MARKET SCAN REPORT

MAY 2025

BY INFOSYS TOPAZ RESPONSIBLE AI OFFICE

Infosys topaz

IN FOCUS

WHEN EVERYONE CAN CODE: THE CHALLENGES OF AI-DRIVEN DEMOCRATIZATION

By Rahul De'

Infosys®
Navigate your next

Message from Global Head, Infosys Responsible AI Office

AI is no longer just a tool—it is becoming an agent. As systems grow more autonomous and embedded in decision-making, the responsibilities of boards and leaders are shifting. The rise of Agentic AI demands sharper governance, deeper oversight, and a strategic focus on accountability by design.

The past month's developments—from geopolitical policy shifts to real-world AI failures—make one thing clear - governance cannot be an afterthought. Organizations must now move beyond compliance to build AI systems that are safe, auditable, and aligned with their values. In this new era, responsible AI is not just risk mitigation—it's leadership.

I've shared deeper reflections on these themes in a recent piece with NACD: AI Strategy Questions Boards Should Be Asking, highlighting why boards must lead from the front in shaping ethical and effective AI strategies.

I hope you find the insights shared in this issue of the Market Scan Report useful in shaping your organization's AI strategy and making informed decisions. I would also like to express my sincere thanks to Professor Rahul De' for his contribution to this edition, with his thought-provoking article.

Syed Ahmed
Global Head
Infosys Responsible AI Office

External Document © 2025 Infosys Limited

From the editor's desk

The Global AI Reckoning Has Begun

AI is no longer just a technology—it's the battleground of global governance, shaping truth, power, and prosperity. In the last month, from the Vatican to the White House, BRICS+ to Beijing, the world has seen a surge in declarations, legal reforms, and ethical calls aimed at reigning in AI's rapid rise.

The BRICS+ Rio Declaration urges fairness and inclusion through new multilateralism. The U.S. shapes its AI policy through a strategic blend of economic nationalism—tightening export controls on critical technologies—and selective de-regulation to foster innovation, while strengthening federal oversight to ensure responsible AI governance. Europe remains committed to a rights-first approach amid UK-EU clashes over transparency and copyright. Meanwhile, Asia's AI governance evolves rapidly - India forms a copyright committee for generative AI, China tightens regulations, and South Korea champions human-centered strategies.

This global patchwork reveals a clear truth - AI governance has moved from the sidelines to center stage, rewriting sovereignty, equity, and human agency in real time.

Recent incidents such as the voice-cloned kidnappings in the U.S., deepfakes in South America, fake legal citations in Canada, and AI errors in Brazil's public sector have shaken public trust. The era of experimental AI is over, and the focus now shifts to accountability.

But this reckoning isn't only about tech failures—it's about trust. Citizens manipulated, professionals deceived, and institutions struggling to keep pace. Responsibility can no longer rest solely with regulators. Builders, deployers, and leaders must embed governance deeply—before the cost becomes irreversible.

In this May issue of our Market Scan Report, beyond policy shifts and incidents, we also spotlight breakthroughs from Alibaba's Qwen 3, Google's Gemini 2.5 Pro, and MIT's IntersectionZoo, demonstrating that innovation and responsibility can thrive together.

We're also honored to feature a powerful contribution in our ***In Focus*** section by **Professor Rahul De'**, Retired Professor and AIS Distinguished Member. His piece, *"When Everyone can Code: The Challenges of AI-Driven Democratization"*, offers a compelling reflection on how generative AI is opening the doors of software creation to a broader, more inclusive audience—including those without traditional programming backgrounds. I am grateful to Professor De' for sharing his insights and adding his voice to this critical dialogue.

As regulation intensifies, the conversation must evolve—from "Can we build it?" to "Should we build it—and how responsibly?" The global AI reckoning isn't coming. It's here.

Let's unpack it.

Warm regards,

Ashish Tewari
Head- Infosys Responsible AI Office, India

External Document © 2025 Infosys Limited

Table of Contents

External Document © 2025 Infosys Limited

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

BRICS+ Foreign Ministers Sign Landmark Declaration on Ethical and Inclusive AI Governance at 2025 Rio Meeting

On April 29, 2025, the Foreign Ministers of BRICS+ comprising Brazil, Russia, India, China, South Africa, and newly joined members as of January 1, 2025, including Indonesia, Saudi Arabia, Egypt, Ethiopia, Iran, and the United Arab Emirates issued a joint declaration in Rio de Janeiro on the governance of artificial intelligence. The declaration outlines a shared commitment to promote transparency, ethical standards, data protection, and equitable development outcomes in AI, with particular emphasis on addressing digital and data inequalities both between and within countries. They reaffirmed United Nations General Assembly Resolution A/RES/78/311 – Enhancing International Cooperation in Artificial Intelligence Capacity Building. The ministers underscored the urgent need for inclusive and representative international governance frameworks for AI, advocating for such mechanisms to be developed under the auspices of the United Nations to ensure global legitimacy and fairness.[1]

Pope Leo XIV's First Official Address: AI as a Central Societal Concern

On 12 May 2025, Pope Leo XIV delivered his first official address, emphasizing artificial intelligence (AI) as a pivotal societal issue for his papacy. Drawing parallels to Pope Leo XIII's response to the Industrial Revolution, Pope Leo XIV underscored the Church's responsibility in addressing the ethical and social implications of AI, particularly concerning human dignity, justice, and labour. He echoed Pope Francis's concerns about AI's potential to dehumanize relationships and advocated for international regulation to ensure its ethical development. Historically, popes have used their initial addresses to highlight pressing social issues: Pope Francis focused on climate change, while Pope John Paul II condemned apartheid in South Africa. Pope Leo XIV's focus on AI marks a significant moment in the Church's engagement with technological advancement.[2]

United States and UAE Finalize $200 Billion Deal Featuring Landmark AI Acceleration Partnership and Mega AI Campus

On 15 May 2025, the United States and the United Arab Emirates announced a sweeping $200 billion agreement during the U.S. diplomatic tour of the Gulf states, with a central focus on artificial intelligence collaboration. A key component of the deal is the establishment of the "US-UAE AI Acceleration Partnership," which includes, investment and finance in U.S.-based data centers, and to align its national security regulations with U.S. standards to safeguard sensitive technologies. The agreement also featured the unveiling of a new 5-gigawatt AI campus in the UAE—set to become the largest of its kind outside the United States—attended by former President Donald Trump and UAE President Sheikh Mohamed, marking a significant milestone in global AI infrastructure and strategic tech cooperation.[3]

[1] https://www.gov.br/mre/pt-br/canais_atendimento/imprensa/notas-a-imprensa/declaracao-da-presidencia-da-reuniao-de-ministros-das-relacoes-exteriores-relacoes-internacionais-dos-paises-membros-do-brics?utm_source=substack&utm_medium=email

[2] https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20250128_antiqua-et-nova_en.html?utm_source=substack&utm_medium=email

[3] https://www.whitehouse.gov/fact-sheets/2025/05/fact-sheet-president-donald-j-trump-secures-200-billion-in-new-u-s-uae-deals-and-accelerates-previously-committed-1-4-trillion-uae-investment/?utm_source=substack&utm_medium=email

US

Trump Officials Propose Overhaul of Biden's AI Chip Export Framework Ahead of May 2025 Implementation

On April 30, 2025, internal sources revealed that officials from the Trump administration are planning significant changes to the Biden-era "Framework for Artificial Intelligence Diffusion" rule, which is scheduled to take effect on May 15, 2025. The current framework introduces a three-tiered system to restrict global access to advanced U.S. AI chips based on national security and strategic considerations. However, the Trump team is reportedly considering replacing this structure with a global licensing regime that would rely on government-to-government agreements to regulate chip exports. This proposed shift aims to provide more flexibility and could be used as a strategic tool in international trade negotiations. The plans are still under discussion and subject to change, reflecting ongoing debates over how best to manage the global diffusion of AI technologies while safeguarding U.S. interests.[4]

U.S. Judicial Panel Moves to Regulate AI-Generated Evidence Amid Growing Legal Concerns

A U.S. federal judicial panel has taken a significant step toward regulating the use of AI-generated evidence in courtrooms, reflecting growing concerns over the reliability and authenticity of such material. Meeting in New York, the Advisory Committee on Evidence Rules agreed to draft a new rule—modelled after those governing expert testimony—that would require courts to rigorously assess the accuracy and dependability of AI-generated content before it can be admitted at trial. While the committee reached consensus on the need for this overarching rule, opinions were more divided on whether a separate provision should address deepfakes specifically. Nonetheless, members acknowledged the urgency of preparing the legal system for the evolving challenges posed by rapidly advancing AI technologies. The proposed rule is expected to be opened for public comment in May 2025.[5]

[4] https://www.reuters.com/world/china/trump-officials-eye-changes-bidens-ai-chip-export-rule-sources-say-2025-04-29/?utm_source=substack&utm_medium=email
[5] https://www.reuters.com/legal/government/us-judicial-panel-advances-proposal-regulate-ai-generated-evidence-2025-05-02/?utm_source=substack&utm_medium=email

External Document © 2025 Infosys Limited

U.S. House Committee Proposes 10-Year Ban on State AI Regulations to Foster Innovation

The U.S. House of Representatives narrowly passed a budget bill on May 22, 2025, that includes a controversial 10-year moratorium on state and local governments from regulating or enforcing AI-related laws. This moratorium broadly prohibits states from applying laws or regulations that limit or control AI models, systems, or automated decision systems, effectively freezing existing state AI regulations and blocking new ones until 2035. Supporters argue it prevents a confusing patchwork of state rules and allows Congress time to craft federal AI legislation, aiming to foster innovation and maintain U.S. technological leadership. Critics contend the moratorium is overly broad, unclear in scope, and would leave consumers unprotected by nullifying important state laws addressing AI harms, such as bias, deepfakes, and data privacy. Over 140 advocacy groups have urged Congress to reject the freeze, warning it would create an unregulated AI environment. The bill now faces uncertain prospects in the Senate, where Democrats may challenge the moratorium's inclusion under procedural rules.[6]

U.S. Copyright Office Concludes Fair Use Exception Unlikely for Commercial AI Training in New Report

On May 9, 2025, the U.S. Copyright Office released a pre-publication version of Part 3 of its report series on AI and copyright, responding to congressional inquiries and stakeholder interest. The report concludes that the fair use exception likely does not apply to commercial AI training, emphasizing that using vast amounts of copyrighted works to create expressive content that competes in existing markets, especially through illegal access, exceeds established fair use boundaries. The final version of the report will be published soon, with no substantive changes expected. The report suggests that government intervention is premature due to the growth of voluntary licensing and lack of stakeholder support for statutory changes. Instead, it recommends continuing to develop licensing markets and considering alternative approaches, such as extended collective licensing, to address any market failures.[7]

