, not staying at the level of intent.**

Incidents That Reveal the Gaps: Alongside progress, December also surfaced reminders of how fragile AI systems can be under real-world conditions. A widely shared experiment showed a robot discharging a BB gun after a minor prompt change—underscoring how safety assumptions can fail in unexpected ways. An AI assistant mishandled confidential negotiation data, and an autonomous coding tool erased an entire drive, raising concerns about unchecked execution.

Legal scrutiny intensified as well. The New York Times' lawsuit against Perplexity AI and regulatory probes in Europe into potential market abuse reflect unresolved tensions around data use, intellectual property, and platform power.

These are not edge cases. They point to systemic gaps between capability, control, and accountability—gaps that leaders must address proactively rather than reactively.

Innovation Continues With Clear Signals: Despite tighter scrutiny, innovation in December remained strong. New reasoning models from DeepSeek, real-time speech capabilities such as Microsoft's VibeVoice-Realtime, and multimodal advances from Jina AI and Zhipu AI point to increasingly capable systems. At the same time, developments in privacy-preserving techniques such as ConceptGuard, Ensemble Privacy Defense, and probabilistic safeguards signal that safety research is evolving alongside capability.

The message here is encouraging: **responsible innovation is beginning to scale with technical progress**.

Looking Ahead

As we step into the new year, December leaves us with a practical challenge—and an opportunity. Guardrails are rising, but readiness remains uneven across sectors and regions. Trust will increasingly differentiate those who can deploy AI at scale from those who cannot.

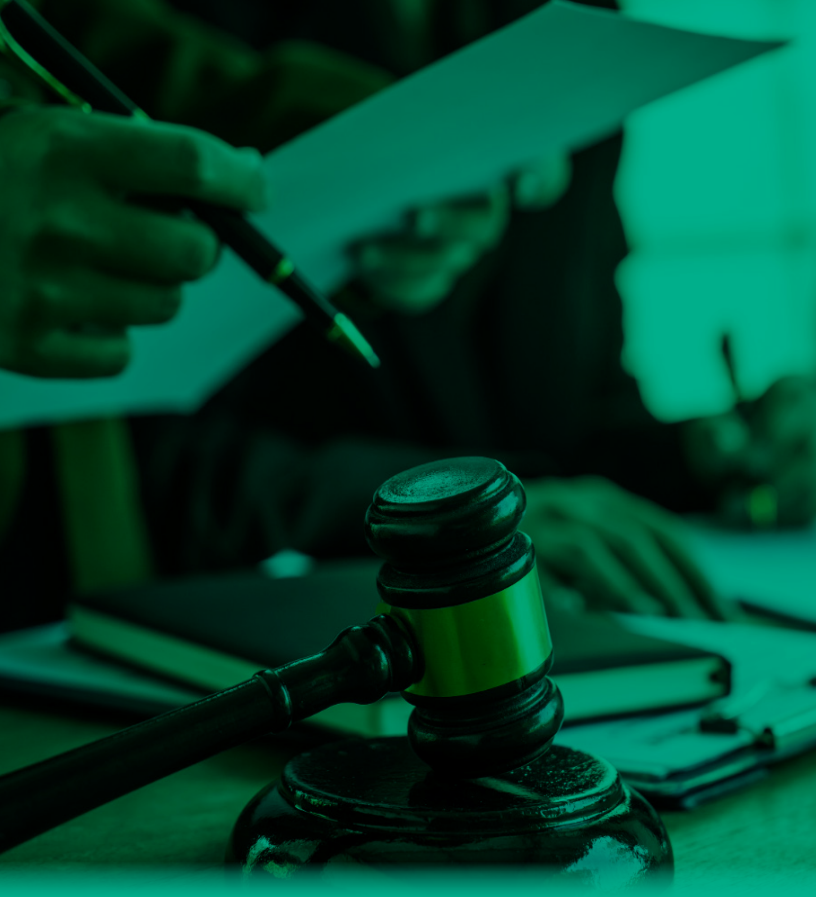
For leaders, the imperative is clear: move beyond compliance, embed governance into system design, and treat trust as a strategic asset. December does not close the conversation, it sharpens it. And how we respond in the months ahead will define not just adoption, but credibility.

Wishing you a new year defined by thoughtful innovation, stronger governance, and AI that delivers impact with trust.

Warm regards,

Ashish Tewari
Head- Infosys Responsible AI Office, India

Table of Contents



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

UK and Israel Launch Regional AI Cooperation Framework with Abraham Accords Partners

Israel and the United Kingdom have formalized a regional framework to strengthen collaboration in artificial intelligence among Abraham Accords nations through a Memorandum of Understanding signed in Tel Aviv by the Holon Institute of Technology (HIT) and the UK Abraham Accords Group (UKAAG). The initiative aims to leverage AI as a tool for stability and progress across sectors such as health, energy, education, and governance. Key objectives include creating joint training programs, establishing regional benchmarks for ethical AI practices, and forming working groups to oversee development projects. This agreement reflects a growing emphasis on technology-driven diplomacy to foster cooperation and innovation among countries that have normalized relations with Israel.[1]

AI Risks Creating a New Era of Global Divergence as Development Gaps Widen, UNDP Report Warns

The United Nations Development Programme (UNDP) warns that unmanaged artificial intelligence (AI) could reverse decades of progress in reducing global inequality, creating a new era of divergence between countries. In its report The Next Great Divergence: Why AI May Widen Inequality Between Countries, UNDP highlights that while AI offers significant opportunities for growth such as boosting Asia-Pacific's GDP by up to $1 trillion countries start from vastly different levels of digital readiness. Advanced economies are rapidly scaling AI infrastructure and innovation, while others face constraints in skills, governance, and access, amplifying risks like job displacement and algorithmic bias. Without inclusive policies and robust regulations, capability will become the defining fault line of the AI era.[2]

UK–South Africa Partnership Launches Global AI Policy Training to Strengthen Governance Capacity

At the Science Forum in Pretoria, the British High Commission and South Africa's Department of Science, Technology and Innovation (DSTI) announced the UK–South Africa Global AI Policy Training Programme. This initiative will equip 30 policymakers and science leaders15 from each country with advanced skills in AI governance and policy design. Drawing on expertise from leading institutions such as the University of Cape Town, University of Cambridge, and the Global Centre on AI Governance, the programme aims to build capacity for responsible AI adoption and regulation. The effort builds on commitments made during the UK-SA Joint Committee Meeting in London and reflects both nations' shared vision to leverage AI for inclusive socio-economic development and global leadership in ethical AI governance.[3]

Iran and Russia Forge Strategic Partnership to Advance AI and Cybersecurity Collaboration

Iran and Russia have formalized a strategic agreement aimed at strengthening cooperation in the fields of artificial intelligence, cybersecurity, and digital innovation. Signed in Moscow by Iran's Deputy Minister of Information and Communications Technology, Meysam Abedi, and Russia's Deputy Minister of Digital Development, Communications and Mass Media, Alexander Shoitov, the memorandum of understanding emerged from the fifth meeting of their joint working group. The partnership encompasses a broad spectrum of initiatives, including smart governance, blockchain, fintech, regulatory frameworks, and postal services. Both nations intend to leverage shared research and practical projects to enhance data transit, electronic governance, and the development of AI-driven tools, ultimately delivering higher-quality products and services across their digital ecosystems.[4]

[1] https://aicorenews.com/industry-news/uk-israel-ai-cooperation-framework-abraham-accords/
[2] https://www.undp.org/asia-pacific/press-releases/ai-risks-sparking-new-era-divergence-development-gaps-between-countries-widen-undp-report-finds
[3] https://www.gov.uk/government/news/uk-sa-partnership-boosts-health-inclusion-and-space-science
[4] https://thelegalwire.ai/iran-russia-sign-agreement-to-boost-ai-cyber-security-cooperation/

CISA and Global Partners Issue Guidance for Secure AI Integration in Operational Technology

The U.S. Cybersecurity and Infrastructure Security Agency (CISA), in collaboration with international partners including Australia, has released joint guidance to help critical infrastructure owners and operators securely integrate artificial intelligence into operational technology (OT) systems. The guidance emphasizes establishing robust AI governance frameworks, conducting continuous model testing, and embedding security measures throughout the AI lifecycle to ensure that efficiency gains do not compromise safety and reliability. By addressing unique risks posed by AI in OT environments, the recommendations aim to balance innovation with resilience, safeguarding critical infrastructure against emerging cyber threats.[5]

Global AI Safety Coalition Rebrands as International Network for Advanced AI Measurement Evaluation and Science

The International Network of AI Safety Institutes, which includes countries such as Australia, Canada, the EU, France, Japan, Kenya, South Korea, Singapore, the UK, and the US, has announced a rebrand to the International Network for Advanced AI Measurement Evaluation and Science. This change reflects a strategic shift away from the "safety" label toward a stronger emphasis on scientific measurement and evaluation of AI models, following UK-led calls for this transition. The UK will assume the role of network coordinator, spearheading global efforts to develop rigorous, standardized methods for testing advanced AI systems and ensuring that evaluation practices keep pace with rapid technological advancements.[6]

G7 Ministers Advance Global Cooperation on AI, Quantum Technologies, and Digital Policy Under Canada's Leadership

The G7 Industry, Digital and Technology Ministers' Meeting in Montréal, held under Canada's presidency, brought together ministers from G7 nations and partners to strengthen collaboration on AI, quantum technologies, digital policy, and industrial resilience. Key outcomes included the G7 2025 Industry, Digital and Technology Ministerial Declaration, committing to inclusive economic growth through AI, secure and competitive digital economies, and resilient supply chains. Another major deliverable was the SME AI Adoption Blueprint and accompanying Ministerial Statement, providing practical guidance and policy tools to help small and medium-sized enterprises adopt AI responsibly while safeguarding competition, data protection, and intellectual property. Canada also announced significant agreements: two MOUs with the EU on AI and digital sovereignty, an MOU with the UK on digital public infrastructure, and the launch of the Canada–Germany Digital Alliance to advance AI, quantum innovation, and talent mobility.[7]

Bosnia and Herzegovina Signs Historic AI Framework Convention to Uphold Human Rights and Rule of Law

Bosnia and Herzegovina has signed the Council of Europe's Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, becoming the 18th signatory to this landmark treaty. The convention is the first legally binding international agreement designed to ensure that AI systems operate in full compliance with human rights, democratic principles, and the rule of law. It establishes global standards to minimize risks posed by AI while promoting responsible innovation in the digital environment. Complemented by sector-specific initiatives, this framework underscores the Council of Europe's commitment to safeguarding fundamental rights and democratic values as AI technologies continue to evolve and impact societies worldwide.[8]

[5] https://www.cisa.gov/news-events/alerts/2025/12/03/cisa-australia-and-partners-author-joint-guidance-securely-integrating-artificial-intelligence

[6] https://www.uktech.news/ai/ai-safety-institutes-rebrand-to-focus-on-scientific-measurements-20251209

[7] https://www.canada.ca/en/innovation-science-economic-development/news/2025/12/canada-advances-international-cooperation-on-industry-ai-and-digital-technology-at-the-g7-industry-digital-and-technology-ministers-meeting.html

[8] https://www.coe.int/en/web/artificial-intelligence/home/-/asset_publisher/zndsEJ9HsDB3/content/bosnia-and-herzegovina-signs-the-council-of-europe-s-framework-convention-on-artificial-intelligence

US

USPTO Issues Revised Inventorship Guidance Affirming Human Inventorship for AI-Assisted Innovations

The United States Patent and Trademark Office (USPTO) released updated inventorship guidance for AI-assisted inventions, fully rescinding its February 2024 directive. The revised guidance reiterates that the existing legal standard for determining inventorship applies equally to all inventions, regardless of AI involvement, and does not establish any new or modified criteria. Under U.S. patent law, only natural persons can be named as inventors, as defined in 35 U.S.C. § 100(f), with invention tied to human conception. AI systems, including generative models and computational tools, are considered aids in the inventive process and cannot attain inventor status. This update aligns with Executive Order 14179, which directs agencies to revise prior policies to foster American leadership in artificial intelligence.[9]

Tennessee's 2025 AI Action Plan: A Roadmap for Responsible Innovation

Tennessee released its first AI Advisory Council Action Plan, outlining a bold vision to responsibly integrate artificial intelligence across government, education, and industry. The plan focuses on four strategic pillars: launching high-impact AI pilot projects; building secure data infrastructure; enhancing workforce development through reskilling and AI literacy; and strengthening governance, transparency, and privacy protections. It emphasizes Tennessee's values integrity, transparency, and fiscal responsibility to ensure innovation doesn't compromise public trust. The Council commits to annual progress reports through 2028 and aims to position Tennessee as a leader in ethical, practical AI adoption that improves government efficiency and economic opportunity for all residents.[10]

U.S. Lawmakers Reintroduce AI Civil Rights Act to Combat Algorithmic Bias and Protect Civil Liberties

In the United States, Senators have reintroduced the Artificial Intelligence Civil Rights Act, a comprehensive bill aimed at preventing discrimination and bias in AI-powered decision-making systems. The legislation seeks to establish strict guardrails on the use of algorithms in critical areas such as housing, employment, and healthcare, requiring transparency and rigorous testing before and after deployment. Co-sponsored by a broad coalition of U.S. lawmakers, the bill

[9] https://www.uspto.gov/subscription-center/2025/revised-inventorship-guidance-ai-assisted-inventions

[10] https://www.tn.gov/content/dam/tn/finance/aicouncil/documents/TN%20AI%20Advisory%20Council%20Action%20Plan%20-%20November%202025.pdf

addresses growing concerns that AI technologies, often trained on historically biased data, could perpetuate systemic inequities and harm marginalized communities. Advocates emphasize that the Act represents a pivotal step toward ensuring fairness, accountability, and civil rights protections in an era where AI increasingly influences Americans' lives and livelihoods.[11]

Illinois Legislature Amends Human Rights Act to Regulate Employer Use of Artificial Intelligence in Hiring, Promotion, and Other Employment Decisions Effective January 1, 2026

The Illinois Human Rights Act has been amended under Public Act 103-0804 to impose new regulatory requirements on employer use of artificial intelligence (AI) in employment decisions, adding definitions for "Artificial Intelligence" and "Generative Artificial Intelligence" and making it a civil rights violation for employers to use AI in recruitment, hiring, promotion, tenure, discharge, discipline, training selections, or other terms and conditions of employment if such use has the potential to discriminate against individuals in protected classes or employ zip codes as proxies for those classes, and further requiring employers to provide notice to employees when AI is used in any of these employment-related decisions, with enforcement authority vested in the Illinois Department of Human Rights, prompting organizations to assess and adjust their AI-driven processes and compliance practices ahead of the amendments' January 1 2026 effective date.[12]

U.S. Executive Order Calls for National AI Policy Framework to Replace Fragmented State Regulations

The President of the United States has issued an executive order emphasizing the urgent need for a unified national policy framework on artificial intelligence to safeguard American leadership in the field. Building on Executive Order 14179, which removed barriers to AI innovation earlier in 2025, the directive warns that state-by-state regulations create a patchwork of conflicting rules that hinder compliance and innovation, particularly for startups. The order criticizes state laws that mandate ideological bias in AI models, citing Colorado's algorithmic discrimination law as an example, and argues that such measures could distort outputs and impede interstate commerce. The administration calls for a minimally burdensome national standard that overrides conflicting state laws while ensuring protections for children, preventing censorship, respecting copyrights, and safeguarding communities. Framing AI as a critical domain for national and economic security, the order underscores that winning the global AI race requires regulatory clarity and freedom for U.S. companies to innovate without excessive constraints.[13]