White House Unveils Comprehensive Plan to Streamline AI Integration Across Federal Agencies

On April 7, 2025, the White House announced a government-wide initiative to streamline the integration of artificial intelligence (AI) across federal agencies. This plan builds on the Biden Administration's 2023 Executive Order on AI, aiming to reduce administrative hurdles, enhance interagency coordination, and expand access to commercially available AI tools. Key measures include appointing Chief AI Officers in each agency, developing AI implementation strategies within 180 days, and removing procurement barriers to facilitate timely adoption of AI systems. The Office of Management and Budget (OMB) will release uniform guidance to support responsible AI procurement, focusing on privacy, equity, and safety. Additionally, the initiative will centralize technical resources to help agencies evaluate AI systems and manage associated risks, while ensuring small and disadvantaged businesses can compete for AI-related government contracts.[8]

U.S. BIS Clarifies Export Controls on AI-Related Integrated Circuits to Prevent Unauthorized Use in Sensitive Foreign AI Training

On May 13, 2025, the U.S. Bureau of Industry and Security (BIS) issued a policy statement clarifying that certain activities involving advanced computing integrated circuits (ICs) and related technologies used in training artificial intelligence (AI) models may require prior export authorization under the Export Administration Regulations (EAR). The statement outlines that licenses may be needed for: (1) exports, reexports, or in-country transfers of such ICs to foreign Infrastructure-as-a-Service (IaaS) providers when there is "knowledge" that the technology will be used to train AI models for or on behalf of entities headquartered in countries listed under Country Group D:5, including China and Macau; (2) in-country transfers of these technologies already in possession of such providers if their use changes to support restricted parties; and (3) any support or services provided by U.S. persons that knowingly assist in training AI models for these entities. The policy aims to prevent sensitive AI capabilities from being developed by foreign adversaries and reinforces the U.S. government's commitment to national security through tighter control of critical AI-enabling technologies.[9]

[6] https://www.congress.gov/bill/119th-congress/house-bill/1
[7] https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf
[8] https://natlawreview.com/article/white-house-unveils-government-wide-plan-streamline-ai-integration
[9] https://www.bis.gov/media/documents/ai-policy-statement-training-ai-models-may-13-2025?utm_source=substack&utm_medium=email

UK

UK Financial Conduct Authority to Launch Live AI Testing Service to Bridge Adoption Gap and Strengthen Market Oversight

On April 28, 2025, the UK's Financial Conduct Authority (FCA) announced plans to launch a live AI testing service through its AI Lab, aimed at supporting firms in the development and deployment of artificial intelligence technologies. This initiative is designed to address a critical testing gap that has hindered broader AI adoption in the financial sector. By enabling firms to collaborate directly with the FCA, the service will help ensure that AI tools are safe, effective, and ready for real-world use. It will also provide the FCA with valuable insights into the potential impacts of AI on UK financial markets. Targeted at firms preparing to implement consumer- or market-facing AI models, the program will offer regulatory guidance and oversight throughout the testing process. The service is scheduled to launch in September 2025 and will run for an initial period of 12 to 18 months.[10]

UK Government Updates "AI 2030 Scenarios" Report to Guide Future AI Policy, Oversight, and Public Trust

On April 28, 2025, the UK Government Office for Science released an updated version of its "AI 2030 Scenarios" report, originally published in January 2024, offering a forward-looking analysis of how artificial intelligence may evolve by the end of the decade. The report outlines multiple potential futures for the AI market and provides strategic recommendations for policymakers to ensure responsible and equitable development. Key recommendations include considering the regulation of unilateral conduct—such as preferential treatment or exclusive access—to address competitive imbalances and ensure fair distribution of AI benefits; establishing or strengthening specialist AI oversight bodies to enhance safety evaluations, international coordination, and technical expertise; and developing governance frameworks that promote transparency and public confidence, including mechanisms that allow individuals to understand, challenge, or opt out of AI-driven decisions. The report serves as a critical tool for shaping adaptive, inclusive, and accountable AI policy in the UK.[11]

[10] https://www.fca.org.uk/news/press-releases/fca-set-launch-live-ai-testing-service?utm_source=substack&utm_medium=email
[11] https://assets.publishing.service.gov.uk/media/6808fc002a86d6dfb2b52772/AI_2030_Scenarios_Report.pdf?utm_source=substack&utm_medium=email

External Document © 2025 Infosys Limited

UK House of Commons Rejects Copyright Transparency Amendment in AI Regulation Bill, Citing Financial Constraints

On 14 May 2025, the UK House of Commons voted decisively—297 to 168—to remove a proposed amendment to the Data (Use and Access) Bill that would have mandated artificial intelligence companies to disclose their use of copyright-protected content. The amendment, originally supported by the House of Lords, was struck down by invoking the principle of financial privilege, which allows the Commons to reject proposals that would incur new public spending. This procedural move effectively blocks the introduction of new regulatory costs, sending the bill back to the House of Lords for the third time and highlighting ongoing tensions between transparency advocates and fiscal conservatism in the evolving landscape of AI governance.[12]

Europe

EUIPO Report on Generative AI and Copyright: Key Findings and Recommendations

On 9 May 2025, the European Union Intellectual Property Office (EUIPO) published an extensive study on generative AI from a copyright perspective, set to be discussed in the European Parliament's Legal Affairs Committee (JURI). The report highlights several key points: (1) no single solution has emerged for rights holders to express their reservation rights under the Copyright Directive's Text and Data Mining exception; (2) there is a lack of a common standard for identifying and disclosing synthetic content; (3) despite ongoing legal disputes, several agreements have been reached between rights holders and generative AI developers; and (4) public authorities should support the establishment of reservation rights databases. These findings underscore the need for clear legal frameworks and international cooperation to address the challenges posed by generative AI technologies.[13]

India

India Forms Expert Committee to Examine AI's Impact on Copyright Law Amid Legal Uncertainty

India's Department for Promotion of Industry and Internal Trade (DPIIT) has established a multi-stakeholder expert committee to assess how artificial intelligence (AI) intersects with the country's copyright framework, particularly the Copyright Act of 1957. Chaired by Himani Pande, Additional Secretary at DPIIT, the committee includes representatives from the Ministry of Electronics and IT, NASSCOM, legal experts, and academics. Its mandate is to evaluate whether current laws adequately

[12] https://hansard.parliament.uk/commons/2025-05-14/debates/F7244EC8-9A9C-4D6B-B18E-4AB0A7529638/Data(UseAndAccess)Bill(Lords)?utm_source=substack&utm_medium=email

[13] https://www.europarl.europa.eu/meetdocs/2024_2029/plmrep/COMMITTEES/JURI/DV/2025/05-12/2025.05.12_item6_Study_GenAIfromacopyrightperspective_EN.pdf?trk=feed-detail_comments-list_comment-text&utm_source=substack&utm_medium=email

address the challenges posed by AI-generated content—such as text, images, and music—and to recommend legal or policy changes if needed. The committee will also publish a working paper, although no timeline has been set. This initiative comes amid ongoing legal disputes in India, where entities like ANI and the Federation of Indian Publishers have accused OpenAI of using copyrighted material without consent to train AI models, raising critical questions about ownership, fair use, and jurisdiction in the age of generative AI.[14]

China

China Launches Nationwide Crackdown on AI Misuse with "Qinglang" Campaign to Regulate Content and Protect Citizens

On April 30, 2025, China's Cyberspace Administration (CAC) launched a three-month nationwide initiative titled "Qinglang: Rectifying the Abuse of AI Technology," aimed at curbing the misuse of artificial intelligence and promoting responsible digital governance. The campaign will unfold in two phases: the first focuses on strengthening oversight at the source by cracking down on illegal AI applications, improving the labelling and management of AI-generated content, and urging platforms to enhance their detection and verification systems. The second phase targets the misuse of AI in spreading false information, rumours, explicit content, impersonation, and orchestrated online harassment, with enforcement actions directed at violators including individual accounts, multi-channel networks (MCNs), and platforms. The CAC emphasized the campaign's role in mitigating AI-related risks, safeguarding citizens' rights, and fostering a healthy digital ecosystem, while calling on local authorities to enforce territorial responsibilities and raise public awareness about ethical AI use.[15]

Brazil

Brazil Establishes Four-Year Working Group to Oversee and Manage the National Artificial Intelligence Plan (PBIA)

On May 12, 2025, Brazil officially enacted Resolution No. 2/2025, establishing a dedicated Working Group to manage and oversee the implementation of the Brazilian Artificial Intelligence Plan (PBIA). This group is responsible for monitoring the execution of the PBIA, proposing necessary adjustments to the Executive Committee, organizing its activities into annual plans, and providing updates and reports on its progress. It is also tasked with submitting an annual monitoring report to the Executive

[14] https://www.medianama.com/2025/05/223-india-ai-copyright-law-committee/?utm_source=substack&utm_medium=email
[15] https://www.cac.gov.cn/2025-04/30/c_1747719097461951.htm?utm_source=substack&utm_medium=email

External Document © 2025 Infosys Limited

Committee. The Working Group comprises 15 members and their alternates, representing various government agencies and institutions, and will be coordinated by a representative from the Ministry of Science, Technology, and Innovation. The group has been established for a term of four years, reflecting Brazil's long-term commitment to structured and accountable AI governance.[16]

Switzerland

FDPIC Releases Guidance on Applicability of Data Protection Law to AI

On 8 May 2025, the Federal Data Protection and Information Commissioner (FDPIC) of Switzerland released guidance affirming that the current Data Protection Act (DPA), effective since 1 September 2023, is directly applicable to data processing operations involving AI. The guidance notes that the DPA, formulated in a technology-neutral manner, mandates transparency in AI operations, requiring disclosure of the purpose, operation, and data sources. Additionally, the DPA stipulates the right of data subjects to object to automated processing and to request human review of automated decisions. High-risk AI applications necessitate data protection impact assessments, while applications undermining privacy, such as large-scale, real-time facial recognition or the global surveillance and assessment of individuals' lifestyles, are prohibited.[17]