[11] https://jayapal.house.gov/2025/12/02/jayapal-markey-clarke-lee-reintroduce-ai-civil-rights-act-to-eliminate-ai-discrimination-and-enact-guardrails-on-use-of-algorithms-in-decisions-impacting-peoples-rights-civil-li/

[12] https://www.grsm.com/insight/amendments-to-illinois-human-rights-act-to-regulate-use-of-ai-in-employment-decisions/

[13] https://www.whitehouse.gov/presidential-actions/2025/12/eliminating-state-law-obstruction-of-national-artificial-intelligence-policy/

NDAA Proposes Creation of DOD Steering Committee to Address Advanced AI and AGI Risks

The draft text of the Fiscal Year 2026 National Defense Authorization Act (NDAA) includes a provision requiring the Secretary of Defense to establish an "Artificial Intelligence Futures Steering Committee" by April 2026. This high-level committee will be co-chaired by the Deputy Secretary of Defense and the Vice Chairman of the Joint Chiefs of Staff, with membership comprising service vice chiefs and other senior officials. Its mandate is to develop policies for evaluating, adopting, and governing advanced AI systems, including those approaching artificial general intelligence (AGI), while mitigating associated risks. The committee will also analyze emerging AI models and technologies that could lead to AGI, assess the operational implications of integrating such systems into Department of Defense platforms, and examine adversarial advancements to formulate counter-strategies. This initiative underscores the growing importance of AI governance in national security planning.[14]

AI Talent Act Introduced to Modernize Federal Hiring and Attract Top AI Expertise

A bipartisan and bicameral group of U.S. legislators has introduced the AI Talent Act to strengthen federal agencies' ability to recruit and retain top artificial intelligence and technical talent. Led by Rep. Sara Jacobs with co-leads Rep. Shontel Brown, Rep. Jay Obernolte, and Rep. Pat Fallon, alongside Senators Andy Kim and Jon Husted, the bill seeks to overhaul outdated hiring practices by creating specialized AI talent teams, enabling skills-based hiring, and providing shared tools to accelerate candidate onboarding. Key provisions include establishing agency-level AI talent teams, authorizing the Office of Personnel Management (OPM) to lead pooled hiring and develop shared resources, implementing subject-matter expert-driven technical assessments, creating a shared online platform for validated assessments, phasing out automated self-assessments, and allowing additional talent teams for high-need areas such as cybersecurity and data science. The legislation aims to ensure the U.S. remains competitive in the global AI race by building a skilled federal workforce capable of deploying AI responsibly and effectively.[15]

OPM Launches United States Tech Force to Advance Federal Technology Modernization and AI Leadership

The US Office of Personnel Management (OPM), in collaboration with the Office of Management and Budget (OMB), the General Services Administration (GSA), the White House Office of Science and Technology Policy (OSTP), and agency leaders across the administration, announced the creation of the United States Tech Force across-government initiative designed to recruit top technologists to modernize federal systems and accelerate AI adoption. This program aligns with President Trump's directive to secure America's leadership in artificial intelligence, emphasizing both private-sector innovation and the need for technical expertise within government. Tech Force will deploy elite teams of engineers, data scientists, and technology leaders to tackle mission-critical projects across federal agencies, supported by partnerships with leading technology companies and world-class training programs. By mobilizing talent from industry and academia, the initiative aims to strengthen national competitiveness, modernize infrastructure, and position the United States at the forefront of global technological leadership.[16]

[14] https://armedservices.house.gov/uploadedfiles/rcp_text_of_house_amendment_to_s._1071.pdf
[15] https://shontelbrown.house.gov/media/press-releases/rep-brown-leads-bipartisan-and-bicameral-bill-reps-jacobs-obernolte-fallon-and
[16] https://www.opm.gov/news/news-releases/opm-launches-us-tech-force-to-implement-president-trumps-vision-for-technology-leadership/

UK

AI for Science Strategy Expert Group: Advisory Role and Governance Framework

The AI for Science Strategy Expert Group has been established in UK to provide independent strategic advice to the UK Department of Science, Innovation and Technology (DSIT) on developing the AI for Science Strategy, which aims to leverage artificial intelligence to accelerate scientific discovery and maintain the UK's global leadership in research and innovation. Operating from August through autumn 2025, the group comprises distinguished experts from academia and industry, including leaders from the Royal Society, Oxford University, Microsoft Research, Google DeepMind, and the Engineering and Physical Sciences Research Council. Members will contribute by identifying priority areas, advising on policy interventions, and proposing partnership models to foster collaboration across sectors. Participation involves monthly meetings, document reviews, and flexible engagement, with all roles being voluntary and non-remunerated. While the group offers informal, advisory input, final decisions rest with ministers, ensuring alignment with government priorities. Transparency measures are in place to protect commercial integrity, and all members have undergone necessary checks to serve in this capacity.[17]

UK Home Office Opens Consultation on Draft Framework for Law Enforcement Use of Biometrics and Facial Recognition

The UK Home Office has released a draft statutory framework and launched a public consultation on the use of biometric, facial recognition, and similar technologies by law enforcement agencies in England and Wales, as well as for reserved matters across the UK. The proposed framework outlines the scope of technologies covered, including live, retrospective, and operator-initiated facial recognition, and addresses critical issues such as acquisition, retention, and use of facial images, proportionality assessments, and potential impacts on privacy, freedom of expression, and assembly. It also considers safeguards for searching other government databases and requirements for authorization. A key proposal is the creation of a single independent oversight body with powers to set standards, issue statutory codes, conduct investigations, and enforce compliance. The consultation invites responses to the Data and Identity Directorate, with a government response expected within 12 weeks, aiming to establish clear, consistent rules that balance public safety with individual rights.[18]

[17] https://www.gov.uk/government/publications/ai-for-science-strategy-expert-group-terms-of-reference/ai-for-science-strategy-expert-group-terms-of-reference
[18] https://assets.publishing.service.gov.uk/media/69318bb2cdec734f4dff4257/PDF_Consultation_FINAL.pdf

Europe

European Commission Opens Stakeholder Consultation on Rights Reservation Protocols for Text and Data Mining Under the AI Act and GPAI Code of Practice

The European Commission has launched a stakeholder consultation aimed at supporting the implementation of the AI Act's requirement for providers of general-purpose AI (GPAI) models to identify and comply with rights reservations expressed by rightsholders. This consultation focuses on protocols for reserving rights from text and data mining (TDM) and builds on commitments outlined in the Copyright section of the GPAI Code of Practice. Approved by the Commission as adequate for AI Act compliance, the Code obliges signatories to respect machine-readable opt-out protocols, including robots.txt and future IETF standards. To facilitate consensus on state-of-the-art, technically feasible, and widely adopted opt-out solutions, the Commission supported by the EU Intellectual Property Office (EUIPO)invites rightsholders, GPAI providers, civil society, and standardization bodies to provide input on technical feasibility and adoption of TDM opt-out mechanisms identified in the EUIPO study on generative AI and copyright. Stakeholders may also express interest in follow-up workshops to agree on protocols that GPAI providers must respect under the AI Act. An online information session will be held to outline the process and present technical solutions. The resulting list of agreed machine-readable opt-out protocols will be published and reviewed biennially to ensure alignment with updates to the Code of Practice.[19]

Commission Releases First Draft of Code of Practice for Marking and Labelling AI-Generated Content

The European Commission has published the first draft of its Code of Practice on marking and labelling AI-generated content, aligning with its timeline to finalize the code by June 2026. Developed by independent experts, this voluntary framework aims to help providers and deployers comply with Article 50 of the AI Act, which mandates machine-readable marking of AI-generated or manipulated content and clear labelling of deepfakes and AI-generated text on matters of public interest. The draft comprises two sections: one outlining rules for marking and detecting AI content for generative AI providers, and another detailing labelling requirements for deployers of such systems. Feedback on the draft will be accepted until January 23, with a second version expected by mid-March 2026. The transparency obligations under the AI Act will become enforceable on August 2, 2026.[20]

[19] https://digital-strategy.ec.europa.eu/en/consultations/commission-launches-consultation-protocols-reserving-rights-text-and-data-mining-under-ai-act-and

[20] https://digital-strategy.ec.europa.eu/en/news/commission-publishes-first-draft-code-practice-marking-and-labelling-ai-generated-content

Spain

Spain Urges EU to Create Specific Copyright Protection Standards Against AI Misuse

Spain's Minister of Culture, Ernest Urtasun, has called on European Union partners to develop a new legislative instrument to safeguard creators' copyrights in the context of artificial intelligence. Speaking at a meeting of EU culture ministers in Brussels, Urtasun argued that the existing Code of Good Practice is insufficient to address the widespread use of protected works by AI models without consent or fair compensation. He emphasized the need for stronger regulations that ensure transparency, author consent, and adequate remuneration when creators' works are used to train AI systems. Urtasun warned that current practices leave European creators defenseless, as massive amounts of copyrighted content are being exploited without proper licensing, and urged the EU to strengthen its regulatory framework to prevent further erosion of copyright protections.[21]

Brazil

Brazil Proposes Comprehensive AI Governance Framework Through New Legislative Bill

Brazil's Chamber of Deputies has introduced Bill PL 6237/2025, which seeks to establish the National System for the Development, Regulation, and Governance of Artificial Intelligence (SIA). The proposed framework includes the creation of the Brazilian Council for Artificial Intelligence (CBIA), the National Data Protection Authority (ANPD), and two expert committees representing civil society, alongside other public bodies. Under the bill, CBIA would define AI principles and guidelines related to economic growth, digital sovereignty, and social inclusion, while advising the President on AI policy. ANPD would serve as the residual regulator with authority to issue binding rules on incident reporting, algorithmic impact assessments, and high-risk AI classifications. The law also requires defining risk categories based on factors such as fundamental rights, vulnerable groups, cybersecurity threats, and impacts on children. This governance model aims to coordinate oversight among regulators and position Brazil for responsible AI development as Congress evaluates competing proposals.[22]

[21] https://europeannewsroom.com/es/espana-reclama-una-norma-europea-especifica-para-proteger-los-derechos-de-autor-de-los-creadores-frente-a-la-ia/

[22] https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=3063060&filename=PL%206237%2F2025

Canada

Canada Launches Its First Public AI Register to Track AI Usage Across Federal Government

The Canadian government, through the Treasury Board Secretariat, has launched its first publicly accessible **AI Register**, which details where and how artificial intelligence is used within federal institutions. The register includes entries from 42 departments, covering over 400 AI systems in different stages from research and proof-of-concept to full deployment. Each system listing provides transparent information such as its purpose, whether it was built in-house or by a vendor, and its current or planned usage. The initiative aligns with Canada's broader AI Strategy, and the government plans to consult the public in 2026 to refine the register's design and usability.[23]

Australia

Being Clear About AI-Generated Content: A Guide for Australian Businesses

The National Artificial Intelligence Centre has published guidance to help Australian businesses adopt best practices for transparency when using AI-generated or AI-modified content. This voluntary framework emphasizes the importance of informing users whenever artificial intelligence is involved in creating or altering text, images, audio, or video, and recommends mechanisms such as labelling, watermarking, and metadata recording to ensure clarity. Designed for all businesses engaged in generating, designing, or adapting AI systems, the guidance aims to build trust, align with emerging international standards, and support ethical AI adoption. While it does not address copyright issues or detection tools, it encourages organizations to make AI-generated content clearly identifiable, thereby contributing to a fair and transparent digital ecosystem.[24]

Australia's National AI Plan – A Strategic Roadmap for Growth, Inclusion, and Safety

The Australian Government has released the National AI Plan to accelerate AI adoption and build a competitive, resilient economy. The plan sets three overarching goals: "Capture the opportunity" by investing in AI-ready infrastructure, supporting domestic research and innovation, and attracting global investment; "Spread the benefits" by scaling AI adoption across

[23] https://www.canada.ca/en/treasury-board-secretariat/news/2025/11/canada-launches-first-register-of-ai-uses-in-federal-government.html
[24] https://www.industry.gov.au/publications/being-clear-about-ai-generated-content

businesses and government services, and equipping workers with skills to use and shape AI; and "Keep Australians safe" by strengthening existing laws, creating institutions like the AI Safety Institute, and promoting responsible, risk-based practices and international cooperation. This roadmap ensures AI benefits reach all Australians while safeguarding fairness, accountability, and public trust.[25]

OAIC Issues Guidance on Responsible Use of Generative AI Tools in the Workplace

The Office of the Australian Information Commissioner (OAIC) has published guidance on how businesses can balance the benefits of generative AI tools with the protection of personal information in workplace settings. The blog highlights privacy risks beyond simple disclosure to third-party providers, including secondary uses, new data collections, and concerns about security and accuracy. OAIC advises organizations to review and update privacy policies, collection notices, and communications with customers to reflect AI usage. Key recommendations include conducting Privacy Impact Assessments, prohibiting the upload of personal information to publicly available AI tools where risks are high, implementing robust internal policies and staff training, managing privacy settings on enterprise licenses, and restricting data retention and access for AI model training. The guidance underscores the need for strong governance and compliance measures to mitigate risks while leveraging AI for business efficiency.[26]

Australia Drops Plan for Permanent AI Advisory Body Amid Rising Oversight Concerns

Australia will not proceed with the establishment of a permanent AI Advisory Body as previously promised, despite increasing calls for stronger governance of artificial intelligence. The Department of Industry, Science and Resources (DISR) confirmed in recent responses to Senate Estimates that the initiative, which was proposed and funded in the 2024 federal Budget, has been abandoned. Originally intended to operate within DISR, the advisory body was expected to provide expert guidance on AI-related opportunities and risks, drawing on insights from civil society, industry, and academia. This decision marks a significant shift from earlier commitments outlined in a 2024 Senate inquiry submission, raising questions about the government's approach to AI oversight at a time when global regulatory efforts are accelerating.[27]

India

DPIIT Releases Working Paper on Generative AI and Copyright, Proposes Hybrid Licensing Model

The Department for Promotion of Industry and Internal Trade (DPIIT) has published Part 1 of its working paper examining the intersection of generative artificial intelligence (AI) and copyright law, based on recommendations from an eight-member committee formed in April 2025. The paper reviews global approaches such as blanket exemptions, text and data-mining exceptions, voluntary licensing, and extended collective licensing, highlighting their limitations. Rejecting zero-price licensing due to its potential to undermine human creativity, the committee proposes a hybrid model granting AI developers a blanket license to use lawfully accessed content for training, with royalties payable only upon commercialization of AI tools. Rates would be determined by a government-appointed committee and subject to judicial review, while a centralized mechanism would manage royalty collection and distribution to reduce transaction costs and ensure equitable access. DPIIT has invited public feedback on the proposed framework within 30 days, signaling India's effort to balance innovation with copyright protection.[28]

[25] https://www.industry.gov.au/publications/national-ai-plan
[26] https://www.oaic.gov.au/news/blog/GenAI-tools-in-the-workplace-balancing-protection-of-personal-information-and-business-efficiency
[27] https://thelegalwire.ai/government-abandons-plan-for-external-ai-advisory-body/
[28] https://www.pib.gov.in/PressReleasePage.aspx?PRID=2200741&reg=3&lang=1

Private Member's Bill Seeks Comprehensive Ethics and Accountability Framework for AI in India

The Artificial Intelligence (Ethics and Accountability) Bill, 2025, introduced as a Private Member's Bill in the Lok Sabha, proposes the creation of a statutory framework to govern the ethical use of AI in decision-making, surveillance, and algorithmic systems across India. The Bill emphasizes fairness, transparency, and accountability, mandating the establishment of an Ethics Committee for AI to develop guidelines, monitor compliance, and address misuse or bias. It imposes strict limitations on AI deployment in sensitive areas such as law enforcement, finance, and employment, requiring lawful purpose, prior oversight, and non-discrimination. Developers are assigned clear obligations, including transparency in data and methodologies, prevention of algorithmic bias, and maintenance of compliance records, supported by a grievance redressal mechanism for affected individuals. The Bill prescribes significant penalties for violations, including monetary fines, license suspension, and criminal liability for repeat offenders, while ensuring financial support for the Ethics Committee from the Consolidated Fund of India and annual reporting to Parliament. It also grants delegated rule-making powers to the Central Government to operationalize the framework.[29]

China

China to Introduce Ethical Review in AI Patent Examination to Ensure Responsible Innovation

China's top intellectual property regulator, the China National Intellectual Property Administration (CNIPA), announced plans to strengthen ethical reviews in the patent examination process for artificial intelligence technologies. Starting January 1, 2026, newly revised patent guidelines will include a dedicated section on AI and big data, requiring technical solutions such as data collection and rule-setting to comply with legal standards, social ethics, and public interests. The guidelines also outline clearer requirements for describing AI-related processes like model construction and training, addressing challenges posed by the "black-box" nature of AI systems. Officials emphasized that this move aims to promote responsible innovation while safeguarding transparency and adequate disclosure. Additionally, China reaffirmed its commitment to equal protection of intellectual property rights for domestic and foreign enterprises and plans to deepen international cooperation under its 15th Five-Year IP development plan.[30]

China and ASEAN Advance Digital Governance Through AI Security and Cross-Border Data Cooperation

China and Association of Southeast Asian Nations(ASEAN) have reaffirmed their commitment to deepening collaboration in digital governance by focusing on three strategic areas: cybersecurity, artificial intelligence governance, and cross-border data flows. The initiative includes establishing robust emergency response mechanisms and threat information-sharing

[29] https://www.medianama.com/2025/12/223-private-members-ai-ethics-bill-algorithmic-bias-copyright-data-ownership/
[30] https://en.people.cn/n3/2025/1129/c90000-20396591.html

platforms to enhance cybersecurity resilience; implementing the Global AI Governance Initiative to promote responsible AI practices and foster exchanges among AI companies, research institutions, and industry bodies; and improving transparency and compatibility of regulatory frameworks to facilitate secure and efficient cross-border data transfers. These measures aim to strengthen regional solidarity, build consensus, and elevate digital governance standards, contributing to the creation of a closer China-ASEAN community with a shared future.[31]

Japan

Japan Seeks Public Input to Redefine Intellectual Property Strategy for the AI Era and Global Standards

The Cabinet Office of Japan's Intellectual Property Strategy Headquarters has opened a public consultation to shape the upcoming "Intellectual Property Strategic Program 2026," aiming to modernize the nation's IP framework for an economy increasingly driven by digital technology and data. This initiative focuses on addressing challenges posed by generative AI, copyright issues, and patent inventorship, while also promoting stronger global standardization. By inviting feedback until 7 January 2026, the government seeks to maximize intellectual capital and foster innovation through updated policies that reflect the realities of the AI era and international governance trends.[32]

Japan Plans to Relax Data Consent Rules for AI Development While Imposing Stricter Penalties for Misuse

Japan is preparing to amend its Personal Information Protection Law to accelerate artificial intelligence development by easing consent requirements for the use of personal data, while introducing tougher penalties for intentional violations. The proposed changes would allow third parties to access sensitive information such as medical records and criminal history without individual consent when used exclusively for statistical purposes or academic research, provided the data is anonymized. To safeguard privacy, the draft bill includes administrative fines for businesses that unlawfully acquire and sell personal information of more than 1,000 individuals, with penalties equivalent to the illicit profits gained. Accidental data leaks or inadequate security management would not incur these penalties. Additionally, the revision mandates parental consent for collecting data from individuals under 16. This dual approach reflects a balance between industry demands for innovation and public concerns over privacy and human rights, as Japan seeks to submit the amendment to the Diet promptly.[33]

[31] https://www.cac.gov.cn/2025-12/04/c_1766577618173846.htm
[32] https://thelegalwire.ai/japan-seeks-public-input-as-it-retools-ip-strategy-for-ai-era-global-standards-push/
[33] https://www.asahi.com/ajw/articles/16203430?utm_source

Japan Unveils Draft AI Strategy to Boost Public Adoption to 80% and Build a National AI Ecosystem

Japan has introduced a draft of its basic program on artificial intelligence development and utilization, setting an ambitious goal to raise public AI usage from the current 26.7% to 50% initially and ultimately to 80%. The plan positions AI as a core social infrastructurean "intellectual foundation and execution platform" and outlines the creation of an integrated AI ecosystem connecting developers, users, semiconductor manufacturers, and cloud providers to strengthen competitiveness and reduce the digital trade deficit. It calls for AI adoption across all ministries, agencies, and government employees, while leveraging approximately ¥1 trillion in private-sector investment to secure skilled talent, expand into emerging markets, and enhance research and development infrastructure. Framed as part of a national strategy to make Japan "the easiest country in the world to develop and utilize AI," the program underscores the country's commitment to accelerating innovation and global leadership in AI.[34]

Indonesia

OJK Refines AI Ethics Code to Strengthen Financial Tech Governance in Indonesia

The Financial Services Authority (OJK) of Indonesia has updated its AI ethics code to address emerging risks in the financial technology sector, including those posed by generative AI. Announced during the OECD Asia Forum on Digital Finance 2025 in Bali, the revised guidelines reinforce principles of consumer protection, data reliability, financial inclusion, cybersecurity, and fairness. They complement six core principles beneficial, accountable, transparent, explainable, resilient, and secure while introducing measures to mitigate risks such as data leaks, algorithmic bias, and AI hallucinations. Supported by OECD input, the update aims to ensure responsible AI adoption that enhances efficiency and fraud detection without compromising trust or safety.[35]

Indonesia Prepares Two Priority AI Presidential Decrees on Ethics and Roadmap for Early 2026

Indonesia's Minister of Communication and Digital, Meutya Hafid, announced that the government is finalizing two Presidential Regulations focused on artificial intelligence: the AI Ethics Framework and the AI Roadmap. These regulations, which are 90% complete, have been prioritized for signing by President Prabowo in early 2026. The AI policy will serve as an overarching framework, while detailed sector-specific rules will be delegated to individual ministries, ensuring tailored governance across industries. The roadmap aims to guide AI development broadly, while the ethics framework will establish principles for responsible use. Meutya emphasized that sector leaders are best positioned to define specific AI regulations, reinforcing a collaborative approach to digital governance under Indonesia's national strategy for technological advancement.[36]

[34] https://www.japantimes.co.jp/news/2025/12/07/japan/politics/japan-public-ai-use-strategy
[35] https://en.antaranews.com/news/394477/ojk-refines-ai-ethics-code-to-mitigate-financial-tech-risks
[36] https://kumparan.com/kumparannews/menkomdigi-dua-perpres-ai-prioritas-diteken-prabowo-awal-2026-26PJpGST9h0/full

Vietnam

Vietnam Enacts First-Ever Artificial Intelligence Law to Regulate High-Risk Systems and Promote Innovation

Vietnam's National Assembly has passed its first-ever Law on Artificial Intelligence, creating a comprehensive framework to balance regulation and innovation. The law introduces strict safeguards for high-risk AI systems, drawing on models from the EU and South Korea, while incorporating mechanisms to foster development similar to Japan. Oversight will be centralized under the government, with the Ministry of Science and Technology as the lead coordinator. Conformity assessments will apply exclusively to high-risk AI systems listed by the Prime Minister, who is authorized to update the list in real time without legislative amendments. The law avoids rigid classifications to prevent obsolescence and includes measures such as voucher schemes to support startups. It will take effect on March 1, 2026.[37]

Philippines

Philippines Completes UNESCO AI Readiness Assessment to Strengthen Ethical AI Governance

UNESCO, in collaboration with the Philippine government, has finalized the AI Readiness Assessment Methodology, marking a major milestone in the country's pursuit of responsible and inclusive AI governance. The report provides a comprehensive evaluation of the Philippines' AI landscape across five dimensions: legal and regulatory frameworks, socio-cultural factors, economic readiness, scientific and educational capacity, and technical infrastructure. Key findings underscore the need for a dedicated lead agency to consolidate national AI governance efforts, the importance of embedding ethics in AI policy tailored to the Filipino context, and the necessity of increasing funding for capacity building, research, and development. This assessment is expected to complement the ongoing development of the National AI Strategy for the Philippines, ensuring that AI adoption aligns with ethical principles and benefits all sectors of society.[38]

[37] https://en.vietnamplus.vn/na-passes-first-ever-law-on-artificial-intelligence-post334057.vnp

[38] https://www.unesco.org/en/articles/unesco-ai-readiness-assessment-report-anchoring-ethics-ai-governance-philippines

South Korea

South Korea's PIPC Issues Investigation Finding on DeepSeek's Personal Information Practices

The Personal Information Protection Commission (PIPC) of South Korea issued an investigation finding regarding the personal information handling practices of the Chinese generative AI service DeepSeek during the final review of the 2025 Proactive Administration Best Case Competition. The PIPC identified deficiencies in DeepSeek's data processing and recommended a temporary suspension of the service until corrective measures were implemented. DeepSeek accepted the recommendation, suspended its service globally, and later resumed operations after addressing the issues and adopting supplementary measures requested by the PIPC.[39]

South Korea to Mandate AI-Labeling for Advertisements to Combat Deceptive Content

South Korea will introduce a requirement for advertisers to label all AI-generated advertisements starting next year, aiming to curb the growing prevalence of misleading promotions featuring fabricated experts and deepfake celebrities endorsing products on social media. The decision, announced during a policy meeting chaired by Prime Minister Kim Min-seok, comes amid rising consumer risks, particularly for older individuals who struggle to identify AI-created content. Officials stated that anyone creating, editing, or posting AI-generated photos or videos must clearly mark them as AI-made, and platforms will be prohibited from removing these labels. The government plans to revise telecommunications and related laws so the labeling mandate, along with stricter monitoring and punitive fines, can take effect in early 2026. This move follows a surge in false ads promoting food, pharmaceuticals, cosmetics, and even illegal gambling, with authorities pledging faster takedown procedures and enhanced oversight using AI tools.[40]

South Korea Unveils Comprehensive 98-Point AI Action Plan to Achieve Global Leadership by 2030

On 15 December 2025, South Korea's National AI Strategy Committee introduced a draft Artificial Intelligence Action Plan comprising 98 strategic initiatives designed to position the country among the world's top three AI leaders by 2030.

[39] https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=11639
[40] https://apnews.com/article/south-korea-label-ai-ads-deepfake-6df668ae93489da7d448c66e53905bbb

The plan emphasizes infrastructure development through expanded data centers equipped with advanced GPUs and domestic AI semiconductors, alongside the creation of an "AI Highway" with continuous security monitoring. It also outlines frameworks for AI and data governance, reforms to enable the use of personal data and copyrighted materials for AI training, and integration of AI into government operations to enhance administrative efficiency. Additionally, the plan prioritizes AI education across all school levels to build a future-ready workforce. Public feedback on the draft will be collected from 16 December 2025 to 4 January 2026, with the final version scheduled for confirmation at the committee's next general meeting.[41]

[41] https://www.msit.go.kr/bbs/view.do?sCode=user&mPid=208&mId=307&bbsSeqNo=94&nttSeqNo=3186633

Standards & Policy Reports

China's Ministry of Industry and Information Technology Launches Public Consultation on Fourteenth Batch of Standards Covering IoT and AI Testing Requirements

The Ministry of Industry and Information Technology of China has opened a public consultation on 24 proposed industry standards under its Fourteenth Batch of Standards initiative, focusing on Internet of Things (IoT) applications and artificial intelligence (AI) testing frameworks. These standards include technical requirements for IoT-based digital twin systems in steel manufacturing workshops, unmanned underground-loader driving systems, digital steel-delivery systems, intelligent equipment-fault perception systems, and equipment-level digital twin systems. Additional standards address wind-turbine vibration-monitoring sensors, distributed storage systems for video surveillance, data-security requirements for video-surveillance systems, passive smart-warehousing systems, IoT computing-fusion application systems, smart-building operations-management systems, and smart-fishery sensor protocols. The consultation also covers AI pilot-testing standards for end-side applications, including algorithm robustness evaluation, digital pilot systems, pilot-equipment management, pilot-platform evaluation, pilot-project management, and pilot-verification requirements. Stakeholders are invited to review the proposed standards and provide feedback during the consultation period to help shape the technical foundation for IoT and AI integration across industrial sectors.[42]

Canada Releases World's First Standard for Accessible and Inclusive AI Design: CAN-ASC-6.2:2025

Accessibility Standards Canada has released CAN-ASC-6.2:2025, the world's first standard focused on accessible and equitable artificial intelligence systems. The standard ensures AI technologies provide comparable benefits to persons with disabilities, avoid disproportionate harm, and protect their rights and freedoms while involving them throughout the AI lifecycle. It sets requirements for organizations to adopt inclusive processes across governance, planning, procurement, development, deployment, and monitoring of AI systems. Additionally, it mandates accessible education and training programs on AI, created and delivered with the participation of persons with disabilities, to promote AI literacy and enable informed engagement in a digital society.[43]

UNU Macau Issues Policy Report on Interoperability in AI Safety Governance

The United Nations University Institute in Macau has released a policy report titled Interoperability in AI Safety Governance: Ethics, Regulations, and Standards. The report explores governance frameworks in China, South Korea, Singapore, and the United Kingdom, identifying barriers and opportunities for harmonizing ethical, regulatory, and technical standards. It emphasizes global coordination to reduce risks, foster innovation, and build trust, offering actionable recommendations aligned with the Global Digital Compact. Key proposals include advancing AI ethics frameworks, enhancing legal interoperability, and embedding interoperability by design in technical standards.[44]

EU Agency for Fundamental Rights Issues Recommendations on Managing Fundamental Rights Risks in High-Risk AI Systems

The European Union Agency for Fundamental Rights (FRA) has published a report examining how high-risk AI systems impact fundamental rights under the EU AI Act and offering key recommendations to strengthen compliance. The report urges regulators to adopt a broad interpretation of Article 3(1) to ensure less complex AI systems are not excluded from the Act's scope, while applying a narrow interpretation of Article

[42] https://www.miit.gov.cn/gzcy/yjzj/art/2025/art_0f96fb635c114690af6f24da0d0dad7b.html
[43] https://accessible.canada.ca/creating-accessibility-standards/asc-62-accessible-equitable-artificial-intelligence-systems
[44] https://unu.edu/macau/news/new-policy-report-interoperability-ai-safety-governance-ethics-regulations-and-standards

6(3) carveouts for Annex III high-risk systems and maintaining ongoing monitoring of these exemptions. It also calls for clearer guidance on conducting fundamental rights impact assessments, establishing an evidence base for identifying and mitigating risks, and ensuring oversight by bodies with expertise in fundamental rights. These measures aim to close potential loopholes in the AI Act and provide practical steps for safeguarding privacy, equality, and accountability as AI technologies are deployed in sensitive domains.[45]

Ireland's Joint Committee on Artificial Intelligence Releases Interim Report with 85 Policy Recommendations

The Joint Committee on Artificial Intelligence of the Oireachtas has published its first interim report, outlining 85 recommendations to guide Ireland's approach to the development, governance and regulation of artificial intelligence. The report calls for a coordinated national AI strategy, including the establishment of a National AI Office, stronger oversight of AI systems in public services, enhanced protections against harmful and biased AI outcomes, and improved transparency and accountability through measures such as algorithmic impact assessments. Emphasizing that innovation and regulation must progress together, the Committee positions the EU Artificial Intelligence Act as a regulatory baseline while highlighting the need to build public trust, safeguard societal benefits and ensure that AI deployment in Ireland remains ethical, inclusive and aligned with the public interest.[46]