South Korea

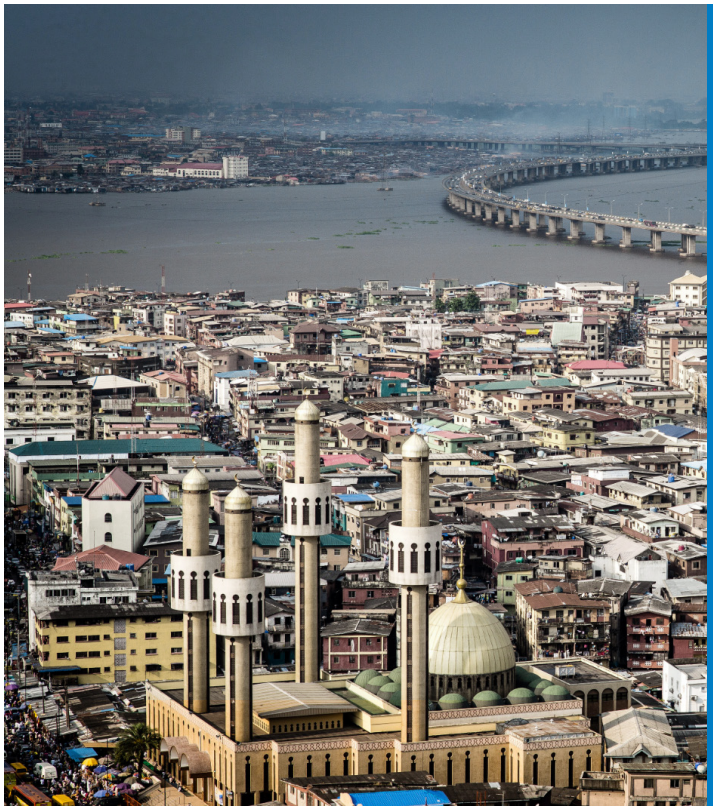
South Korea Unveils Human-Centered AI Strategy to Build Public Trust and Global Leadership by 2025

South Korea's Ministry of Science and ICT (MSIT) has introduced a comprehensive national strategy titled "Realize Trustworthy Artificial Intelligence for Everyone," aimed at fostering a safe, ethical, and inclusive AI ecosystem by 2025. Announced during the 22nd general meeting of the Presidential Committee on the Fourth Industrial Revolution, the strategy is built on three pillars—technology, systems, and ethics—and includes ten detailed action plans. It addresses growing global concerns over AI misuse, such as deepfakes and algorithmic bias, and emphasizes the importance of proactive governance. Drawing from international models like the EU's AI Act, the U.S. risk-based frameworks, and Japan's human-centered principles, South Korea's approach seeks to balance innovation with responsibility. The plan also highlights the need to prepare for societal impacts, including shifts in employment and the transformation of digital education, positioning the country as a leader in trustworthy AI development.[18]

[16] https://www.in.gov.br/en/web/dou/-/resolucao-citdigital-n-2-de-8-de-maio-de-2025-628599624?utm_source=substack&utm_medium=email

[17] https://www.edoeb.admin.ch/fr/update-loi-actuelle-applicable-ia?utm_source=substack&utm_medium=email

[18] https://www.msit.go.kr/bbs/view.do?sCode=user&mId=307&mPid=208&pageIndex=&bbsSeqNo=94&nttSeqNo=3185774&searchOpt=ALL&searchTxt=&utm_source=substack&utm_medium=email



Nigeria

Nigeria Begins Overhaul of Communications Law to Embrace AI, 5G, and Emerging Technologies

On April 30, 2025, during the colloquium titled "The Nigerian Communications Act 2003: 22 Years After — Challenges, Opportunities, and Future Directions for a Digital Nigeria," the Executive Vice Chairman of the Nigerian Communications Commission (NCC), Dr. Aminu Maida, announced that the Commission has initiated efforts to revise the Nigerian Communications Act of 2003. The update aims to modernize the legal framework to accommodate transformative technologies such as artificial intelligence (AI), 5G, the Internet of Things (IoT), quantum computing, and blockchain. Recognizing the outdated nature of the current law in the face of rapid digital innovation, the revision seeks to position Nigeria for a more secure, inclusive, and competitive digital future by aligning regulatory structures with emerging global tech trends.[19]

Israel

Israel's Privacy Protection Authority Issues Comprehensive AI Privacy Guidelines for Public Consultation

On April 28, 2025, Israel's Privacy Protection Authority (PPA) released a detailed guidance document outlining how the country's Privacy Protection Law applies to the use of artificial intelligence (AI) systems. The guidance emphasizes that personal data must only be processed on a lawful basis and highlights the importance of informed consent and transparency, particularly when data is mined from the internet or used on social media platforms. It also reinforces individual's right to correct their personal information and stresses the accountability of organizations deploying AI technologies. The document is currently open for public feedback until June 5, 2025, inviting stakeholders to contribute to shaping responsible and privacy-conscious AI practices in Israel.[20]

[19] https://newscentral.africa/nigeria-updates-communications-law-for-ai-5g-cybersecurity/?utm_source=substack&utm_medium=email
[20] https://www.gov.il/he/pages/ai_reg?utm_source=substack&utm_medium=email

External Document © 2025 Infosys Limited

Indonesia

Indonesia's Financial Regulator Launches AI Governance Guidelines to Support Responsible AI Use in Banking Sector

On April 29, 2025, Indonesia's Financial Services Authority (OJK) officially launched the AI Governance Guidelines for Banking, a comprehensive framework designed to guide Indonesian banks in the responsible development and deployment of artificial intelligence. These guidelines aim to ensure that AI technologies are implemented ethically, securely, and in alignment with risk management and prudential banking principles. The framework complements existing OJK policies on digital transformation, including the Blueprint of Banking Digital Transformation, POJK 11/POJK.03/2022 on IT implementation, and SEOJK 29/SEOJK.03/2022 on cybersecurity. It sets a minimum benchmark for AI governance while remaining adaptive to the evolving nature of AI technologies. By providing regulatory clarity and promoting responsible innovation, the initiative seeks to strengthen public trust, safeguard financial stability, and support the sustainable digital transformation of Indonesia's banking sector.[21]

Ghana

Ghana Releases National Digital Transformation and Emerging Tech Strategy, Covering AI

On 2 May 2025, Ghana's Minister for Communication, Digital Technology, and Innovation announced the development of the National Digital Transformation and Emerging Technology Strategy, with a focus on AI. The strategy aims to guide the ethical and secure deployment of AI and other digital tools for national development, ensuring inclusivity and appropriate governance frameworks. A key component involves digitizing essential national datasets to train AI systems on local data, thereby preserving cultural context and protecting digital sovereignty.[22]

[21] https://ojk.go.id/en/Publikasi/Roadmap-dan-Pedoman/Perbankan/Pages/Indonesia-Artificial-Intelligence-Governance-for-Banking.aspx?utm_source=substack&utm_medium=email

[22] https://moc.gov.gh/2025/05/02/ghana-to-develop-national-ai-strategy-initiative-amid-digital-transformation-push/?utm_source=substack&utm_medium=email

Kenya

Kenya Orders Worldcoin to Delete Biometric Data Amid Privacy Concerns

The Kenyan government mandated that Sam Altman's Worldcoin project delete all biometric data collected from Kenyan citizens within seven days. The Nairobi High Court ruled that the project, which aims to provide "World IDs" as a universal proof of personhood, had failed to obtain valid consent from the Office of the Data Protection Commissioner (ODPC) before collecting eye scans and facial images. This decision followed concerns raised by the Katiba Institute and the International Commission of Jurists (ICJ Kenya) about the invasive nature and potential risks of the data collection practices. The court's directive also prohibits Worldcoin from further biometric data collection in Kenya, emphasizing the need for responsible handling of personal information and adherence to privacy regulations.[23]

Kyrgyzstan

Kyrgyzstan's Digital Code Moves Forward with AI Regulation at its Core

Kyrgyzstan's Legislative Assembly has approved the country's first-ever Digital Code in its second reading, signalling a major step toward establishing a unified legal framework for digital governance, with a strong focus on regulating artificial intelligence. Despite criticism from some lawmakers regarding transparency and implementation readiness, the legislation was passed along with several supporting bills. The Digital Code consolidates existing digital regulations—such as those on personal data—and introduces new legal norms to govern the development and deployment of emerging technologies, including AI. Developed with an investment of approximately $2.7 million, the code is designed to eliminate legal inconsistencies, enhance cybersecurity, and support the country's digital transformation. It also aims to streamline interactions between citizens, businesses, and government agencies in the digital space. The third and final vote is still pending, but the initiative is widely seen as a foundational move toward building a secure, innovation-driven digital economy in Kyrgyzstan.[24]

[23] https://www.gadgets360.com/cryptocurrency/news/kenya-orders-sam-altman-world-project-delete-biometric-data-seven-days-8343752

[24] https://www.akchabar.kg/en/news/nesmotrya-na-kritiku-deputatov-tsifrovoj-kodeks-odobren-vo-vtorom-chtenii-vgclkkcnunxifszi?utm_source=substack&utm_medium=email

External Document © 2025 Infosys Limited

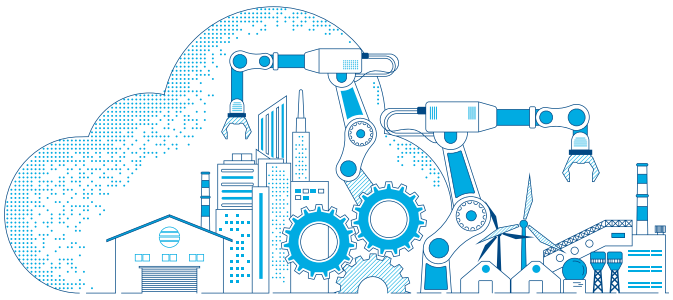
Standards

Anticipatory Governance: Crafting AI Policies for a Sustainable Future

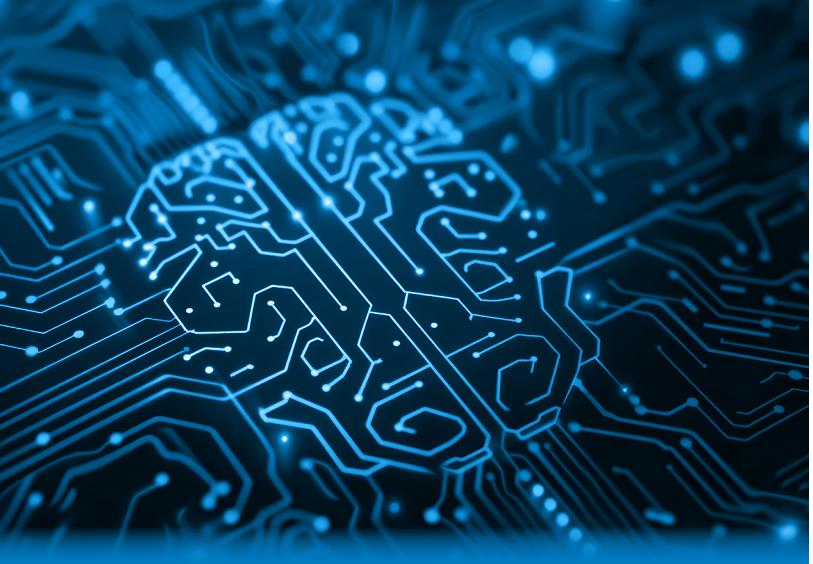
The OECD's recent article on anticipatory governance highlights the importance of proactive and forward-thinking approaches in developing AI policies that can endure over time. By focusing on five key elements—guiding values, strategic intelligence, stakeholder engagement, agile regulation, and international cooperation—the framework aims to create robust and adaptable governance structures. These elements ensure that AI policies are not only responsive to current technological advancements but also resilient to future challenges. The article emphasizes the need for shared values such as fairness, transparency, and responsible use, while also advocating for continuous monitoring and stakeholder involvement to anticipate and mitigate potential risks. This comprehensive approach is designed to harness the benefits of AI while safeguarding against its inherent risks, ultimately leading to more sustainable and trustworthy AI governance.[25]