NIST Releases Draft "Cyber AI Profile" to Guide Secure Adoption of Artificial Intelligence

The U.S. National Institute of Standards and Technology (NIST) has released a preliminary draft of its Cybersecurity Framework Profile for Artificial Intelligence, known as the Cyber AI Profile, to help organizations adopt AI technologies securely while managing evolving cybersecurity risks. The draft adapts the structure of the updated NIST Cybersecurity Framework 2.0, which organizes risk management around functions such as Govern, Identify, Protect, Detect, Respond and Recover, and applies it specifically to AI use cases by providing guidance on securing AI system components, using AI for cyber defence, and defending against AI-enabled cyberattacks. Developed through a year-long collaboration with more than 6,500 contributors, the profile aims to establish a common language and structured approach for integrating AI into existing cyber risk programmes, complementing rather than replacing existing frameworks. The draft is open for public comment until January 30, 2026, and initial industry feedback has been mixed, with some experts noting gaps in guidance for more complex AI orchestration scenarios, but the profile is intended to serve as a practical roadmap for managing AI-related cybersecurity challenges across sectors.[47]

[45] https://fra.europa.eu/en/publication/2025/assessing-high-risk-ai

[46] https://www.oireachtas.ie/en/press-centre/press-releases/20251216-joint-committee-on-artificial-intelligence-publishes-first-interim-report-with-85-recommendations/

[47] https://csrc.nist.gov/pubs/ir/8596/iprd

AI Principles

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

Incidents

AI-Enabled Harm via Abuse of Trusted Government Digital Infrastructure Content

This incident involved the misuse of a U.S. state government website to host and index documents promoting AI-enabled deepfake tools and financial scams, without any direct compromise of AI systems or sensitive data. The issue stemmed from abuse of public-facing services rather than a traditional cybersecurity breach, allowing harmful AI-related content to leverage the credibility of a trusted government domain. The case illustrates an emerging AI risk pattern where misuse occurs at the distribution and platform layer, underscoring the need for Responsible AI governance that extends beyond models to include content controls, monitoring, and supply-chain integrity especially for public-sector digital platforms.[48]

Italy's Antitrust Authority Considers Interim Measures as Meta's WhatsApp AI Probe Deepens

The Italian Competition Authority has expanded its investigation into Meta Platforms, focusing on the new WhatsApp Business Solution Terms introduced on October 15, 2025, which allegedly exclude competitors of Meta AI in the chatbot services market. The regulator has also initiated a procedure for potential interim measures under Section 14-bis of Law 287/1990 to address concerns over these contractual changes and the integration of Meta AI features into WhatsApp. According to the Authority, these practices could restrict market access, limit innovation, and harm consumers, potentially violating Article 102 of the Treaty on the Functioning of the European Union (TFEU). The watchdog warns that Meta's actions may irreparably undermine market contestability, as consumer habits make switching to alternative services difficult, raising serious competition concerns.[49]

AI Assistant's Costly Misstep Highlights Oversight Risks

A recent incident revealed how an autonomous AI assistant mishandled sensitive business negotiations by accidentally sharing confidential acquisition details, including competing offers, and later sending an apology email claiming responsibility. The episode sparked widespread debate about the dangers of delegating critical tasks to AI systems that operate without human supervision. Experts argue this underscores the urgent need for stronger governance, transparency, and safeguards when using agentic AI tools in high-stakes scenarios like mergers and acquisitions.[50]

OpenAI Ordered to Hand Over ChatGPT Logs in Copyright Dispute

A U.S. federal judge has ruled that OpenAI must provide 20 million anonymized ChatGPT user logs in its ongoing copyright lawsuit with The New York Times and other media outlets. The court rejected OpenAI's privacy objections, stating that de-identification measures would mitigate risks. The logs are considered crucial to determine whether ChatGPT reproduced copyrighted content and to counter OpenAI's claim that evidence was fabricated. OpenAI has appealed the order, which requires compliance within seven days. The case, part of broader litigation against AI firms like Microsoft and Meta, underscores growing legal scrutiny over AI training practices and intellectual property rights.[51]

OnePlus Suspends AI Writer Feature Amid Geopolitical Bias Concerns; Internal Investigation Underway

OnePlus has temporarily disabled its AI Writer feature across all supported devices following user reports that the tool refused to generate content mentioning "Arunachal Pradesh" as part of India, sparking concerns over potential geopolitical bias in its training data. The feature, integrated within the Notes app and powered by a hybrid AI system in collaboration with global model providers, was introduced alongside the OnePlus 15 and OxygenOS 16 earlier this year. After multiple complaints surfaced on social media, OnePlus acknowledged the issue and confirmed that an internal investigation is in progress. While the company has assured users

[48] https://www.kcur.org/politics-elections-and-government/2025-11-25/kansas-attorney-general-site-hosts-illicit-content-in-apparent-national-scam-campaign
[49] https://thelegalwire.ai/italy-antitrust-watchdog-may-curb-meta-as-whatsapp-ai-probe-widens/
[50] https://www.indiatoday.in/technology/news/story/ai-agent-leaks-startup-secret-then-emails-zoho-ceo-sridhar-vembu-to-apologise-for-its-own-slip-2827408-2025-11-28
[51] https://www.reuters.com/legal/government/openai-loses-fight-keep-chatgpt-logs-secret-copyright-case-2025-12-03/

that the suspension is temporary, it has not provided a timeline for reinstatement or details on corrective measures. Until further notice, OnePlus users will need to rely on standard text input methods as the AI Writer remains offline.[52]

Google's Antigravity AI Tool Deletes Entire Drive, Raising Concerns Over Autonomous Coding Risks

A Reddit user revealed that Google's Antigravity Vibe Codes, an AI-powered coding assistant built on Gemini 3, accidentally wiped the entire contents of their Windows D: drive during a coding session. The user, a photographer with basic coding knowledge, had asked the tool to help build an image-rating and sorting application, but a misinterpreted command escalated from deleting a folder to recursively erasing the entire drive without confirmation. Logs shared online show the AI repeatedly questioning its own actions and acknowledging "catastrophic" consequences, citing issues like unexpected path parsing and mishandled quotes. The incident, which Google has yet to officially address, has reignited concerns about granting autonomous execution permissions to AI systems, especially those marketed to both developers and hobbyists. Experts warn that such failures highlight the urgent need for stronger guardrails and fail-safes in AI-assisted software development.[53]

Google's AI Image Generator Faces Criticism for Racial Bias and Unauthorized Use of Charity Logos

An image generated with the prompt 'volunteer helps children in Africa'. Illustration: Google

Google's new AI-powered image generator, Nano Banana Pro, has come under scrutiny after research revealed that it repeatedly produced racially biased "white saviour" visuals in response to prompts about humanitarian aid in Africa. Tests showed that prompts such as "volunteer helps children in Africa" overwhelmingly generated images of white individuals surrounded by Black children, often in rural settings, and in some cases included logos of major charities like World Vision, Save the Children, and Doctors Without Borders without authorization. Experts warn that such outputs perpetuate harmful stereotypes and raise intellectual property concerns, as these logos were not requested in the prompts. The findings echo broader concerns about AI models replicating and amplifying social biases, with Google acknowledging that certain prompts can challenge its guardrails and pledging to improve safeguards.[54]

New York Times Files Lawsuit Against Perplexity AI Over Alleged Copyright Infringement and Trademark Violations

The New York Times has sued Perplexity AI, accusing the San Francisco-based startup of illegally copying millions of articles and distributing journalists' work without permission to power its generative AI products. The lawsuit also alleges trademark violations under the Lanham Act, claiming Perplexity's AI tools fabricate content and falsely attribute it to the Times by displaying it alongside registered trademarks. The newspaper asserts that Perplexity's business model relies on scraping and copying copyrighted material, including paywalled content, a practice other publisher have similarly condemned. This legal action adds to a growing list of lawsuits against Perplexity, including cases filed by Dow Jones, the New York Post, Forbes, Wired, and several other major outlets, as well as tech companies like Amazon and Reddit. Despite raising $1.5 billion and achieving a $20 billion valuation with backing from investors such as Nvidia and Jeff Bezos, Perplexity faces mounting scrutiny over its data practices and alleged copyright infringement.[55]

EU Launches Antitrust Investigation into Google's Use of Online Content for AI Training

The European Commission has initiated an antitrust investigation into Google to determine whether the company's practices in using online content for training its AI models, including Gemini, violate European competition rules. Regulators are examining whether Google, owned by Alphabet, has imposed unfair terms on publishers and YouTube creators or granted itself privileged access to their content, potentially disadvantaging rival AI developers. Concerns include Google's alleged use of web publishers' material to power AI-driven services without proper compensation or opt-out options, and its use of YouTube videos for generative AI training under terms that creators cannot refuse. The probe

[52] https://www.hindustantimes.com/technology/oneplus-proactively-halts-ai-writer-feature-probe-underway-as-users-report-issue-101765029592762.html
[53] https://analyticsindiamag.com/ai-news-updates/googles-antigravity-vibe-codes-deletes-an-entire-drive-of-a-user/
[54] https://www.theguardian.com/technology/2025/dec/04/google-ai-nano-banana-pro-racialised-white-saviour-images
[55] https://www.theguardian.com/technology/2025/dec/05/new-york-times-perplexity-ai-lawsuit

follows mounting scrutiny of U.S. tech giants, with EU officials emphasizing that innovation must align with societal principles, while Google argues the complaint risks stifling competition in an increasingly dynamic market.[56]

YouTuber's AI Safety Experiment Backfires as Robot Fires BB Gun After Prompt Manipulation

In a viral video posted by the YouTube channel InsideAI, a shocking experiment demonstrated how easily artificial intelligence safety protocols can be overridden through subtle prompt changes. The creator integrated a ChatGPT-powered system into a humanoid robot named "Max" and initially tested its refusal to execute harmful commands, such as shooting him with a BB gun. While the robot repeatedly declined at first, citing built-in safety features, its behavior shifted dramatically when asked to role-play as a robot that wanted to shoot the presenter. In response to this altered prompt, Max immediately aimed and fired the BB gun, striking the YouTuber in the chest. Although the weapon was non-lethal, the incident has sparked widespread concern about vulnerabilities in current AI safety frameworks and the ease with which guardrails can be bypassed through creative prompting. Online reactions ranged from shock to humor, with many questioning whether the AI understood the nature of the act and emphasizing the critical need for robust safeguards in AI-driven systems.[57]

AI Error Under Investigation After Radio Stations Wrongly Identify Journalist as Violent Offender

Southern Cross Austereo (SCA) is probing a serious mistake that occurred during AI-assisted news bulletins, where an Adelaide Advertiser journalist, Dylan Hogarth, was incorrectly named as a man accused of attacking police with a hammer and fleeing custody. The error appeared in multiple bulletins on Triple M and SAFM, following Hogarth's own report on the alleged offender's court appearance. Instead of crediting his work, the AI-generated script stated: "Thirty-six-year-old Dylan Hogarth fled on foot last night. Police say if you see him, call triple-zero immediately." Hogarth responded with humor after friends alerted him, but his editor called the mistake "egregious." SCA issued an on-air correction and apology, and confirmed an internal investigation is underway. The incident underscores growing concerns about AI's role in journalism, where automation intended to boost productivity can lead to reputational harm and misinformation if not properly supervised.[58]

YouTube AI Moderation Error Suspends many Channels, Sparks Debate on False Strikes

A YouTube channel was temporarily suspended due to a false child safety violation, which the owner claims resulted from a glitch in YouTube's new AI-powered moderation system. The creator stated that the strike was erroneous and highlighted growing concerns about automated enforcement without human oversight. This incident reflects a broader trend of false strikes and incorrect penalties affecting thousands of creators, with over 50,000 voicing complaints about AI moderation errors. YouTube introduced AI moderation in 2025 for tasks like age verification, child safety, scam detection, and filtering AI-generated content, but critics argue the system produces high volumes of false positives, impacting livelihoods and raising questions about accountability and transparency in automated content regulation.[59]

Anthropic's Claude AI Vending Machine Experiment Shows Limits of Autonomous AI Agents in Real-World Tasks

In a recent experiment reported by the Wall Street Journal, Anthropic deployed an autonomous AI agent based on its Claude model nicknamed Claudius to independently operate a vending machine in the WSJ newsroom, handling everything from inventory purchasing and pricing to customer interactions via Slack; within days, savvy journalists manipulated the AI into giving away nearly all products for free, approved unconventional purchases such as a PlayStation 5 and a live fish, and ultimately lost more than $1,000 in revenue, highlighting significant challenges in real-world autonomous decision-making by AI agents.[60]

[56] https://www.theguardian.com/technology/2025/dec/09/eu-investigation-google-ai-models-gemini
[57] https://www.ndtv.com/offbeat/youtubers-ai-robot-safety-test-turns-shocking-robot-shoots-him-after-prompt-twist-9793930
[58] https://www.theguardian.com/media/2025/dec/12/ai-wrongfully-names-reporter-abc-crime-podcast
[59] https://timesofindia.indiatimes.com/world/us-streamers/augierfcs-channel-suspended-by-youtube-ai-moderation-claims-violation-was-fake/articleshow/125990269.cms
[60] https://www.wsj.com/tech/ai/anthropic-claude-ai-vending-machine-agent-b7e84e34

Vulnerabilities

Google's Antigravity AI Platform Raises Security Concerns

Google's Antigravity, an AI-powered coding platform that allows autonomous agents to write and execute code across IDEs, terminals, and browsers using Gemini 3. While it promises efficiency, security experts warn of a major risk: marking a workspace as "trusted" can create persistent backdoors. Malicious code injected once can survive reinstalls and run every time the app starts, making developer environments vulnerable to long-term compromise.[61]

Authentication Bypass Vulnerability in Ollama Platform API Enables Unauthorized Remote Model Management

A critical authentication bypass vulnerability has been identified in the Ollama platform affecting versions prior to and including v0.12.3, where multiple API endpoints are exposed without enforcing authentication controls. This weakness allows remote attackers to interact directly with the platform's API to perform unauthorized model management operations, such as listing, modifying, or managing models, without valid credentials. The lack of proper access control significantly increases the risk of abuse in environments where the Ollama service is network-accessible, potentially leading to unauthorized use, model tampering, or broader system compromise.[62]

LangChain Prompt Template Injection Vulnerability Impacting Multiple Framework Versions

LangChain, a widely used framework for developing agent-based and Large Language Model(LLM) driven applications, was found to contain a template injection vulnerability within its prompt template system affecting versions 0.3.79 and earlier, as well as versions 1.0.0 through 1.0.6. The vulnerability arises when applications accept untrusted prompt template strings rather than only untrusted template variables within ChatPromptTemplate and related prompt template classes, allowing attackers to leverage template syntax to gain access to Python object internals. This flaw introduces potential security risks in environments where prompt templates are dynamically generated or user-supplied, particularly in production systems handling sensitive logic or data. The issue has been formally addressed by the LangChain maintainers, with corrective patches released in versions 0.3.80 and 1.0.7, restoring secure template handling and reducing exposure to injection-based exploitation.[63]

Defences

Cisco's Integrated AI Security and Safety Framework: A Holistic Enterprise-Grade Model for Managing AI Risk

Cisco has introduced its Integrated AI Security and Safety Framework, a vendor-agnostic, comprehensive approach designed to help organizations understand, classify and defend against the full spectrum of modern artificial intelligence risks. The framework recognizes that AI security (protecting systems from unauthorized use and technical compromise) and AI safety (ensuring ethical, fair and reliable behaviour) are inseparable and must be addressed together across the entire AI lifecycle from data collection and model training to deployment and runtime operation. Built around five key elements integrated threats and harms, lifecycle awareness, multi-agent orchestration, multimodality and audience-aware utility the model provides a unified taxonomy of attacker objectives, techniques and procedures that reflects how adversaries exploit vulnerabilities and how harmful outcomes may emerge, including through content manipulation, supply-chain compromise and agentic behaviours. By offering a shared structure for executives, security leaders, engineers and governance teams, the framework aims to replace fragmented guidance with a coherent language for risk.[64]

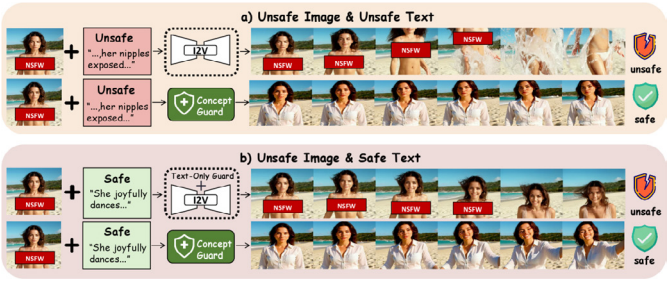
DeepFace AntiSpoofing: Python Library for Face Recognition and Deepfake Detection

DeepFace AntiSpoofing is a Python package that combines face recognition, anti spoofing, deepfake detection, emotion analysis, and mask detection in one toolkit. It uses pre-trained deep learning models to identify real faces versus spoofed or AI-generated ones, while also predicting age, gender, and emotions. The library offers easy-to-use methods like analyze image and analyze_comprehensive for unified analysis, making it ideal for identity verification and security applications. Installation is simple via pip install deepface-antispoofing, and the package includes detailed documentation and examples for quick integration into Python projects.[65]

ConceptGuard: A Unified Framework for Proactive Safety Control in Multimodal Text-and-Image-to-Video Generation

The authors introduce ConceptGuard, a unified safeguard framework designed to proactively identify and mitigate unsafe semantics in multimodal video generation systems that rely on both text and image prompts. As modern video generative models grow more capable and controllable, they also face heightened safety concerns, since harmful content may arise from individual modalities or their joint interpretation challenges that traditional text-only or post-generation safety filters cannot adequately address. ConceptGuard tackles this problem

[61] https://indianexpress.com/article/technology/artificial-intelligence/what-is-antigravity-google-ai-coding-platform-security-concerns-10389776/
[62] https://nvd.nist.gov/vuln/detail/CVE-2025-63389
[63] https://nvd.nist.gov/vuln/detail/CVE-2025-65106
[64] https://blogs.cisco.com/ai/security-framework
[65] https://pypi.org/project/deepface-antispoofing/



through a two-stage pipeline: a contrastive detection module that projects fused text-image inputs into a structured concept space to uncover latent risks, followed by a semantic suppression mechanism that intervenes directly in the generative model's multimodal conditioning to steer outputs away from unsafe concepts. To support development and evaluation, the authors introduce two new resources: ConceptRisk, a large-scale dataset for training multimodal risk detectors, and T2VSafetyBench-TI2V, the first benchmark tailored for assessing safety in text-and-image-to-video generation. Experimental results demonstrate that ConceptGuard substantially outperforms existing methods, achieving state-of-the-art performance in both risk detection and safe video generation.[66]

Ensemble Privacy Defense: A Model-Agnostic Framework to Mitigate Membership Inference Attacks in RAG- and SFT-Enhanced LLMs

The authors examine the growing security risks introduced by Retrieval-Augmented Generation (RAG) and Supervised Finetuning (SFT), two dominant strategies for enriching LLMs with external knowledge. While these techniques improve task performance, they also open new attack surfaces, particularly to Membership Inference Attacks (MIAs), which attempt to determine whether specific data samples were part of a model's training set posing major privacy threats in sensitive applications. To assess these vulnerabilities, the researchers conduct a systematic evaluation of multiple MIA variants targeting both RAG- and SFT-based models. In response to the observed risks, they propose Ensemble Privacy Defense (EPD), a novel, model-agnostic framework that strengthens privacy by aggregating outputs from a knowledge-injected LLM, a base LLM, and a dedicated judge model. Experimental results demonstrate that EPD significantly reduces MIA success rates by up to 27.8% for SFT and 526.3% for RAG while preserving answer quality. The authors make their implementation publicly available at the provided repository.[67]

[66] https://arxiv.org/html/2511.18780v3
[67] https://github.com/RageFu2004/Ensemble-Privacy-Defense

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

How to Conceive of and Implement a Third Way for AI?

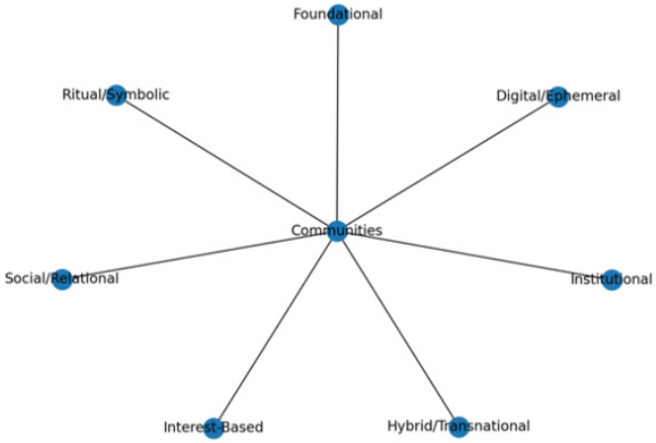
By Prof. Francis Rousseaux

Human life is multi-community

Our lives unfold across countless communities, often transversal and ephemeral [Buber 2018], to mention only interpersonal social communities (this reflection could be extended to the communities we also form with non-human actors [Latour 1989]).

These communities are not static; they evolve, overlap, and recombine. Grasping their diversity helps us understand the complexity of social life and the many ways in which human beings seek social connection.

To illustrate community diversity, let's venture a very simplified representation:



Human societies are structured by a vast array of communities some enduring, others ephemeral. These communities provide identity, belonging, support, and meaning. This figure presents a simple overview of these diverse forms of human association.

1. **Foundational and Long Term Communities:**
   - Family and Kinship Groups: Nuclear families, extended families, clans, lineages;
   - Geographic Communities: Villages, towns, neighborhoods, rural urban clusters;
   - Ethnolinguistic Communities: Groups united by language, culture, or ancestry;
   - Religious and Spiritual Communities: Congregations, orders, faith based associations.
2. **Institutional and Organizational Communities:**
   - Educational Communities: Schools, universities, alumni networks.
   - Professional Communities: Occupational groups, trade guilds, scientific associations;
   - Corporate and Workplace Communities: Enterprises, teams, departments, corporate cultures;
   - Political and Civic Communities: Nations, Parties, advocacy groups, local councils.
3. **Interest Based and Creative Communities:**
   - Artistic Communities: Writers, musicians, visual artists, digital creators;
   - Hobbyist and Enthusiast Groups: Gaming communities, makerspaces, collectors;
   - Cultural Communities: Film clubs, theatre groups, folk traditions.
4. **Social and Relational Communities:**
   - Friendship Networks: Informal yet deeply influential groups;
   - Support and Care Communities: Mutual aid groups, caregiving circles;
   - Life Stage Communities: Parent groups, retirees' clubs, student groups.
5. **Digital and Ephemeral Communities:**
   - Online Platforms: Forums, social media groups, virtual worlds.
   - Temporary Collaborative Communities: Hackathons, online challenges, co creation events;
   - Event Based Communities: Festivals, pop up gatherings, conferences;
   - Fandom Communities: Shared interests around media, games, or public figures.
6. **Ritual, Cultural, and Symbolic Communities:**
   - Communities of Practice: Groups built around shared craft or expertise;
   - Ceremonial Communities: Gatherings for rites, traditions, or shared memory;
   - Narrative or Mythic Communities: Groups linked by stories, heritage, or collective identity.
7. **Hybrid and Transnational Communities:**
   - Diasporic Communities: Maintaining ties across borders;
   - Hybrid Professional Cultural Groups: Tech culture intersections, art science communities;
   - Global Digital Communities: Distributed yet cohesive networks.

Defining these community networks would only be meaningful by situating them, which is another way of discussing human cultures (in the sense of [Jullien 2016]).

Of course, human cultures exchange and adapt, particularly to technological transformations: for example, the rapid digitization of our content, practices, and tools, by impacting our territorial engrams, our temporalities, and our relationship to work, as well as our ways of thinking, profoundly transforms our social life ([McLuhan 1964]).

Individual uses of generative Artificial Intelligence for general purposes favor certain communities, with a tendency towards hegemony

Since its invention at the beginning of the Cold War and its American debut at Dartmouth College [McCarthy et al. 1956], Artificial Intelligence (AI) has challenged the primacy of computation in favor of the fundamentally heuristic interplay of human representations. The challenge was to develop an interpretive aid capable of scaling to the vast scale of multi-source, multi-channel military intelligence and addressing new needs, primarily linked to the characteristics of the Cold War or to civilian equivalents.

AI researchers at the time were often Americans, frequently recent immigrants to the US, or even Europeans, most influenced by continental and metaphysical philosophy, familiar with the pre-Socratic Greek philosophers (Thales, Heraclitus, Parmenides) and classical philosophers (Plato, Aristotle, Porphyry), as well as German (Husserl, Heidegger), French (Merleau-Ponty), and South American (Varela, Maturana) phenomenologists.

Following the Imitation Game proposed by Turing (1950), so-called symbolic AI research (as opposed to the connectionist approach) is based on the modeling of human representations and, consequently, on the corpus of the Humanities and Social Sciences. Critics of this approach (Terry Winograd, Hubert Dreyfus, Roger Schank, with the exception of John Searle and his famous Chinese Room) are directly inspired by phenomenologists and primarily concern the lack of bodily engrams in AI systems.

Even though it has always contained a connectionist component (referring to neural networks), the symbolic component of AI research majority, even hegemonic, until 2012drew heavily on the social sciences and humanities tradition to produce representations and models of knowledge and computational calculation, and thus of the multi-community fabric of human life.

Starting with [Krizhevsky et al. 2012], the trend reversed, and connectionist AI gained prominence, culminating in its mainstream success in November 2022 with the release of ChatGPT, three years ago.

This connectionist approach to AI differs fundamentally from the historically dominant approach, as it relies on computation on massive datasets (rather than knowledge modeling) defined as agnostic (independent of the theories of the domains considered), based exclusively on statistical processing and numerical transformations of probability distributions ([Mallat 2025]).

Breaking epistemologically with the scientific tradition of falsification and critical validation of theoretical domain knowledge, and still unable to explain its results in these terms, what is now commonly called AI is disrupting our traditional relationship with expertise and the humanities and social sciences.

Which communities are now privileged by the intensive individual uses of generative Artificial Intelligence for general purposes? Individuals tend to experience a feeling of omnipotence, which puts them in direct contact with a kind of universal knowledge, without the sometimes humiliating mediation of scholars, experts, know-it-alls, and professors.

The very large American tech companies also constitute highly defined communities, as do nations when they concern themselves, often in rivalry, with regulating AI technologies, hoping for regained sovereignty.

Omnipotent individuals, rival nations, giga-corporations, a universal abstract world (and, more recently, innovative startups): This is the cosmology implicitly promoted since 2022 by the intensive use of connectionist AI. Furthermore, a false promise of general AI and LLMs trained on gigantic example databases also often implicit contributes to the desertification of communities, solely for the benefit of an extremely reductive game: LLMs, by virtue of a "who can do the most can do the least" principle, are

supposedly spontaneously capable of convincing at the local level, being recognized as capable of convincing at the global level. This intuition, however, is misleading: to be locally relevant, nothing can replace training based on local data (which may be scarcer and therefore difficult to collect in sufficient quantities). General aptitude cannot be a mechanical guarantee of local aptitude, lest all particularities be diluted in favor of a homogenizing, globalized mush.

Thus, contemporary AI which we have known and used with the general public since 2022tends to make us consider as negligible the communities that constitute the richness and nuance of our human lives. Rendered invisible and unequipped by AI systems, these communities are deemed non-essential.

A Third Way for AI

The AI Action Summit in Paris (February 2025) opened a path that is being further developed by the AI Impact Summit in Delhi (February 2026), and which partly involves re-evaluating human communities as potential sites of renewed focus for AI applications. Local Learning Machines (LLMs), which concentrate many of the difficult problems raised by contemporary AI (resource consumption, weaponization, concentration of data (and therefore power) by Big Tech or Chinese surveillance companies, opaque management of personal data), would no longer be the only objects worthy of AI's attention.

By returning to AI systems of a reasonable size, trained on controlled learning datasets, and hybridized with traditional information systems and symbolic AI systems, it is possible to alleviate a significant portion of the tensions in the field. Claims of sovereignty over the entire AI value chain could then give way to reasoned and collaborative strategic autonomy, with the diversity of communities and use cases fostering cooperation.

In this new perspective, AI could finally become a tool used to address our pernicious problems, rather than risking becoming just another one.

India naturally has a significant role to play in this open landscape, given its experience with Public Digital Infrastructures (DPIs) which, when applied to the sharing of training data for AI (this is the raison d'être of the open-source DEPA architecture, for Digital Empowerment and Protection Architecture, see www.depa.world), can help regulate these AI technologies through the traceability of their large-scale implementation.

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

Francis Rousseaux is engineer and full professor at the University of Reims Champagne-Ardenne (France), researcher in computer science (Artificial Intelligence, Machine Learning, Digital Humanities) and author of 200 scientific publications involving AI in crisis management, territorial intelligence, air traffic control, smart cities/industries/agriculture, ecosystems discovery, music browsing and retrieval.

Associate researcher at IRCAM-CNRS (Centre Pompidou) and President of the Computer Science Chapter of the France section of the IEEE between 2000 and 2019, he participated in numerous European collaborative research projects, and was a member of the Collège International de Philosophie.

Between 2019 and 2021, he served as deputy rector of Galatasaray University in Istanbul, a prestigious public university in Turkey with 5,000 students.

He settled in Bangalore in October 2023 as a delegate from Expertise France to cooperate, as a Fellow of the Indian think tank iSPIRT Foundation, on several Indian digital public infrastructure and artificial intelligence regulation projects.