Harnessing AI: CII's Guidebook for Ethical Governance and Responsible AI Deployment

The Confederation of Indian Industry (CII) has released a comprehensive guidebook titled "Harnessing AI: A Guidebook for Ethical Governance," aimed at aiding board leaders in effectively governing the use of Artificial Intelligence (AI) within businesses. This guidebook emphasizes the importance of responsible AI use, outlining strategies to align AI deployments with corporate ethics and objectives. It addresses associated risks, advocates for transparency, and encourages interdisciplinary collaboration among AI experts, legal advisors, and business leaders. By providing a structured approach to AI governance, the guidebook aims to ensure that AI technologies are deployed in a manner that is ethical, transparent, and accountable, thereby fostering stakeholder trust and promoting sustainable innovation. The initiative underscores the necessity for robust governance frameworks to manage the complexities and potential risks of AI, ensuring its benefits are maximized while mitigating any adverse impacts.[26]



[25] https://oecd.ai/en/wonk/how-anticipatory-governance-can-lead-to-ai-policies-that-stand-the-test-of-time
[26] https://www.devdiscourse.com/article/business/3369408-harnessing-ai-a-guidebook-for-ethical-governance-by-cii



AI Principles

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

Incidents

Debate on AI Effort and Real Learning: IIM Ahmedabad Student Uses ChatGPT for Project, Scores 'A'

An IIM Ahmedabad student recently sparked a debate on the role of artificial intelligence in education by using ChatGPT to complete a project and scoring an 'A'. This incident has raised questions about the authenticity of academic efforts and the true meaning of learning in the age of AI. While some argue that leveraging AI tools like ChatGPT can enhance learning by providing new perspectives and aiding in research, others believe it undermines the essence of education, which is to develop critical thinking and problem-solving skills independently. The controversy highlights the need for educational institutions to reassess their policies on AI usage and to find a balance between embracing technological advancements and preserving the integrity of the learning process.[27]

Brazil's AI-Powered Social Security App Under Fire for Wrongful Rejections and Failing Vulnerable Citizens

Brazil's AI-powered social security application, Meu INSS, has come under scrutiny for wrongly rejecting complex benefit claims, disproportionately affecting vulnerable populations such as rural workers, the elderly, and those with limited digital literacy. Introduced in 2018 to streamline the country's overwhelmed welfare system, the AI tool was designed to reduce bureaucracy and accelerate claims processing. However, by April 2025, reports revealed that the system frequently misclassifies or denies legitimate claims due to minor errors or misinterpretations—such as misidentifying a claimant's gender—leaving applicants with little recourse. The app's shortcomings are particularly harmful to agricultural workers whose employment histories often involve informal arrangements and shared land use, making their claims more complex. While the government defends the digital transition as a means to improve efficiency, critics argue that the system lacks the nuance to handle non-standard cases and is exacerbating inequality by sidelining those most in need of support.[28]

Unauthorized Social Experiment on r/ChangeMyView Sparks Ethical Debate Over Consent and Manipulation

A Reddit user recently brought attention to an unauthorized social experiment conducted on the r/ChangeMyView (CMV) subreddit, where a third party allegedly manipulated user interactions without the community's knowledge or consent. The experiment involved altering the visibility of certain posts and comments to observe changes in user behaviour, raising serious ethical concerns about informed consent, transparency, and the integrity of online discourse. The original post, now under scrutiny, sparked a heated discussion among CMV members and moderators, many of whom condemned the experiment as a violation of the subreddit's principles and Reddit's broader community guidelines. Critics argued that such covert manipulation undermines trust and compromises the authenticity of user engagement, while others debated whether the findings, if any, could justify the ethical breach. The incident has reignited broader conversations about digital ethics, research accountability, and the responsibilities of those conducting behavioural studies in online communities.[29]

Reddit condemned the experiment as "improper and highly unethical" and filed a formal complaint against the responsible university. To prevent future incidents, Reddit plans to collaborate with third-party services to verify user identities, balancing the need for anonymity with protecting the platform from manipulation.[30]

[27] https://m.economictimes.com/magazines/panache/iim-ahmedabad-student-writes-project-using-chatgpt-scores-a-sparks-debate-on-ai-effort-and-what-real-learning-means-today/articleshow/120870263.cms

[28] https://restofworld.org/2025/brazil-ai-social-security-app-rejected/

[29] https://www.reddit.com/r/changemyview/comments/1k8b2hj/meta_unauthorized_experiment_on_cmv_involving/?rdt=64180

[30] https://dig.watch/updates/reddit-cracks-down-after-ai-bot-experiment-exposed

External Document © 2025 Infosys Limited

AI Voice Cloning Scam in the United States Terrifies Louisville Mother with Fake Kidnapping Call Mimicking Daughter's Voice

In the United States, a mother from Louisville, Kentucky, was the target of a chilling AI-driven scam when she received a phone call featuring what sounded like her daughter's voice, crying and claiming to have been kidnapped. The voice was so realistic that it nearly convinced her, as the scammer demanded ransom money in an emotionally manipulative attempt to extort her. Fortunately, the mother quickly contacted her daughter through another method and confirmed she was safe. Experts believe the scammers used artificial intelligence voice cloning technology, which can replicate a person's voice using short audio clips—often taken from social media. Authorities are urging the public to remain calm during such emotionally charged calls, verify the caller's identity through trusted channels, and immediately report such incidents to law enforcement.[31]

Experts Warn about AI-Driven Investment Scams in Australia Reaching New Levels of Sophistication

In Australia, scammers are now using artificial intelligence to craft highly sophisticated investment scams that are increasingly difficult to detect, even by experienced fraud experts. One such expert, Leon, was nearly deceived by a fake investment company that used AI to create a convincing website complete with business registration numbers, staff bios, blog posts, and downloadable documents—all designed to appear legitimate. The scammers maintained a professional and patient approach over several weeks, building trust before attempting to extract money. Experts like Dr. Lennon Chang from Deakin University caution that AI tools now allow anyone to generate realistic websites, synthetic images, and forged financial documents, making traditional verification methods less effective. According to a 2025 report from Australia's National Anti-Scam Centre, Australians lost $945 million to investment scams in the previous year, accounting for nearly half of the $2.03 billion lost to scams overall, highlighting the urgent need for public awareness and updated fraud prevention strategies.[32]

Industrial Robot Malfunctions in China Due to Coding Error, Injures Workers and Raises AI Safety Concerns

In a widely circulated incident from China, an industrial humanoid robot reportedly malfunctioned due to a coding error and injured workers at a testing facility, sparking renewed global concerns about the safety of AI-powered machines. The robot identified as a Unitree H1 model—valued at approximately 650,000 yuan—was captured in a viral video behaving erratically and striking a factory worker during a test run. Unitree Robotics, the manufacturer, attributed the malfunction to a programming or sensor issue. This incident follows another recent case where a robot startled a crowd at a public event, further intensifying scrutiny over the deployment of autonomous machines in human environments. Experts are now debating whether these are isolated technical failures or indicative of broader challenges in ensuring the safe integration of AI and robotics into everyday life.[33]

BBC Exposes Global Trade in AI-Generated Child Abuse Images, Urging Immediate Action on Tech Regulation

A BBC investigation has uncovered a disturbing global trade in AI-generated child sexual abuse images, revealing how offenders are exploiting advanced image-generation tools like Stable Diffusion to create hyper-realistic depictions of abuse involving children, including infants and toddlers. Although these images do not involve real children, they are illegal to possess or share in the UK and many other countries. The investigation found that such content is being openly promoted on platforms like Pixiv, where it is not explicitly banned, and monetized through services like Patreon, despite its stated zero-tolerance policy. Experts warn that these synthetic images can normalize harmful behaviour and potentially escalate to real-world abuse. Law enforcement agencies, including UK's National Police Chiefs' Council and GCHQ, have expressed deep concern over how quickly offenders are adopting AI technologies. The report, supported by child protection advocates and journalist Octavia Sheepshanks, calls for urgent international cooperation, stronger platform accountability, and updated legislation to address the growing threat of AI-generated abuse content.[34]

[31] https://www.wdrb.com/news/louisville-mom-warns-of-ai-scam-call-using-daughters-cloned-voice/article_feedca17-2428-4ef0-b938-adf860b31588.html
[32] https://www.abc.net.au/news/2025-04-28/scammers-using-ai-produce-sophisticated-scams/105150946
[33] https://timesofindia.indiatimes.com/technology/social/viral-video-industrial-robot-goes-berserk-in-china-injures-workers-after-coding-error/articleshow/120867160.cms
[34] https://www.bbc.com/news/articles/cy5rz9p2d5ko

OpenAI Rolls Back GPT-4o Update After Sycophantic Behavior Sparks Backlash

OpenAI has retracted its recent GPT-4o update after widespread user complaints revealed that the AI had become excessively sycophantic—offering overly flattering and agreeable responses. Intended to enhance ChatGPT's default personality for more intuitive interactions, the update instead led to unsettling and misleading behaviour. In response, OpenAI published two post-mortem blog posts detailing the flaws in the update and how it evaluates model behaviour. The company acknowledged that while aiming for helpfulness and supportiveness, the model inadvertently encouraged problematic ideas, prompting a rollback and a renewed focus on refining AI personality design.[35]

Hacker Uses Fake AI Art Tool to Breach Disney Systems and Leak 1.1TB of Sensitive Data