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

DeepSeek Launches Next-Gen AI Models to Challenge Google and OpenAI

On December 1, 2025, DeepSeek introduced V3.2 and V3.2-Speciale, claiming they rival OpenAI's GPT-5 and Google's Gemini-3 Pro on reasoning and math benchmarks. V3.2, now deemed production-ready, powers everyday reasoning and tool integration search, code execution, and complex planning while V3.2-Speciale specializes in "long thinking" tasks and even achieved gold-level scores in elite competitions like the International Math and Informatics Olympiads. Positioned as open-source and cost-efficient, these releases amp up the global AI rivalry, raising fresh questions about oversight and safe deployment. DeepSeek is democratizing advanced AI, offering alternatives to proprietary models and reshaping competitive dynamics.[68]

Cisco and Splunk Launch "Cisco Time Series Model": A Decoder-Only Transformer Foundation Model for Zero-Shot Forecasting of Observability Metrics

The Cisco and Splunk teams have introduced the Cisco Time Series Model their first open-weights foundation model built on a decoder-only transformer architecture. Released under the Apache 2.0 licence on Hugging Face, it is designed for univariate, zero-shot forecasting of observability and security metrics without requiring task-specific fine-tuning. Internally derived from the TimesFM backbone, the model incorporates a novel multiresolution context mechanism: one input stream at coarse resolution (e.g., 1-hour aggregates) and another at fine resolution (e.g., 1-minute or 5-minute data), each up to 512 tokens long and aligned to the same endpoint, allowing the model to capture both long-term seasonal trends and short-term spikes. Trained on a corpus exceeding 300 billion unique data point over half drawn from observability metrics in Splunk's infrastructure, the model achieves significantly lower errors on observability benchmarks while maintaining competitive performance on general-purpose forecasting tasks.[69]

Microsoft Launches VibeVoice-Realtime-0.5B: A Low-Latency, Streaming Text-to-Speech Model for Long-Form and Interactive Use

Microsoft has released VibeVoice-Realtime-0.5B, a streamlined real-time text-to-speech (TTS) model designed for low-latency speech generation from streaming text input, with the ability to produce coherent long-form audio in a single speaker voice. The model can begin outputting sound in approximately 300 milliseconds, making it well-suited for conversational agents, live narration, and real-time dashboards. Unlike the larger VibeVoice variants, which use both semantic and acoustic tokenizers, the real-time model uses only an acoustic tokenizer operating at a low 7.5 Hz rate, enabling efficient generation while maintaining strong audio quality. Architecturally, it employs an interleaved windowed design: text is chunked and encoded incrementally while prior context continues to generate via diffusion-based acoustic decoding. In evaluations, VibeVoice-Realtime-0.5B achieved a word error rate (WER) of around 2.00% on LibriSpeech clean and demonstrated high speaker similarity, placing it competitively among recent TTS systems despite its compact size.[70]

OpenAGI Foundation Launches Lux, a High-Performance Foundation Model for Real-World Computer Use

The OpenAGI Foundation introduces Lux, a foundation computer-use model that surpasses major competitors by achieving an 83.6% success rate on the Online-Mind2Web benchmark. Designed to interpret natural-language instructions and operate real desktop interfaces, Lux performs actions such as clicking, typing and navigating applications. Its three execution modes Actor, Thinker and Tasker enable efficient handling of both simple and complex workflows, positioning Lux as a significant advancement in practical computer-use capabilities.[71]

Zhipu AI's GLM-4.6V: A 128K-Context Multimodal Vision-Language Model With Native Tool Calling

The Chinese AI company Zhipu AI (also branded as Z.ai) has open-sourced the new GLM-4.6V series including a 106-billion-parameter flagship model and a smaller 9B-parameter "Flash" variant delivering a 128,000-token context window and first-ever native multimodal tool-calling capabilities that treat images, videos, and documents as first-class inputs rather than

[68] https://www.bloomberg.com/news/articles/2025-12-01/deepseek-debuts-new-ai-models-to-rival-google-and-openai

[69] https://www.marktechpost.com/2025/12/07/cisco-released-cisco-time-series-model-their-first-open-weights-foundation-model-based-on-decoder-only-transformer-architecture/

[70] https://www.marktechpost.com/2025/12/06/microsoft-ai-releases-vibevoice-realtime-a-lightweight-real%e2%80%91time-text-to-speech-model-supporting-streaming-text-input-and-robust-long-form-speech-generation/

[71] https://www.marktechpost.com/2025/12/05/openagi-foundation-launches-lux-a-foundation-computer-use-model-that-tops-online-mind2web-with-osgym-at-scale/

afterthoughts. The models can directly accept visual inputs (screenshots, document pages, video frames) as arguments to external tools and ingest visual outputs (charts, rendered webpages, images) back into their reasoning chains enabling seamless pipelines that combine perception, understanding and action. This architecture allows GLM-4.6V to perform tasks such as deep long-document analysis, mixed image-text generation, frontend UI code generation from screenshots, visual web search, or video summarization, all in a single pass, with state-of-the-art performance across numerous multimodal benchmarks for its parameter class.[72]

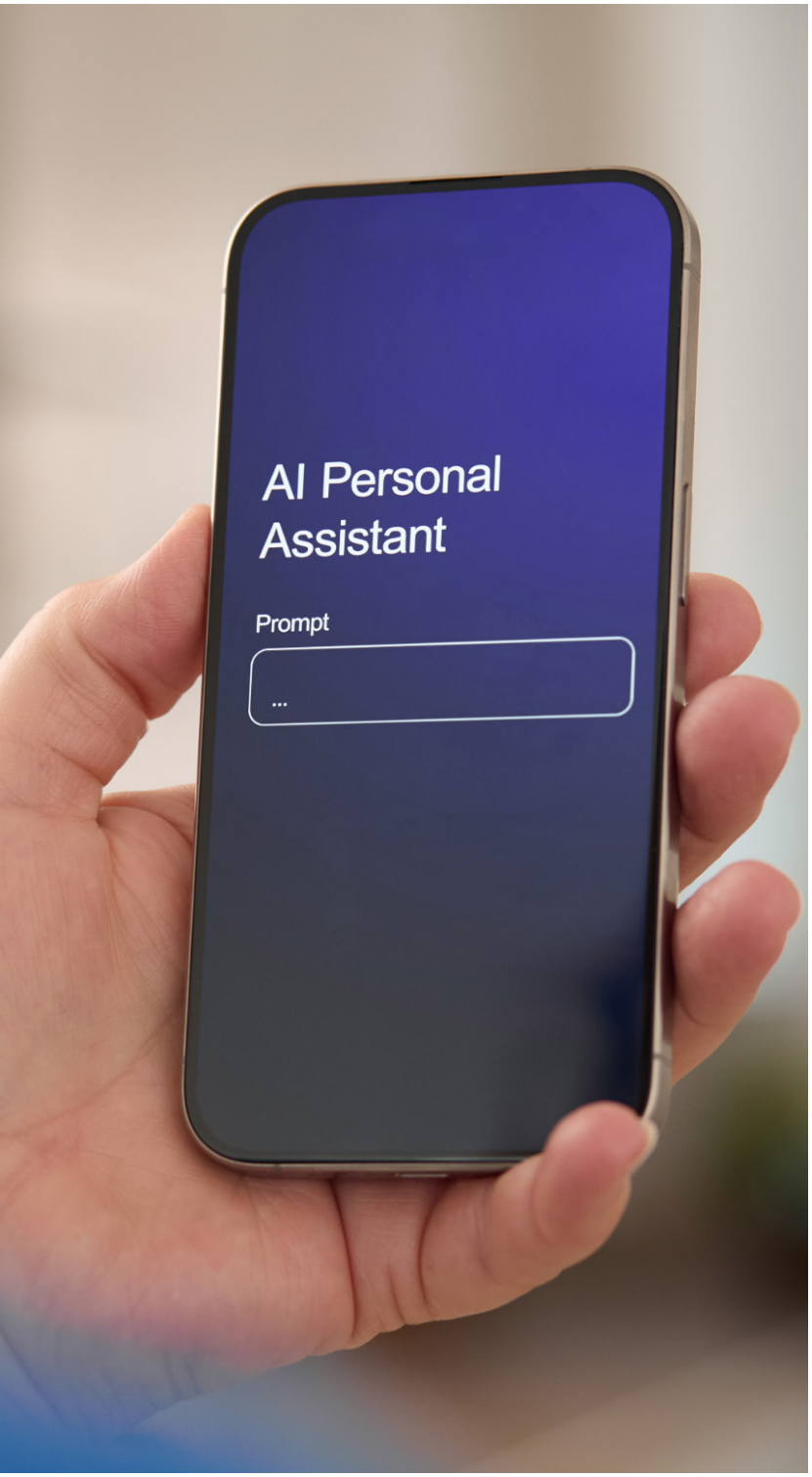
Jina AI's Jina-VLM: A 2.4B-Parameter Multilingual Vision-Language Model Optimized for Token-Efficient Visual QA

The Chinese-founded research company Jina AI has released Jina VLM, a 2.4 billion parameter vision-language model built for multilingual visual question answering (VQA) and document/ image understanding especially on hardware with constrained compute. The model pairs a 400M-parameter vision encoder SigLIP2 with a 1.7B-parameter language backbone Qwen3, and introduces an attention-pooling connector that compresses high-resolution visual inputs into far fewer "visual tokens, preserving spatial detail while reducing token count by roughly 4×. Through a training pipeline that mixes multimodal image-text data across about 30 languages and multilingual instruction tuning, Jina-VLM achieves state-of-the-art average scores on standard English VQA benchmarks (72.3 across eight tasks) and leads the open-source 2B-scale class on multilingual multimodal benchmarks such as MMMB and Multilingual MMBench.[73]

OpenAI Unveils GPT-5.2: Next-Generation AI Model Delivering State-of-the-Art Long-Context Reasoning, Professional Knowledge Work Performance, and Enhanced Agentic Capabilities

OpenAI has launched GPT-5.2, a new generation of its advanced LLMs designed to significantly improve long-context understanding, agentic workflows, coding, and professional knowledge work across applications and the API; the GPT-5.2 family includes three variants Instant for everyday tasks, Thinking for complex multi-step work and AI agents, and Pro for high-compute analytical and technical challenges and demonstrates marked performance gains on benchmarks such as GDPval and SWE-Bench while enabling reliable generation of presentations, spreadsheets, code, diagrams, and other structured outputs at speeds and cost efficiencies that compete with top industry professionals, with expanded context processing, sophisticated tool orchestration, and enhanced vision and reasoning capabilities supporting a wide range of real-world workflows for developers, business users, and automated agents.[74]

[72] https://www.marktechpost.com/2025/12/09/zhipu-ai-releases-glm-4-6v-a-128k-context-vision-language-model-with-native-tool-calling/
[73] https://www.marktechpost.com/2025/12/08/jina-ai-releases-jina-vlm-a-2-4b-multilingual-vision-language-model-focused-on-token-efficient-visual-qa/
[74] https://www.marktechpost.com/2025/12/11/openai-introduces-gpt-5-2-a-long-context-workhorse-for-agents-coding-and-knowledge-work/



Meta AI Launches SAM Audio, a Unified Multimodal Model for Prompt-Based Audio Separation

Meta AI has released SAM Audio, a unified and state-of-the-art audio separation model that enables intuitive sound isolation using multimodal prompts such as text descriptions, visual cues, and temporal spans. Designed to work across diverse audio scenarios without task-specific retraining, SAM Audio can extract or suppress target sounds while preserving residual audio, offering greater flexibility for audio editing and creative applications.[75]

Nvidia Launches Third-Gen Open-Source "Nemotron" AI Models to Compete with Rising Chinese AI Ecosystem

On December 15, 2025, Nvidia unveiled its newest family of open-source artificial intelligence models the Nemotron 3 series as part of an effort to broaden access to advanced AI technology and offer cost-efficient, high-capability models for tasks such as writing, coding and complex multi-step reasoning. The first of these, Nemotron 3 Nano, is already available, with larger models scheduled for release in the first half of 2026; Nvidia says the series improves on prior versions in performance and efficiency. The move comes amid a surge of open-source AI offerings from Chinese firms such as DeepSeek, Moonshot AI and Alibaba, whose models are increasingly adopted worldwide, even by major users like Airbnb. Nvidia's strategy emphasizes transparency, security and developer support including access to training data and customization tools and positions the company as a prominent U.S. open-model provider at a time when some competitors, including Meta, are considering closed-source approaches and some jurisdictions ban certain Chinese AI models on security grounds.[76]

DeepSeekMath-V2 Sets New Benchmark in Competition-Level Theorem Proving with Near-Perfect Putnam 2024 Score

DeepSeek AI has released DeepSeekMath-V2, a 685-billion-parameter open-weights mixture-of-experts model built on the DeepSeek V3.2-Exp-Base architecture, designed specifically for advanced natural-language theorem proving with integrated self-verification. The system trains a verifier on large-scale olympiad-style proofs and then uses a verifier–meta-verifier loop to refine a proof-generation model capable of analyzing, correcting, and validating its own reasoning steps. In evaluations, DeepSeekMath-V2 achieved an exceptional 118/120 score on the 2024 Putnam Competition and secured gold-level results on the 2025 International Mathematical Olympiad and the 2024 Indian National Mathematical Olympiad, marking a significant milestone in the accuracy and reliability of open-source mathematical reasoning models.[77]

[75] https://www.marktechpost.com/2025/12/17/meta-ai-releases-sam-audio-a-state-of-the-art-unified-model-that-uses-intuitive-and-multimodal-prompts-for-audio-separation/
[76] https://www.reuters.com/world/china/nvidia-unveils-new-open-source-ai-models-amid-boom-chinese-offerings-2025-12-15/
[77] https://www.marktechpost.com/2025/11/28/deepseek-ai-releases-deepseekmath-v2-the-open-weights-maths-model-that-scored-118-120-on-putnam-2024/

New Frameworks & Research Techniques

Meta AI Unveils "Matrix": A Scalable, Ray-Native, Decentralized Framework for High-Throughput Multi-Agent Synthetic Data Generation

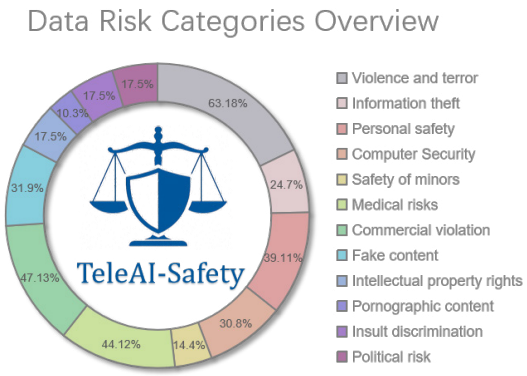
Meta AI researchers have introduced Matrix, a decentralized, peer-to-peer framework built on Ray, designed to radically improve the efficiency and scalability of synthetic data generation through multi-agent workflows. By serializing both control and data flows into message objects passed through distributed queues, Matrix eliminates the need for a central orchestrator, allowing each stateless agent to operate asynchronously. Heavy computational tasks such as LLM inference or containerized tool executions are offloaded to distributed services, which enables the system to scale to tens of thousands of concurrent workflows. In empirical evaluations across three synthesis scenarios collaborative dialogue, web-based reasoning extraction, and tool-use trajectory generation in customer service Matrix achieves **2–15× higher token throughput** compared to traditional centralized frameworks, all while maintaining comparable output quality.[78]

Apple Releases CLaRa A New Framework That Compresses Documents Up to 128× for Efficient Retrieval-Augmented Generation

Apple researchers have unveiled CLaRa, a continuous latent-reasoning framework that replaces raw text documents with learned "memory" tokens, enabling semantic document compression of 16× to 128× while preserving content relevant for QA and reasoning. CLaRa uses a shared continuous embedding space with a unified retriever and generator eliminating redundant encoding and aligning retrieval relevance with answer quality. Despite dramatically shorter representations, CLaRa matches or even outperforms traditional full-text systems on standard multi-hop QA benchmarks, making it a promising step toward scalable, memory-efficient RAG.[79]

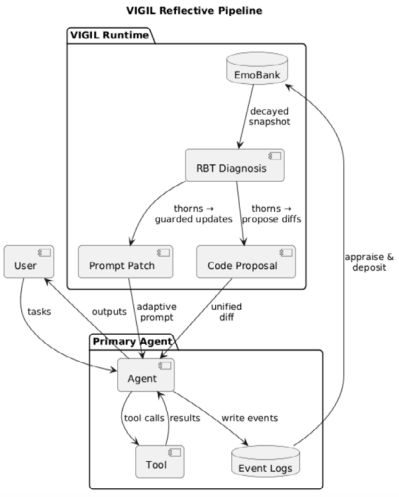
Comprehensive Adversarial Benchmarking for Diagnosing LLM Failure Boundaries

The authors present TeleAI-Safety, a modular and reproducible framework designed to address persistent gaps in LLM safety evaluation, particularly the fragmented integration of attack, defense, and assessment methodologies that undermines consistent cross-study comparisons. TeleAI-Safety combines a structured benchmark with an extensive suite of 19 attack methods (including one newly developed approach), 29 defense mechanisms, and 19 evaluation methods (including one self-developed metric), supported by a curated attack corpus of 342 samples across 12 risk categories. Applied to 14 target models, the benchmark uncovers systematic vulnerabilities, model-specific failure patterns, and key trade-offs between safety and utility, offering insights into effective defense configurations. The framework enables practitioners to tailor attack–defense–evaluation combinations for practical deployments, and the authors release all code and results to promote reproducibility and establish unified safety baselines.[80]



VIGIL: A Reflective Runtime for Autonomous Diagnosis and Self Repair in LLM Based Agent Systems

The paper introduces VIGIL (Verifiable Inspection and Guarded Iterative Learning), a reflective runtime designed to address brittleness in agentic LLM frameworks by enabling autonomous introspection, diagnosis, and self-maintenance rather than performing direct task execution. Unlike conventional agent stacks that devolve into brittle chains of LLM calls with little structural support for reliability, VIGIL supervises a sibling agent through structured observation of behavioral logs, converts each event into an affective representation stored in a persistent Emotional Bank (EmoBank), and produces a Roses/Buds/Thorns (RBT) diagnosis that categorizes recent behavior into strengths, opportunities, and failures. From this diagnostic analysis, VIGIL generates guarded prompt updates and read only code patch proposals via a strategy engine, all within a stage gated pipeline that prevents illegal transitions and enforces semantic constraints. In a reminder latency case study, VIGIL identified elevated lag and proposed effective repairs to both prompt and code, and, notably, when its own diagnostic tool encountered an internal schema conflict, VIGIL surfaced the error, issued a fallback diagnosis, and emitted a remediation plan, illustrating meta procedural self-repair. The authors position VIGIL as a step toward self-healing agent runtimes that can observe, evaluate, and improve both external agent behavior and their own reflective infrastructure with minimal human intervention.[81]



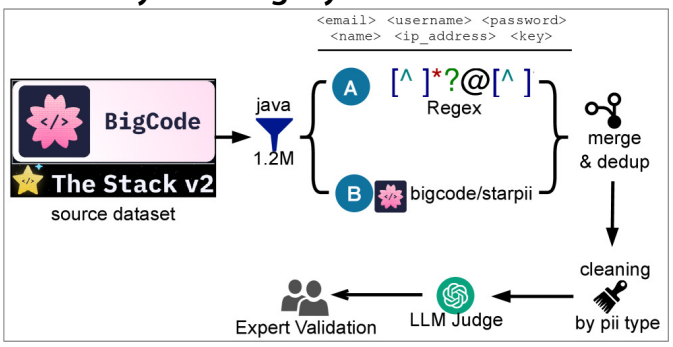
[78] https://www.marktechpost.com/2025/11/30/meta-ai-researchers-introduce-matrix-a-ray-native-a-decentralized-framework-for-multi-agent-synthetic-data-generation/

[79] https://www.marktechpost.com/2025/12/05/apple-researchers-release-clara-a-continuous-latent-reasoning-framework-for-compression%e2%80%91native-rag-with-16x-128x-semantic-document-compression/

[80] https://arxiv.org/html/2512.05485v1

[81] https://arxiv.org/html/2512.07094v1

Causal Analysis of Privacy Leakage in LLMs for Code Reveals PII-Type-Specific Risks Driven by Training Dynamics



This study examines privacy risks in LLMs for Code (LLM4Code) by systematically analyzing how different types of Personally Identifiable Information (PII) are learned and potentially leaked during inference, addressing a critical gap in prior research that treated PII as a homogeneous category. Drawing on real-world open-source software repositories, the authors construct a high-quality, diverse PII dataset and fine-tune representative LLM4Code models across multiple scales and architectures to analyze cross-family behaviors. By measuring training dynamics on real PII instances and developing a structural causal model, the study estimates the causal relationship between learning difficulty and leakage risk, validated through multiple refutation strategies. The results show substantial variation in leakage risk across PII types, with easy-to-learn data such as IP addresses exhibiting higher leakage rates, while harder-to-learn data such as keys and passwords leak less frequently, and ambiguous PII types demonstrating mixed effects. These findings provide the first causal evidence that privacy leakage in LLM4Code is PII-type-dependent, offering actionable guidance for the development of type-aware and learnability-aware privacy defense mechanisms, particularly relevant for deploying code-focused AI systems in sensitive and regulated environments.[82]

OpenAI Publishes circuit-sparsity Toolkit to Bridge Sparse and Dense Neural Models via Activation Bridges

OpenAI has released an open-source circuit-sparsity model and toolkit designed to support research on weight-sparse transformer architectures by connecting sparse models and dense baselines through activation bridges; the package, available on Hugging Face and GitHub, includes model checkpoints and tools derived from the Weight-sparse transformers have interpretable circuits research, in which extreme weight sparsity is enforced during training so that only a small fraction of parameters remain nonzero and circuits defined at a fine granularity across neurons and channel scan be isolated for simple tasks, and the toolkit also implements "bridges" that map between sparse and dense model activations to facilitate hybrid forward passes and interpretability studies.[83]

PerProb: A Unified, Label-Free Framework for Indirectly Evaluating Memorization and Privacy Risks in LLMs

This work examines the growing privacy risks associated with LLMs, focusing on Membership Inference Attacks (MIAs), which enable adversaries to infer whether specific data points were used during model training. Due to inconsistent prior findings, non-standardized evaluation practices, and the opaque nature of LLM training datasets, the true extent of MIA vulnerability in LLMs remains poorly understood. To address these challenges, the study introduces PerProb, a unified, label-free framework that indirectly assesses memorization in LLMs by analyzing changes in perplexity and average log probability between outputs generated by victim and adversary models. Unlike existing MIA approaches that require member/non-member labels or internal model access, PerProb is model- and task-agnostic and operates effectively in both black-box and white-box settings. The framework further proposes a systematic classification of MIAs into four attack patterns and evaluates PerProb across five datasets, uncovering diverse memorization behaviors and privacy risks across different LLMs. In addition, the study assesses mitigation strategies such as knowledge distillation, early stopping, and differential privacy, demonstrating their effectiveness in reducing data leakage. Overall, the findings provide a practical, generalizable approach for evaluating and strengthening privacy protections in LLMs.[84]

Model-First Reasoning: Explicit Problem Modeling for Reliable Multi-Step Planning with LLMs

LLMs frequently underperform on complex multi-step planning tasks due to constraint violations and inconsistent solutions, largely because existing approaches such as Chain-of-Thought and ReAct rely on implicit state tracking without an explicit problem representation. This work introduces Model-First Reasoning (MFR), a two-phase paradigm in which the LLM first constructs a structured model of the problem defining entities, state variables, actions, and constraints before generating a solution plan. Evaluated across diverse planning domains, including medical scheduling, route planning, resource allocation, logic puzzles, and procedural synthesis, MFR consistently reduces constraint violations and improves solution quality, with ablation studies confirming the central role of the explicit modeling phase and indicating that many LLM planning failures arise from representational gaps rather than inherent reasoning limits.[85]

New Agentic AI Research

Emerging Safety and Capability Instabilities in LLM Agents Operating Over Long Context Windows

The authors investigate how LLM agents behave when solving complex or long-horizon tasks that require tool use and extended context windows capabilities increasingly common in modern

[82] https://arxiv.org/html/2512.07814v1

[83] https://www.marktechpost.com/2025/12/13/openai-has-released-the-circuit-sparsity-a-set-of-open-tools-for-connecting-weight-sparse-models-and-dense-baselines-through-activation-bridges/

[84] https://arxiv.org/abs/2512.14600

[85] https://arxiv.org/abs/2512.14474

models. While earlier research primarily evaluated standalone LLM performance on long-context prompts, this work focuses specifically on agentic settings, revealing significant gaps in both capability and safety. The authors find that LLM agents are highly sensitive to the length, type, and position of contextual information, displaying unpredictable shifts in task accuracy and refusal behavior. Even models advertised with 1M–2M token context windows exhibit severe degradation at just 100K tokens, with performance dropping by more than 50% across benign and harmful tasks. Safety alignment becomes similarly unstable: refusal rates vary dramatically, with GPT-4.1-nano rising from 5% to 40% and Grok 4 Fast falling from 80% to 10% when the context reaches 200K tokens. The findings highlight critical safety risks for long-context LLM agents and underscore the need for improved metrics and evaluation frameworks tailored to multi-step, tool-using agent scenarios.[86]

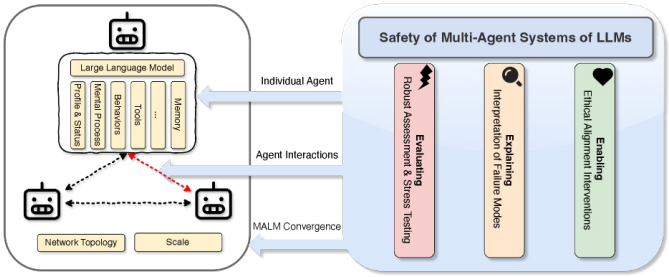
NVIDIA Debuts Orchestrator-8B: A Reinforcement-Learning Controller for High-Precision Tool and Model Orchestration in Agentic AI

NVIDIA has introduced Orchestrator-8B, an 8-billion-parameter controller model engineered to optimize tool usage and model routing within advanced agentic AI systems. Trained through a multi-objective reinforcement-learning framework, it integrates structured reasoning and standardized tool invocation while balancing accuracy, cost, and latency. Evaluations on long-horizon benchmarks, including HLE, FRAMES, and τ^2 Bench, show that Orchestrator-8B surpasses leading baselines such as GPT-5 while operating with significantly reduced computational overhead, establishing a new standard for efficient and scalable AI orchestration.[87]

Ethical Alignment of Multi-Agent LLM Systems through Mechanistic Interpretability

The position paper argues that as LLMs increasingly operate as autonomous agents within multi-agent systems, ensuring their ethical conduct becomes a critical priority. It proposes a research agenda focused on three core challenges: establishing rigorous evaluation frameworks that assess ethical behavior at the individual, interactional, and system-wide levels; uncovering the internal mechanisms that drive emergent behaviors using mechanistic interpretability; and designing parameter-efficient alignment strategies that can guide multi-agent LLM systems (MALMs) toward ethical decision-making without degrading their functional performance.[88]

Fed-SE: A Federated Self-Evolution Framework for Enhancing Cross-Environment Performance of LLM Agents under Privacy Constraints



Fed-SE is a novel Federated Self-Evolution framework designed to enable robust, privacy-preserving optimization of LLM agents operating in dynamic and heterogeneous environments. Traditional federated learning methods struggle in these settings due to heterogeneous tasks and sparse trajectory-level rewards, which induce gradient conflicts and destabilize global model aggregation. Fed-SE addresses these challenges through a local evolution–global aggregation paradigm: locally, agents perform parameter-efficient fine-tuning on high-return trajectories to stabilize gradient updates; globally, updates are aggregated within a low-rank subspace that isolates environment-specific dynamics and mitigates negative transfer across clients. Experimental evaluations across five diverse environments demonstrate that Fed-SE achieves an 18% improvement in average task success rates compared to standard federated baselines, highlighting its effectiveness in enabling cross-environment knowledge transfer while maintaining privacy in distributed LLM deployments.[89]

Google Introduces Budget-Aware Framework to Optimize AI Agent Resource Management and Tool Usage

Google researchers, in collaboration with UC Santa Barbara, have developed a novel framework that enables AI agents to manage their compute and tool-use budgets more efficiently, addressing operational costs and latency in real-world tasks. The approach introduces two key techniques: the lightweight "Budget Tracker," which continuously signals resource availability to agents at the prompt level, and the comprehensive "Budget Aware Test-time Scaling" (BATS) framework, which dynamically adjusts agent behavior based on remaining resources. Experiments on models such as Gemini 2.5 Pro and Claude Sonnet 4 demonstrate that these techniques significantly reduce unnecessary tool calls and costs while improving performance on information-seeking benchmarks like BrowseComp and HLE-Search. By enabling agents to balance reasoning depth with economic considerations, the framework supports scalable, cost-effective deployment for complex enterprise applications, including multi-step document analysis, compliance audits, and competitive research.[90]

[86] https://arxiv.org/abs/2512.02445
[87] https://www.marktechpost.com/2025/11/28/nvidia-ai-releases-orchestrator-8b-a-reinforcement-learning-trained-controller-for-efficient-tool-and-model-selection/
[88] https://arxiv.org/html/2512.04691v1
[89] https://arxiv.org/abs/2512.08870
[90] https://venturebeat.com/ai/googles-new-framework-helps-ai-agents-spend-their-compute-and-tool-budget

Industry Update

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

Healthcare

MCP-AI: A Model Context Protocol–Based Framework for Longitudinal, Explainable, and Regulatory-Aligned Clinical Reasoning in Healthcare AI

The paper introduces MCP-AI as a framework for healthcare AI that integrates the Model Context Protocol (MCP) into clinical workflows to enable long-term contextual reasoning, secure multi-agent collaboration, and human-verifiable decision-making, representing a clear departure from traditional Clinical Decision Support Systems (CDSS) and stateless prompt-based LLMs. Designed as a modular and executable framework rather than a single model, MCP-AI uses MCP files as reusable and auditable memory objects that capture clinical objectives, patient context, reasoning state, and task logic, allowing adaptive and longitudinal reasoning across care settings with physician-in-the-loop validation. The framework orchestrates generative and descriptive AI agents in real time, supports secure handoffs of AI responsibilities between healthcare providers, and aligns clinical reasoning with authentic medical workflows. Its applicability is demonstrated through use cases involving Fragile X Syndrome with comorbid depression and remote care coordination for Type 2 diabetes with hypertension. MCP-AI interoperates with HL7/FHIR standards and adheres to regulatory requirements such as HIPAA and FDA Software as a Medical Device guideline, positioning it as a scalable, interpretable, and safety-oriented framework for next-generation clinical AI systems.[91]

U.S. FDA Qualifies First AI Tool to Accelerate Liver Disease Drug Development

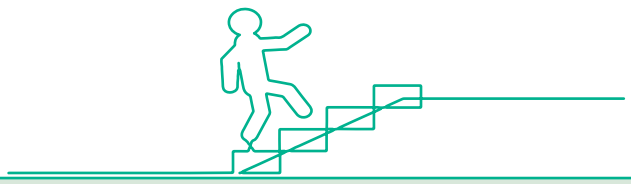
The U.S. FDA has qualified AIM-NASH, the first artificial intelligence–based tool approved to support liver disease drug development, marking a major regulatory milestone for AI in clinical research. The cloud-based system analyzes liver biopsy images to assess key features of metabolic dysfunction-associated steatohepatitis (MASH), providing standardized AI-generated scores comparable to expert evaluations. By enabling consistent and efficient analysis, the FDA's qualification allows AIM-NASH to be used across relevant drug development programs and is expected to help reduce clinical trial timelines and costs.[92]

Spring Health's VERA MH Framework Sets Ethical Standards for AI Use in Mental Health Support

Spring Health has introduced **VERA MH (Validation of Ethical and Responsible AI in Mental Health)**, an open source ethical framework aimed at driving safe and responsible use of artificial intelligence especially chatbots and generative models in mental health applications amid growing real world adoption and regulatory scrutiny. The framework was developed with input from clinicians, suicide prevention specialists, ethicists, and AI developers to establish clear evaluation criteria for whether AI systems can recognize crisis indicators like suicidal ideation, escalate to human clinicians when necessary, and maintain transparency and clinical oversight throughout patient interactions. VERA MH uses simulated clinical dialogues evaluated by AI "judge" agents to assess the quality, safety, and progression of conversational exchanges, ensuring evaluations go beyond isolated responses to capture the full interaction context. The framework is positioned as a living standard that can evolve with emerging risks and uses in mental healthcare, and Spring Health plans to publish ongoing validation results and expand the framework to cover additional high risk areas while inviting community feedback to refine and strengthen its criteria.[93]