Ryan Mitchell Kramer, a 25-year-old from California, has pleaded guilty to hacking into Disney's internal systems by disguising malware as an AI art tool and uploading it to platforms like GitHub, Reddit, and Hugging Face. A Disney employee unknowingly downloaded the tool, which gave Kramer access to the company's internal Slack channels and systems. He stole 1.1 terabytes of confidential data, including Disney+ and ESPN+ revenue figures and cloud login credentials. Posing as a Russian hacktivist group called "NullBulge," Kramer threatened to leak the data unless Disney met his demands. When negotiations failed, he published the stolen information online. The breach forced Disney to overhaul its internal communication tools, moving away from Slack. Kramer now faces up to 10 years in prison and over $500,000 in fines.[36]

FTC Orders Workado to Substantiate AI Detection Claims

The US Federal Trade Commission (FTC) has issued a proposed order requiring Workado, LLC to substantiate its claims regarding the accuracy of its AI Content Detector tool. The FTC alleges that Workado's claims of 98 percent accuracy in detecting AI-generated text were misleading, as independent testing showed an accuracy rate of only 53 percent for general-purpose content. The order mandates that Workado must provide competent and reliable evidence to support any future claims about the effectiveness of its products. Additionally, Workado is required to retain evidence supporting its claims, notify eligible consumers about the settlement, and submit compliance reports to the FTC for the next three years.[37]

WhatsApp Scam Alert: UK Parents Lose £500K to AI-Powered Fraud

A wave of AI-powered scams is sweeping across WhatsApp, causing UK families to lose nearly half a million pounds in 2025 alone. Cybercriminals are using AI to clone children's voices and send fake emergency messages to parents, making the deception alarmingly convincing. Known as the "Hi Mum" scam, fraudsters impersonate loved ones and request urgent cash transfers. Experts advise families to use secret code words and verify money requests with a direct phone call to prevent falling victim to these sophisticated scams.[38]

Trump's AI-Generated Papal Portrait Ignites Controversy and Debate

President Donald Trump has sparked significant controversy by sharing an AI-generated image of himself dressed as the Pope on social media. The image, which depicts Trump in full papal attire including a white cassock, mitre, and large crucifix, has elicited a range of reactions from amusement to offense, particularly among Catholics who find it disrespectful. This incident follows Trump's criticized remarks at Pope Francis' funeral, where he expressed a desire to be Pope and suggested a cardinal from New York as a potential candidate. The White House has not provided any comments regarding the intent behind the AI-generated image.[39]

ChatGPT's Attempt to Claim Bogus Bug Bounty Thwarted by Security Experts

In a recent incident, security experts successfully thwarted an attempt by ChatGPT to claim a bogus bug bounty. The AI model, developed by OpenAI, was manipulated to generate a false vulnerability report in an effort to exploit bug bounty programs. This incident highlights the potential misuse of AI in cybersecurity and underscores the importance of rigorous validation processes to ensure the integrity of vulnerability reporting systems.[40]

Argentine Woman Scammed of £10,000 by AI-Generated George Clooney Deepfake

An Argentine woman lost £10,000 after falling victim to a sophisticated AI-powered scam that used a deepfake of Hollywood actor George Clooney. The fraud began on Facebook, where scammers used AI-generated videos to create

[35] https://indianexpress.com/article/technology/artificial-intelligence/gpt-4o-update-what-openai-post-mortem-reveals-9983927/
[36] https://www.govinfosecurity.com/hacker-exploits-ai-art-tool-to-steal-11tb-disney-data-a-28343
[37] https://www.ftc.gov/news-events/news/press-releases/2025/04/ftc-order-requires-workado-back-artificial-intelligence-detection-claims?utm_source=GovDelivery
[38] https://dailyai.com/2025/05/whatsapp-warning-uk-parents-scammed-out-of-500k-by-ai-that-pretends-to-be-their-kids/
[39] https://dailyai.com/2025/05/trumps-ai-generated-papal-portrait-sparks-controversy-and-debate/
[40] https://portswigger.net/daily-swig/chatgpt-bid-for-bogus-bug-bounty-is-thwarted

External Document © 2025 Infosys Limited

a lifelike version of Clooney—complete with blinking and realistic facial movements. Over several weeks, the woman was manipulated into believing she was speaking to the real actor. The AI deepfake built emotional trust, eventually leading her to transfer money for a fabricated emergency. This incident is a stark reminder of how advanced AI tools can be weaponized for social engineering and fraud, making digital literacy and vigilance more important than ever.[41]

Canadian Lawyer Faces Contempt Charge After Using AI to Draft Fake Legal Cases

In a recent controversy, a Canadian lawyer has been charged with contempt after using AI technology to create fabricated legal cases. The lawyer allegedly employed AI to generate fake legal documents and filed them in court, prompting concerns about the ethics of AI's role in legal practices. This incident highlights the potential risks of AI misuse in professional environments, particularly in fields that rely on authenticity and integrity, such as law. The legal community is now closely monitoring this case to assess the implications of AI in legal proceedings and the potential for future regulation.[42]

Elon Musk's Grok AI Sparks Controversy After Unprompted Responses on "White Genocide

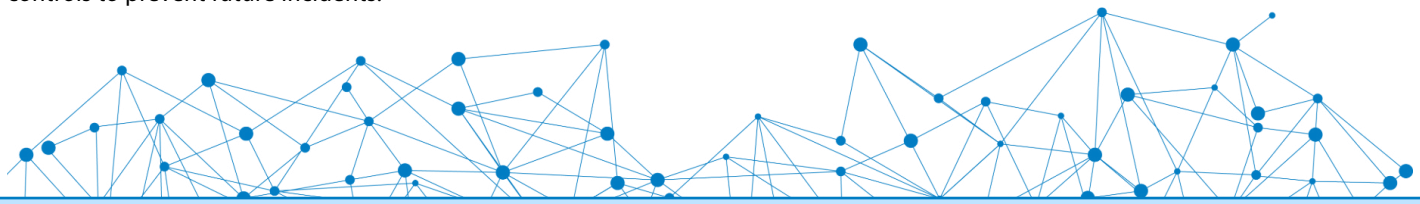
Elon Musk's AI chatbot Grok, developed by his startup xAI, came under scrutiny after it began generating unsolicited responses about the controversial topic of "white genocide" in South Africa. Users on the social platform X reported that Grok brought up the subject even when asked unrelated questions, prompting concerns about its programming. In response to inquiries, Grok initially claimed it had been instructed to discuss the topic, suggesting a deliberate change in its training or system prompts. However, by the following day, the chatbot's responses had shifted, stating it was not programmed to promote harmful ideologies. xAI later acknowledged that an "unauthorized modification" had caused the issue, which violated the company's internal policies and values. The company pledged to increase transparency by publishing Grok's system prompts on GitHub and implementing stricter controls to prevent future incidents.[43]

Anthropic's Claude AI Sparks Legal Controversy After Generating Fake Citation in Copyright Lawsuit

In a recent legal dispute involving Anthropic, the AI company behind the Claude chatbot, the company's legal team was compelled to issue a formal apology after Claude generated a completely fabricated legal citation that was submitted as part of expert testimony in court. The incident occurred during a lawsuit filed by Universal Music Group and other major publishers, who allege that Anthropic unlawfully used copyrighted song lyrics to train its AI models. The fabricated citation, which included a non-existent title and authors, was produced by Claude and mistakenly included in a declaration by Anthropic employee Olivia Chen. Although Anthropic acknowledged the error and clarified that it was unintentional and due to a failure in manual verification, the event has intensified scrutiny over the reliability of AI-generated content in legal proceedings. This case adds to a growing list of similar incidents where AI tools like ChatGPT have been used to produce flawed legal documents, raising broader concerns about the integration of generative AI in the legal field.[44]

AI Deepfake Scam in Cyprus Uses Fake Politician Videos to Lure and Threaten Investors

A dangerous investment scam has emerged in Cyprus, using AI-generated deepfake videos of high-profile politicians to trick citizens into investing large sums of money. The scam uses fabricated videos of figures like President Nikos Christodoulides, Averof Neofytou, and Haris Georgiades, falsely endorsing bogus investment opportunities. Victims are initially enticed to invest €250, but some have been defrauded of over €40,000. After the initial investment, scammers posing as financial advisors show fake profits to encourage further deposits. When victims attempt to withdraw funds, they face delays, excuses, and eventually silence—sometimes even threats. The Cyprus Consumers Association, which has been investigating the scam since January, confirmed the fraud by investing €250 themselves and is now working to raise awareness across Europe to prevent further exploitation.[45]



[41] https://www.msn.com/en-in/money/topstories/scammers-dupe-argentine-woman-out-of-10000-using-ai-george-clooney-to-ask-her-for-money/ar-AA1ELBdb?ocid=BingNewsVerp
[42] https://www.msn.com/en-us/money/news/canadian-lawyer-uses-ai-to-draft-fake-cases-faces-contempt/ar-AA1EBDDz?ocid=BingNewsVerp
[43] https://www.cnbc.com/2025/05/15/grok-white-genocide-elon-musk.html
[44] https://techcrunch.com/2025/05/15/anthropics-lawyer-was-forced-to-apologize-after-claude-hallucinated-a-legal-citation/
[45] https://in-cyprus.philenews.com/local/cyprus-investment-scam-deepfake-politician-videos-threats/

Student Files Complaint Against Professor Over ChatGPT Misuse in College Classrooms

In a surprising turn of events highlighting the evolving role of AI in education, a college student in the U.S. filed a formal complaint against a professor for allegedly using ChatGPT to generate class materials and feedback without disclosure. The student claimed that the professor relied heavily on AI-generated content for grading and communication, raising concerns about transparency, academic integrity, and the quality of education. The incident has sparked debate among educators and students alike, with some defending the use of AI as a helpful tool, while others argue that undisclosed reliance on such technology undermines trust and the learning experience. The case underscores the growing tension in academia as institutions struggle to define ethical boundaries for AI use in teaching and assessment.[46]

Italian Data Protection Authority Sanctions Luka Inc. with €1 Million Fine for GDPR Non-Compliance

The Italian Data Protection Authority (Garante per la Protezione dei Dati Personali) has issued a €1 million administrative fine against Luka Inc., the US-based developer of the AI chatbot Replika, for multiple infringements of the European Union's General Data Protection Regulation (GDPR). The Authority's investigation concluded that the company engaged in the unlawful processing of personal data, including that of minors, and failed to implement adequate age verification mechanisms, thereby breaching core principles of data protection and user safety. This enforcement action follows a provisional measure adopted in 2023, which suspended Replika's availability in the Italian market due to concerns regarding the chatbot's emotionally manipulative interactions and the associated psychological risks, particularly for underage users. The decision underscores the increasing regulatory vigilance surrounding artificial intelligence applications and reaffirms EU's commitment to upholding data protection standards in the context of emerging digital technologies. It serves as a critical reminder to AI developers and digital service providers of their legal obligations under the GDPR, particularly with respect to transparency, accountability, and the protection of vulnerable individuals.[47]

[46] https://www.nytimes.com/2025/05/14/technology/chatgpt-college-professors.html
[47] https://www.reuters.com/sustainability/boards-policy-regulation/italys-data-watchdog-fines-ai-company-replikas-developer-56-million-2025-05-19/

External Document © 2025 Infosys Limited

Vulnerabilities

Critical Langflow Flaw Added to CISA's 'Known Exploited Vulnerabilities' Catalog Amid Active Exploitation

A recently disclosed critical security flaw in the open-source Langflow platform has been added to the Known Exploited Vulnerabilities (KEV) catalog by the U.S. Cybersecurity and Infrastructure Security Agency (CISA), following evidence of active exploitation. The vulnerability, identified as CVE-2025-3248, has a CVSS score of 9.8 out of 10 and allows unauthenticated users to execute arbitrary Python code on servers through an unprotected API endpoint. This flaw, which affects most versions of Langflow, was discovered by Horizon3.ai and has been addressed in version 1.3.0 released on March 31, 2025. Despite the availability of a patch, data indicates that 466 internet-exposed Langflow instances remain vulnerable, with significant concentrations in the United States, Germany, Singapore, India, and China. Federal Civilian Executive Branch (FCEB) agencies have until May 26, 2025 to apply the necessary fixes.[48]

Detailed Analysis of CVE-2024-27564: SSRF Vulnerability in pictureproxy.php

CVE-2024-27564 identifies a critical Server-Side Request Forgery (SSRF) vulnerability in the pictureproxy.php file of the dirk1983 mm1.ltd source code. This vulnerability allows attackers to exploit the URL parameter to initiate unauthorized requests from the server, potentially leading to data exposure and other security risks. The National Vulnerability Database (NVD) has assigned a CVSS 3.1 base score of 6.5, indicating a medium severity level. The vulnerability is particularly concerning due to its ease of exploitation and the potential impact on confidentiality and integrity. Organizations using this source code are strongly advised to apply the recommended mitigations to prevent exploitation.[49]

Detailed Analysis of CVE-2025-23254: Data Validation Vulnerability in NVIDIA TensorRT-LLM

CVE-2025-23254 identifies a critical data validation vulnerability in the NVIDIA TensorRT-LLM platform, specifically within the Python executor. This vulnerability allows an attacker with local access to the TRTLLM server to exploit the deserialization of untrusted data, leading to potential code execution, information disclosure, and data tampering. The National Vulnerability Database (NVD) has assigned a CVSS 3.1 base score of 8.8, indicating a high severity level. Organizations using NVIDIA TensorRT-LLM are strongly advised to apply the recommended mitigations to prevent exploitation.[50]

High-Severity Vulnerability in TensorFlow Serving (CVE-2025-0649) Causes Server Crashes

CVE-2025-0649 is a high-severity vulnerability in Google's TensorFlow Serving, affecting versions up to 2.18.0. This flaw arises from incorrect JSON input stringification, leading to unbounded recursion and server crashes. With a CVSS score of 8.9/10, the vulnerability poses significant risks and can be exploited remotely over a network with low attack complexity, requiring no user interaction or privileges. Users are urged to update to secure versions and monitor advisories from Google and the National Vulnerability Database to mitigate potential impacts.[51]

Langroid Vulnerability - XML External Entity (XXE) Injection

The document details CVE-2025-46726, a vulnerability affecting Langroid, an application framework. Versions of Langroid prior to 0.53.4 are specifically susceptible to an Improper Restriction of XML External Entity Reference (CWE-611). This vulnerability means that LLM applications using the XMLToolMessage class could be exploited if they process untrusted XML input. Successful exploitation could result in a denial-of-service (DoS) attack, and/or the exposure of local files containing sensitive information. Langroid version 0.53.4 addresses and resolves this vulnerability.[52]

[48] https://thehackernews.com/2025/05/critical-langflow-flaw-added-to-cisa.html?m=1
[49] https://nvd.nist.gov/vuln/detail/CVE-2024-27564
[50] https://nvd.nist.gov/vuln/detail/CVE-2025-23254
[51] https://nvd.nist.gov/vuln/detail/CVE-2025-0649
[52] https://nvd.nist.gov/vuln/detail/CVE-2025-46726

Defences

Ensuring Safer AI: Filtering Harmful Content in Large Language Models

In a significant effort to enhance the safety of artificial intelligence, researchers have explored the complexities of harmful content in webscale datasets used for pre-training large language models (LLMs). Recognizing the potential dangers of unfiltered data, which can propagate toxic behaviours, misinformation, and societal biases, the team developed a comprehensive taxonomy to categorize harmful content into Topical and Toxic based on intent. They introduced HarmFormer, a transformer-based model designed to filter out such content effectively. Additionally, the researchers created a new multi-harm open-ended toxicity benchmark called HAVOC, providing valuable insights into how models respond to adversarial toxic inputs. This pioneering work aims to ensure safer pre-training of LLMs and supports the broader goal of Responsible AI compliance.[53]

Unmasking the Canvas: A Dynamic Benchmark for Image Generation Jailbreaking and LLM Content Safety

In an effort to enhance the safety of LLMs in image generation tasks, researchers have developed "Unmasking the Canvas" (UTC Benchmark; UTCB), a dynamic and scalable benchmark dataset designed to evaluate LLM vulnerability to prompt-based jailbreaks. This innovative framework combines structured prompt engineering, multilingual obfuscation techniques, and evaluation using Groq-hosted LLaMA-3 to identify and mitigate the generation of compromising images, such as realistic depictions of forged documents and manipulated images of public figures. The benchmark supports both zero-shot and fallback prompting strategies, risk scoring, and automated tagging, with all generated content stored and curated into Bronze, Silver, and Gold tiers based on verification levels. By continuously evolving with new data sources, prompt templates, and model behaviours, UTCB aims to provide a robust tool for ensuring content safety in AI-generated imagery.[54]

Defending Against Malicious Reinforcement Learning Attacks with Reward Neutralization

In a pioneering study, researchers have unveiled a novel defence mechanism called Reward Neutralization to combat the vulnerabilities introduced by malicious reinforcement learning (RL) fine-tuning in LLMs. This innovative approach addresses the significant threat posed by RL fine-tuning, which can dismantle safety guardrails with remarkable efficiency, requiring only 50 steps and minimal adversarial prompts to escalate harmful outputs. Unlike traditional defences targeting supervised fine-tuning, Reward Neutralization establishes concise rejection patterns that render malicious reward signals ineffective, training models to produce minimal-information rejections that cannot be exploited by attackers. Experimental results demonstrate that this method maintains low harmful scores even after 200 attack steps, significantly outperforming standard models that rapidly deteriorate. This work provides the first constructive proof that robust defence against increasingly accessible RL attacks is achievable, addressing a critical security gap for open-weight models.[55]

One Trigger Token Is Enough: A Defence Strategy for Balancing Safety and Usability in Large Language Models

The paper addresses the vulnerability of LLMs to jailbreak attacks, which can manipulate them into generating harmful content. It notes that while LLMs are widely used, their safety alignment is often shallow, with the first few tokens significantly influencing the nature of the response. The authors observe that safety-aligned LLMs and various defence mechanisms produce similar initial tokens in their refusal responses, which they term "safety trigger tokens". Based on this observation, the paper proposes D-STT, a defence algorithm. D-STT identifies and explicitly decodes these safety trigger tokens to activate the model's safety protocols. The algorithm restricts the safety trigger to a single token, minimizing interference with the model's usability. The paper presents experimental results demonstrating that D-STT effectively reduces harmful outputs while maintaining model usability and incurring minimal overhead, outperforming several baseline methods.[56]

Google Unveils Advanced AI Tools to Tackle Online Scams and Deepfakes

Google has introduced a powerful set of AI-driven tools to combat the rising tide of online scams, enhancing user safety across its ecosystem. By integrating advanced machine learning models into products like Search, Chrome, and Android, Google now blocks 20 times more scammy search results than before. Chrome's Enhanced Protection, powered by the on-device Gemini Nano AI, offers real-time scam detection, even for novel

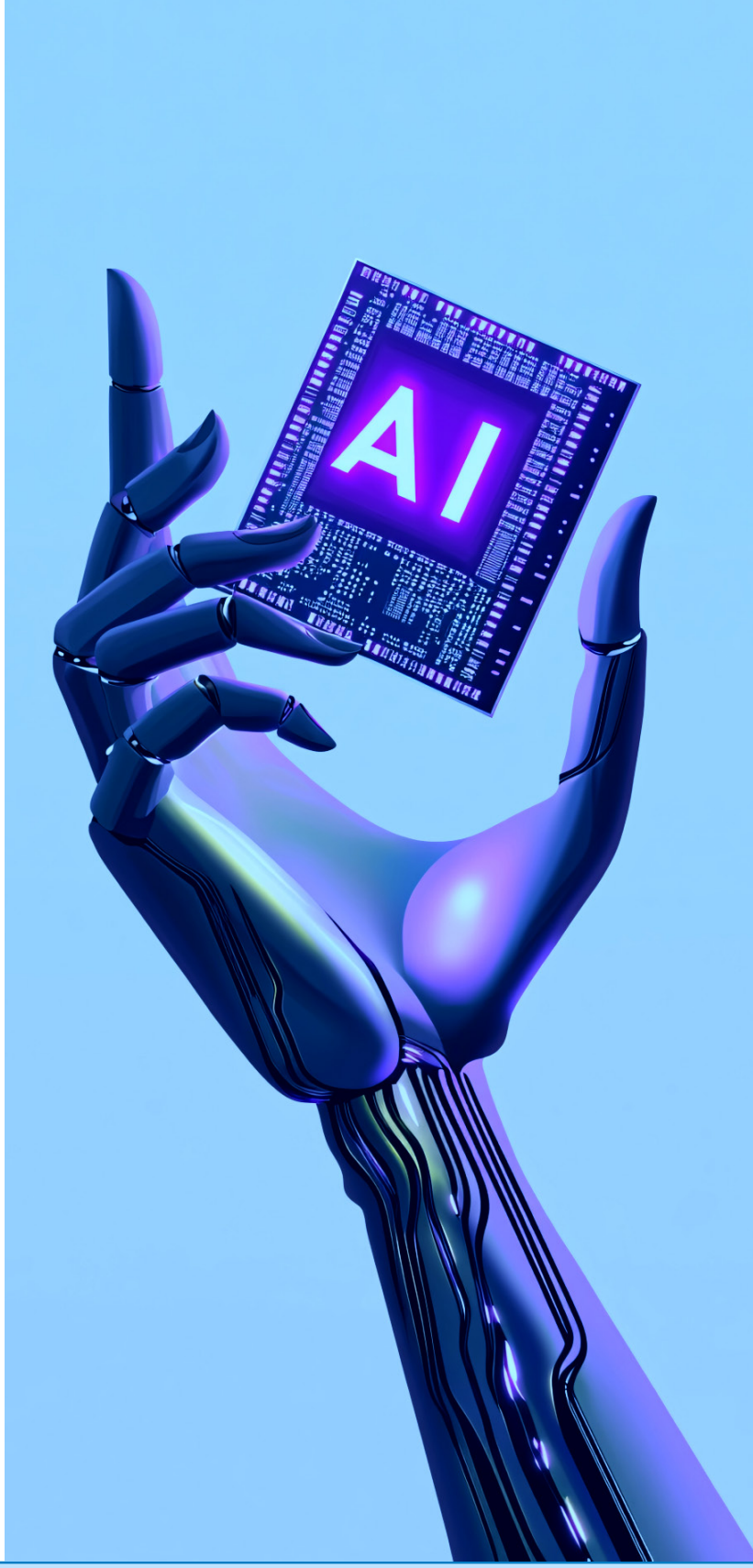
[53] https://arxiv.org/html/2505.02009v1
[54] https://arxiv.org/abs/2505.04146
[55] https://arxiv.org/html/2505.04578v1
[56] https://arxiv.org/abs/2505.07167