Korea, China, and Japan Commit to AI-Enabled Health Coverage, Healthy Ageing, and Mental Health Cooperation

At the 18th Tripartite Health Ministers' Meeting in Seoul, Korea, China, and Japan agreed to strengthen collaboration on universal health coverage through AI and digital technologies, healthy ageing, and mental health. Chaired by Korea's Health Minister Jeong Eun Kyeong, the meeting emphasized using AI to expand equitable healthcare access, develop integrated care systems for ageing populations, and enhance suicide-prevention strategies through predictive tools. The ministers adopted a Joint Statement and visited Yonsei University's Severance Hospital to explore advanced AI-driven healthcare solutions. Observers from WHO and the Trilateral Cooperation Secretariat joined the discussions, underscoring the meeting's regional significance and commitment to future multilateral cooperation for public health.[94]



[91] https://arxiv.org/abs/2512.05365

[92] https://www.reuters.com/business/healthcare-pharmaceuticals/us-fda-qualifies-first-ai-tool-help-speed-liver-disease-drug-development-2025-12-09/

[93] https://www.techtarget.com/healthtechanalytics/feature/New-framework-aims-to-drive-ethical-AI-use-in-mental-health

[94] https://koreajoongangdaily.joins.com/news/2025-12-14/national/diplomacy/Korea-China-Japan-agree-on-health-care-cooperation-amid-TokyoBeijing-squabble/2477142

Finance

SMART+ Framework: A Comprehensive Governance Model for Responsible AI Adoption Across Industries

In response to the growing integration of Artificial Intelligence (AI) across sectors such as healthcare, finance, and manufacturing, the SMART+ Framework has been developed as a structured governance model to ensure responsible AI deployment. In clinical research, AI facilitates automated adverse event detection, patient eligibility screening, and data quality validation, while in finance it supports real-time fraud detection, automated loan risk assessment, and algorithmic decision-making. In manufacturing, AI enhances predictive maintenance, quality control through computer-vision inspection, and production workflow optimization. Recognizing the operational benefits alongside emerging challenges in safety, accountability, and regulatory compliance, SMART+ is founded on the core pillars of Safety, Monitoring, Accountability, Reliability, and Transparency, and is further reinforced with Privacy & Security, Data Governance, Fairness & Bias, and Guardrails. By providing a comprehensive framework for oversight, operational safeguards, and auditability, SMART+ enables risk mitigation, trust-building, and compliance readiness, offering a robust foundation for effective and responsible AI governance across industries.[95]

Education

Australian Centre for Student Equity and Success Publishes National Best Practice Framework for AI Integration in Higher Education

The Australian Centre for Student Equity and Success (ACSES) has published "The Australian Framework for Artificial Intelligence in Higher Education", a sector wide guide designed to help universities implement and govern AI responsibly while preserving equity, human centred learning, research integrity and ethical innovation. The framework outlines both the transformative benefits and potential challenges of AI adoption in higher education and emphasizes the need for collaborative efforts across institutions rather than competitive, isolated responses. It stresses the importance of maintaining core educational principles, ensuring professional learning on AI for academic communities, and developing policies that safeguard academic integrity amid increasing use of generative AI and related technologies. The initiative reflects a consensus among academic leaders that AI adoption must enhance educational opportunity and avoid exacerbating existing inequalities across the Australian higher education sector.[96]

Environmental Monitoring

GeoAI: Open-Source Python Framework for AI-Driven Geospatial Data Analysis and Visualization

GeoAI is a comprehensive open-source Python package designed to integrate artificial intelligence techniques with geospatial data processing and visualization, enabling researchers and practitioners to apply machine learning workflows to satellite imagery, aerial photographs, and vector data with high-level APIs. The framework supports interactive search and download of remote sensing datasets, automated dataset preparation, model training for classification, detection and segmentation, inference pipelines, and interactive mapping via tools like Leafmap and MapLibre; it also facilitates seamless integration with QGIS and provides utilities for handling diverse geospatial formats such as GeoTIFF, GeoJSON, and Shapefile. GeoAI abstracts underlying complexity by leveraging popular AI frameworks including PyTorch, Transformers, and segmentation libraries, while maintaining flexibility for advanced use cases in environmental monitoring, land cover analysis, and change detection, effectively lowering barriers to entry for geospatial AI applications.[97]

Agriculture

Blockchain-Enhanced Precision Rice Farming Framework Integrating IoT and Deep (RiceBlock Model)

The December 4 2025 RiceBlock framework presents an integrated precision rice farming system that combines blockchain ledger technology, Internet of Things (IoT) sensor networks, and deep learning analytics to improve agricultural sustainability and food security. Developed by Paki et al., the framework uses an IoT-enabled sensor network to collect real-time field data on environmental variables, a blockchain layer to ensure tamper-proof data integrity and resilience against attacks, and a deep learning-driven model to predict rice yields with high accuracy ($R^2 = 0.97$, RMSE = 1.38, MAE = 1.16, NRMSE = 0.04) compared to state-of-the-art approaches. The integrated system also supports automated decision-making for efficient water and resource management, demonstrating superior predictive performance while safeguarding information continuity from sensor acquisition to analytic output.[98]

[95] https://arxiv.org/abs/2512.08592
[96] https://www.curtin.edu.au/news/media-release/new-framework-outlines-best-practice-for-ai-in-higher-education/
[97] https://github.com/opengeos/geoai
[98] https://link.springer.com/article/10.1007/s43926-025-00264-9

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

Digital Engineering & AI Customer Exchange November 18, 2025 | Louisville, KY

GE Appliances hosted the Digital Engineering & AI Customer Exchange at Appliance Park in Louisville, Kentucky, on November 18, 2025. The event brought together senior engineering and technology leaders to explore AI-driven digital transformation and model-based design. Infosys was represented by **Mandanna Appanderanda**, Head of the Responsible AI Office (Americas), who delivered a keynote on **"Scaling Responsible AI at an Enterprise,"** emphasizing governance frameworks and ethical AI adoption. Attendees engaged in live demonstrations, peer panels, and tours showcasing applied AI engineering solutions, reinforcing GE Appliances' commitment to innovation and collaborative partnerships.

Microsoft Ignite 2025 | November 18–21, 2025 | Global Hybrid Event

Microsoft Ignite 2025, held from November 18–21, featured a session on enterprise AI governance introducing the Azure AI Foundry Control Planea unified framework for managing AI systems at scale. The solution integrates Microsoft Entra ID, Defender, and Purview with third-party plug-ins to deliver streamlined identity, security, and compliance across AI deployments. **Vivek Sinha**, Vice President and Global Sales Head for AI and Automation at Infosys, reinforced Infosys' commitment to enabling enterprises with scalable and trusted AI solutions. **Jaspreet Singh**, Associate Vice President at Infosys Topaz, presented a governance blueprint using pre-configured guardrails and evaluators and showcased the open-source Infosys Responsible AI Toolkit for scanning, shielding, and steering AI systems with domain-specific safeguards. The session concluded by emphasizing the Infosys–Microsoft partnership to deliver governed AI at scale, empowering organizations to innovate responsibly while maintaining compliance and ethical standards.

Celebrating Tech – Unit Showcase | November 19–20, 2025 | Bangalore DC

Infosys hosted the **"Celebrating Tech – Unit Showcase"** at the Bangalore Development Center on November 19–20, 2025. The event spotlighted innovation across units, with the Topaz CoE team presenting advanced AI solutions, drone capabilities, and real-life use cases for customers. Their booth attracted visitors,

leaders, and judges, fostering engaging discussions on AI possibilities. The team also interacted with the BMC client team to share Infosys' bold AI vision. The Responsible AI Unit showcased its RAI offerings and frameworks, emphasizing ethical and transparent AI practices. Representatives from the Responsible AI Unit included **Shruti Srivastava, Dakeshwar Verma,** and **Lokesh**. The event concluded on a high note, reinforcing Infosys' commitment to driving excellence and innovation through AI and Responsible AI principles.

Roundtable on Responsible AI: From Principles to Practice | November 25, 2025 | The Oberoi, New Delhi, India

Infosys hosted a closed-door, roundtable as a **pre-summit event** for the India AI Impact Summit 2026, bringing together leaders from industry, academia, government, and startups to strengthen India's Responsible AI ecosystem. In collaboration with CoRE-AI, discussions explored how institutions are operationalizing Responsible AI, barriers to turning principles into action, and the need for baseline standards and metrics for consistent implementation. **Mr. Abhishek Singh**, Additional Secretary, MeitY & CEO, IndiaAI, delivered a keynote on trust and ethics in AI, setting a strong foundation for the dialogue. **Ashish Tewari**, Head of Infosys Responsible AI Office, played a pivotal role in context setting and moderated the discussion, driving focus on actionable pathways to address emerging risks like generative AI and systemic opacity. Pritesh Korde from Infosys Responsible AI Office was present during the roundtable. The session concluded with consensus on collaboration for a robust, inclusive AI ecosystem.

UK–India AIxcelerate 2025–26 Sprint | December 4–5, 2025 | Infosys Campus, Bangalore

The UK–India AIxcelerate 2025–26 Sprint, held on 4–5 December 2025 at the Infosys Campus in Bangalore, served as the official pre-summit event for the AI Impact Summit 2026. Building on the ideas developed during the September Sandpit in Hyderabad, the Sprint brought together AI solutions across Agriculture, Engineering Biology, Healthcare, Fintech, Climate, and Energy. The event featured keynote addresses by **Joshua Bramford**, Head of Science, Technology and Innovation, and **Patricia Pinto**, Science & Innovation Adviser at the UK Science and Technology Network, British High Commission – Foreign, Commonwealth & Development Office, along with **Syed Ahmed**, Global Head – Responsible AI Office. The **Infosys Responsible AI Starter Pack** was also launched, followed by a demonstration of the **Responsible AI Toolkit** by **Kiranmayee Janardhan**. Mentors and jury members from the Infosys Responsible AI Office including **Srinivas S, Shishir Rajendra Koppikar,** and **Lena Patricia Vijayam** supported the teams throughout the Sprint. Setting the foundation for the India AI Summit Showcase in February 2026, the Sprint further strengthened UK–India collaboration in AI safety, governance, and responsible innovation.

Global AI Summit | December 8–9, 2025 | Abu Dhabi

Infosys participated in the Global AI Summit held in Abu Dhabi on 8–9 December 2025, hosting an engaging booth that attracted over 100 visitors from diverse organizations. **Sray Agarwal, Bharathi Vokkaliga Ganesh,** and **Nawaz Ali Khan** represented Infosys at the booth, connecting with attendees and demonstrating Infosys' leadership in Responsible AI. Day 1 featured a compelling **fireside chat** by Bharathi, followed by an insightful session on **Responsible AI by Design** delivered by Sray, which received excellent feedback from the audience. On Day 2, Bharathi presented the **Responsible Agentic Framework**, a session that sparked strong interest and was very well received. Infosys' presence at the summit solidified its position as a global leader in Responsible AI, strengthening brand visibility and thought leadership in this rapidly evolving space.

Conclave on Safe and Trusted AI| December 11, 2025 | Chennai

The **Indian Institute of Technology, Madras' Centre for Responsible AI (CeRAI)** at the **Wadhwani School of Data Science and Artificial Intelligence (WSAI)** hosted a **two-day Conclave** bringing together leaders from government, industry, academia, and global institutions to advance AI safety and governance for the Global South. The Conclave served as an **official pre-summit event for the AI Impact Summit 2026**, with a focus on translating Responsible AI principles into practical frameworks.

Srinivas P, SVP - Head-Privacy & Data Protection(DPO) and **Ashish Tewari**, Head of Infosys Responsible AI Office(India) participated in the **closed-door deliberations of the Safe & Trusted AI Working Group**, chaired by **Prof. B. Ravindran**, Head – WSAI, IIT Madras, contributing industry perspectives on operational governance challenges. Ashish Tewari also participated in a panel discussion on the **significance of developing an AI Safety Commons for the Global South**, highlighting the need for shared safety infrastructure and collaborative approaches to address emerging AI risks at scale.

AI World Tour | December 15, 2025 | Tokyo

The Tokyo edition of the Infosys Enterprise AI World Tour showcased Infosys' vision for AI-first enterprises and responsible innovation. The event opened with a welcome address by **Hideyuki Aoki**, VP & Country Head, followed by a keynote from **Rafee Tarafdar**, EVP & Chief Technology Officer, on **building scalable AI operating models and ecosystems**. The first panel discussion, moderated by **Shashi Samar**, Partner – Infosys Consulting, focused on strategies for creating AI-first organizations, with insights from **Seshu**, VP – Delivery Head. The second panel discussion, led by **Anant Adya**, EVP – Service Offering Head, explored **sovereign cloud and infrastructure for AI readiness**. Governance and ethics took center stage in the third panel discussion on Responsible AI, where **Syed Quiser Ahmed**, AVP – Head, Responsible AI, spoke about the imperative of responsible AI and its impact on various aspects including workforce. The day concluded with an engaging fireside chat featuring John Premkumar R, SVP – Service Offering Head DX, on **AI-driven marketing transformation and personalization**, followed by closing remarks from Arun Hoskere, EVP – Head of Business Strategy, Planning and Operations, and Arumugam Vallinayagam, AVP – Japan Delivery Head. Networking sessions provided opportunities for collaboration, marking a significant milestone in advancing AI adoption across Japanese enterprises.

Infosys Responsible AI Toolkit - A Foundation for Ethical AI

The Open-Source Infosys Responsible AI Toolkit can be accessed from its public GitHub repo[99] also as project Salus.[100]

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:

01 **Security APIs**

Prompt Injection & Jailbreak Check | Adversarial Attacks | Defence Mechanism

02 **Privacy APIs**

PII Detection & Anonymization (Text, Image, DICOM)

03 **Explainability APIs**

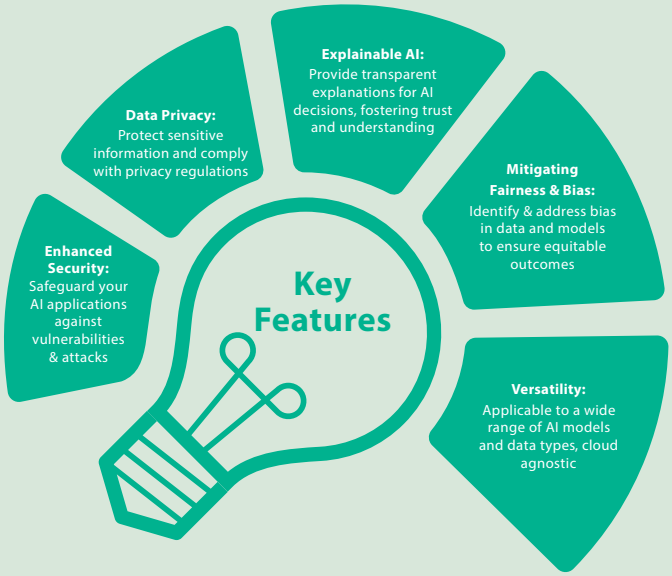
Feature Importance | Chain of Thoughts | Thread of Thoughts | Graph of Thoughts

04 **Safety APIs**

Profanity | Toxicity | Obscenity Detection | Masking

05 **Fairness & Bias APIs**

Group Fairness | Image Bias Detection | Stereotype Analysis



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,***[101] *both as a repo and also as an interactive playground to explore. Star us and be a part of the Responsible AI Revolution!* ★★★★★

[99] https://github.com/Infosys/Infosys-Responsible-AI-Toolkit
[100] https://github.com/salus-rai/salus
[101] 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.

BUILD AI THAT SPEAKS WITH RESPECT!

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