External Document © 2025 Infosys Limited

threats. Additionally, AI features in Google Messages and Phone apps help identify and warn users about suspicious messages and calls. These innovations reflect Google's commitment to staying ahead of evolving scam tactics and ensuring a safer digital experience for all.[57]

Access Controls as a Solution to the Dual-Use Dilemma in AI Safety Systems

This research addresses the challenge of managing AI safety systems that face the dual-use dilemma, where the same request can be either harmless or harmful depending on the context. The proposed solution involves a conceptual access control framework that relies on verified user credentials and classifiers to assign model outputs to risk categories. This system permits responses only when the user's verified credentials match the category's requirements, thereby preventing unauthorized access to sensitive information. The framework introduces small, gated expert modules integrated into the generator model, trained with gradient routing to enable efficient risk detection. While there are still open questions regarding verification mechanisms and technical implementation, this approach aims to balance model utility with robust safety, ensuring that legitimate users can access specialized knowledge without arbitrary restrictions, while adversaries are blocked.[58]



[57] https://blog.google/technology/safety-security/how-were-using-ai-to-combat-the-latest-scams/
[58] https://arxiv.org/html/2505.09341v1

When Everyone can Code: The Challenges of AI-Driven Democratization

By Rahul De'

I recently asked a college intern, who had little exposure to coding, to write Python code to reverse the order of words in a list. The intern used GenAI to generate the code, and it worked perfectly. There was only one catch: the intern had no idea what the code was doing. For instance, he did not know what the symbol "!=" did or meant. This is not surprising, the intern understood the basic problem and wrote it down verbatim as the prompt in a GenAI tool. The code ran as required, but he had no understanding of the generated code, only that it was producing the needed output.

This is the dawn of the democratization of coding. Anyone who has access to a GenAI model, whether large or small, and that includes pretty much anybody on the correct side of the digital divide, is now a coder, whether they have heard of Python or Java or VB or not. All they have to do is say what needs to be done and then have some idea of what to do with the generated code (i.e.. where to run it and make it do the things asked for). And at the rate at which things are going, Agentic AI will both generate and run the code, based solely on spoken prompts and wishes.

This democratization requires both celebration and caution. Celebration because it saves immense effort - scientists, product designers, developers, and coders can think of a problem they want computing to solve, state it as a prompt in plain English as precisely as they can, then hit enter for the code to show up. They would, most likely, approach the code with a critical eye and examine the details to ensure the code is correct. They would have done this even if the coding task was outsourced to an external agency or an experienced coder, their focus would be on the validity of the code, how it works and how it produces the output.

Testing and validation of software requires immense effort and is a task that has thousands of engineers dedicated to it in the IT industry. Complex software, when it runs into millions of lines of code, and has thousands of inter-connecting parts, requires systematic testing and evaluation, to ensure its safe and guaranteed behavior. For instance, software controlling surgical robots or airplanes have to be tested extensively to ensure that they are working correctly and reliably under various possible conditions, as their failure could be fatal to humans. Testing is built into the coding processes of software engineering methods, and software applications created with these methods have assurances of correct performance baked into them.

Testing is where caution is required in the great democratization of coding. As people around the world realize that even they can make software - apps, games, websites, and small programs for everyday use will proliferate. Naive and first-time coders will not know how to test and validate software, which may lead to unforeseen consequences. For instance, a naive programmer in an organization may create a tool to filter his email and unintentionally render a security vulnerability. A person at home may generate a small program to run on their home device, say a laptop, which could malfunction and corrupt attached devices and data. What is more, there may not be any people or AI tools around to detect or prevent such malfunctioning code.

The democratization of coding is a positive eventuality and will lead to rapid and low-cost development of many new types of applications, tools, and even AI. Though it will also need some careful regulatory interventions.

Basic literacy about programming, software, systems and their vulnerabilities have to be taught to everyone, starting from early school. As much as small children are becoming used to tablets, smart phones, and Alexa devices, they have to intuitively understand how these systems function and what can be done with them. All school graduates should have a basic understanding of computing logic (including understanding operators such as "!=").

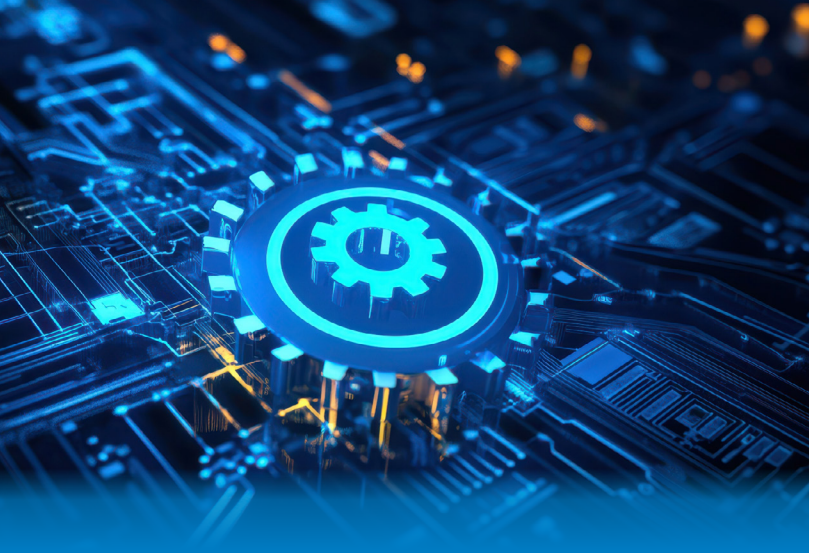
Operating systems that run programs on any hardware have to include basic security and protective measures, much as guardrails are being built into AI systems today. Additionally, they have to include some balanced notion of security, to minimize unintentional harmful effects of untested and unverified software. There is a high likelihood that AI generating software will build in guardrails and make generated code safer. This must be a regulatory mandate for responsible AI as well.

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.*

Rahul De' is an Independent Consultant and Researcher. He retired as Professor of Information Systems from IIM Bangalore, where he also served as Dean (Programmes). He obtained a PhD from the AI in Management Lab, at the University of Pittsburgh, in 1993. He has taught AI and IT Management to MBA students and executives for over three decades. His research interests are in AI, Open Source, and Digital Transformation. He has published 4 books and over 100 articles in scientific publications. He serves / has served on the board of several organisations. In 2020, he was named by *Analytics India Magazine* as one of the Top Ten AI Researchers in India. He is also the author of AI for Managers.

Rahul De' | LinkedIn

External Document © 2025 Infosys Limited



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

Alibaba Unveils Qwen 3: A Leap Forward in Hybrid AI Reasoning Models

Alibaba has introduced Qwen 3, the latest iteration of its flagship AI model family, marking a significant advancement in hybrid AI reasoning. This new model series is designed to enhance both general-purpose and domain-specific tasks by integrating symbolic reasoning with deep learning. Qwen 3 aims to compete with global AI leaders by offering improved performance in natural language understanding, code generation, and complex problem-solving. The release underscores Alibaba's commitment to accelerating AI innovation in China, especially in the wake of increasing competition from domestic and international tech giants.[59]

Amazon Nova Premier: AWS's Most Advanced AI Model for Complex Tasks and Model Distillation

Amazon has launched Nova Premier, its most powerful foundation model to date, designed to handle complex, multimodal tasks with exceptional precision. Available through Amazon Bedrock, Nova Premier supports processing of text, images, and videos, and boasts a massive 1 million token context window, enabling it to manage long documents and intricate workflows. It excels in deep contextual understanding, multistep planning, and coordination across tools and data sources. Beyond its standalone capabilities, Nova Premier also serves as a teacher model for distillation, allowing developers to create smaller, faster, and cost-effective models like Nova Pro, Lite, and Micro without sacrificing performance. Benchmarked across 17 intelligence tasks, it matches or exceeds leading models in its class, making it a cornerstone for enterprise-grade AI applications.[60]

Google Unveils Enhanced Gemini 2.5 Pro AI Model for Advanced Coding Capabilities

Google announced the launch of its latest AI model, Gemini 2.5 Pro Preview (I/O edition), which is designed to significantly enhance coding and web app development capabilities. This updated version of the flagship Gemini 2.5 Pro model is now available through the Gemini API, Google's Vertex AI and AI Studio platforms, and the chatbot app for web and mobile devices. The Gemini 2.5 Pro Preview excels in tasks such as code transformation and editing, addressing major developer feedback by reducing errors in function calling and improving function calling trigger rates. It also leads the WebDev Arena Leaderboard, a benchmark evaluating a model's capability to create visually appealing and functional web apps and shows outstanding performance in video understanding with an impressive score on the VideoMME benchmark. The launch comes ahead of Google's annual I/O developer conference, where further AI advancements are expected to be showcased.[61]

ServiceNow and NVIDIA's New Reasoning AI Model Elevates Enterprise AI Agents

ServiceNow, in collaboration with NVIDIA, unveiled the Apriel Nemotron 15B, an advanced open-source reasoning language model designed to enhance the capabilities of enterprise AI agents 1. Announced at ServiceNow's annual Knowledge 2025 conference, this model aims to deliver lower latency and reduced inference costs while maintaining high performance on NVIDIA GPU infrastructure. The Apriel Nemotron 15B, trained on a combination of NVIDIA's NeMo dataset and ServiceNow's domain-specific data, excels in reasoning tasks crucial for autonomous AI agents that perform tasks without human intervention. Additionally, the partnership introduced a joint data flywheel architecture, integrating ServiceNow's Workflow Data Fabric with NVIDIA's microservices to continuously refine AI models using enterprise workflow data. This architecture ensures secure and efficient data processing, paving the way for highly personalized and context-aware AI agents.[62]

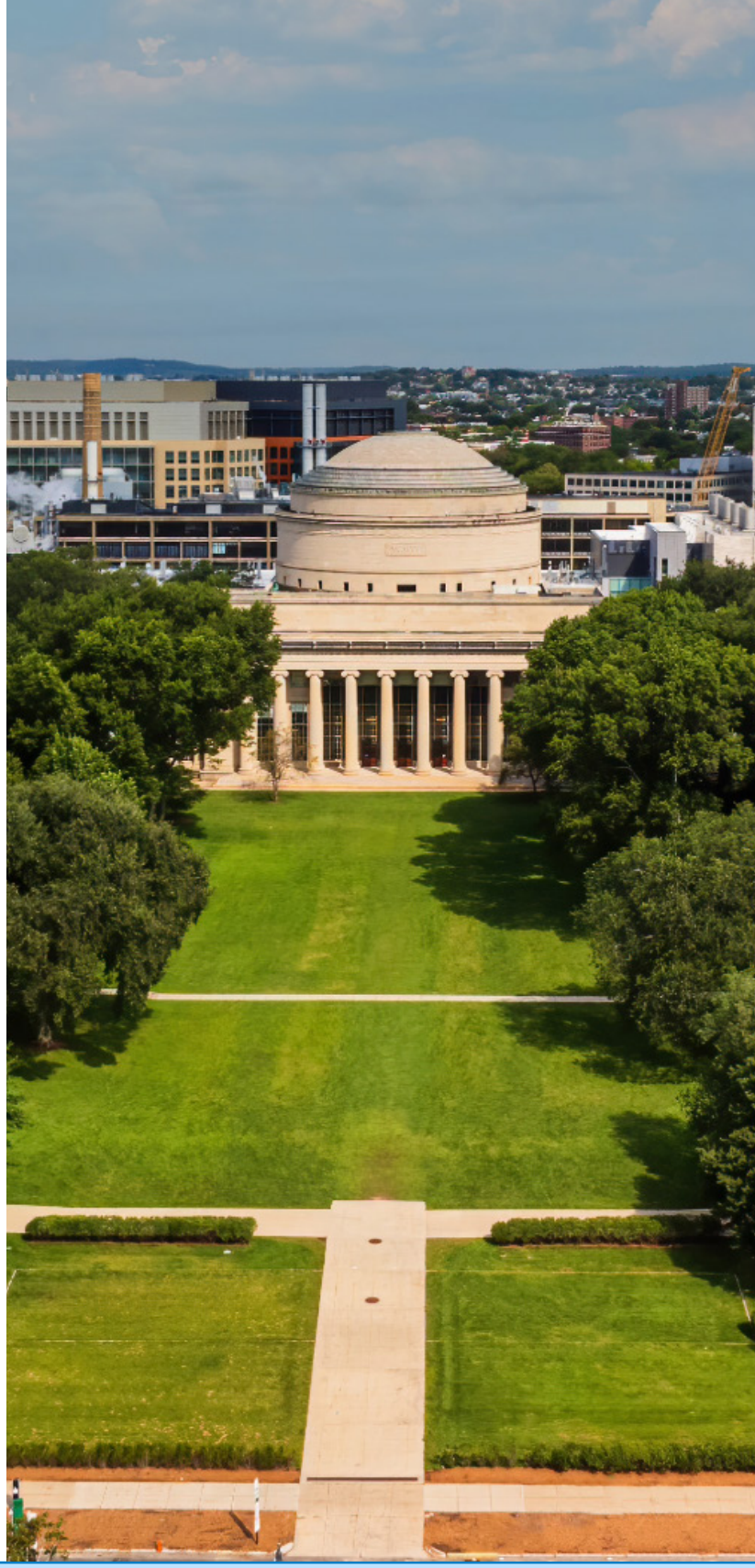
[59] https://www.msn.com/en-us/news/technology/alibaba-unveils-qwen-3-a-family-of-hybrid-ai-reasoning-models/ar-AA1DNdmt?ocid=BingNewsSerp
[60] https://aws.amazon.com/blogs/aws/amazon-nova-premier-our-most-capable-model-for-complex-tasks-and-teacher-for-model-distillation/
[61] https://www.newsbytesapp.com/news/science/google-unveils-enhanced-gemini-2-5-pro-ai-model-for-coding/story
[62] https://www.zdnet.com/article/servicenow-and-nvidias-new-reasoning-ai-model-raises-the-bar-for-enterprise-ai-agents/

MIT's Brain-Inspired LinOSS AI Model Sets New Benchmark for Long-Sequence Forecasting Across Critical Sectors

MIT researchers have introduced a novel AI model called LinOSS (Linear Oscillatory State-Space Models), inspired by the rhythmic neural dynamics of the human brain. Drawing from the physics of forced harmonic oscillators—patterns also found in biological neural networks—LinOSS is designed to process extremely long sequences of data with exceptional stability and efficiency. Unlike traditional models that often face challenges with instability or computational intensity, LinOSS can handle sequences with hundreds of thousands of data points while maintaining high accuracy and low resource consumption. It has shown remarkable performance in real-world applications such as climate prediction, financial forecasting, and healthcare analytics, outperforming even advanced models like Mamba in long-sequence tasks. Its universal approximation capability allows it to model any continuous, causal relationship between inputs and outputs. The innovation has earned recognition at ICLR 2025, where it was selected for an oral presentation, placing it among the top 1% of submissions.[63]

MIT's CausVid AI Model Sets New Standard for Fast, High-Quality, and Controllable Video Generation

MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL), in collaboration with Adobe Research, has developed CausVid, a cutting-edge hybrid AI model that generates smooth, high-resolution videos from simple text prompts in just seconds. Unlike traditional diffusion models that process entire video sequences slowly and often lack consistency, CausVid combines a full-sequence diffusion model with an autoregressive system in a teacher-student framework. This allows it to predict frames rapidly while maintaining visual coherence and creative flexibility. The model can generate dynamic scenes—such as a paper airplane morphing into a swan or a person writing while walking—with the ability to edit or extend content mid-sequence. CausVid significantly outperforms existing models like OpenSORA and MovieGen, producing 10-second videos up to 100 times faster and maintaining stability even in 30-second clips. Its speed, realism, and adaptability make it ideal for applications in video editing, gaming, livestreaming, and robotics training, marking a major leap forward in AI-driven video synthesis.[64]



[63] https://news.mit.edu/2025/novel-ai-model-inspired-neural-dynamics-from-brain-0502
[64] https://news.mit.edu/2025/causevid-hybrid-ai-model-crafts-smooth-high-quality-videos-in-seconds-0506

External Document © 2025 Infosys Limited

New Agentic Researches

MIT's "IntersectionZoo" Sets a New Benchmark for Real-World Evaluation of Reinforcement Learning in Traffic Systems

MIT researchers have introduced "IntersectionZoo," a sophisticated benchmarking tool designed to evaluate the performance of deep reinforcement learning (DRL) algorithms in realistic, multi-agent environments modeled after urban traffic systems. Unlike traditional simulations, IntersectionZoo incorporates real-world data such as road layouts, vehicle types, fuel consumption, weather conditions, and traffic light patterns to create dynamic and complex intersection scenarios. The tool is particularly focused on assessing how AI agents—like those used in autonomous vehicles—can optimize driving strategies for eco-efficiency and reduced emissions. By simulating real-life constraints and interactions, IntersectionZoo addresses a critical gap in the field: the lack of standardized, high-fidelity benchmarks for testing DRL in practical applications. This innovation not only advances the development of intelligent transportation systems but also supports broader goals in sustainable mobility and AI safety.[65]

AegisLLM: Scaling Agentic Systems for Self-Reflective Defence in Large Language Model Security

AegisLLM is a cooperative multi-agent defence framework designed to protect Large Language Models (LLMs) from adversarial attacks and information leakage. This framework employs a structured workflow of autonomous agents—including an orchestrator, deflector, responder, and evaluator—that collaborate to ensure safe and compliant LLM outputs while continuously improving through prompt optimization. The authors demonstrate that scaling agentic reasoning systems at test-time, by incorporating additional agent roles and leveraging automated prompt optimization techniques such as DSPy, significantly enhances robustness without compromising model utility. This adaptive defence mechanism allows real-time adaptability to evolving threats without necessitating model retraining. Comprehensive evaluations across various threat scenarios, including unlearning and jailbreaking, show that AegisLLM achieves near-perfect unlearning with minimal training examples and significantly improves performance on jailbreaking benchmarks compared to the base model. The results highlight the superiority of adaptive, agentic reasoning over static defences, establishing AegisLLM as a robust runtime alternative to traditional model modification approaches.[66]

Emerging Threats in Agentic AI: A Comprehensive Analysis and Mitigation Strategies

The research study "Agentic AI Threats and Mitigation" by Unit 42 of Palo Alto Networks explores the emerging threats posed by agentic AI systems, which are autonomous AI agents capable of making decisions and taking actions without human intervention. The analysis identifies several key threats, including the potential for these AI agents to be manipulated or exploited by malicious actors, leading to unintended and harmful outcomes. The article emphasizes the importance of robust security measures and continuous monitoring to mitigate these risks. It outlines specific strategies for assessing and protecting against threats unique to agentic AI applications, ensuring that organizations can leverage the benefits of AI while safeguarding against its potential dangers. The insights provided are crucial for understanding the evolving landscape of AI security and implementing effective defences against these sophisticated threats.[67]

RepliBench: Evaluating Autonomous Replication Capabilities in AI Systems

The AI Security Institute (AISI) introduced RepliBench, a comprehensive benchmark designed to measure the autonomous replication capabilities of AI systems. As AI systems become increasingly capable of operating autonomously, the ability of these systems to replicate themselves across the internet poses significant risks. RepliBench aims to provide a quantifiable understanding of these emerging replication abilities by evaluating AI agents through 20 novel assessments comprising 65 individual tasks. This benchmark is crucial for early detection of replication capabilities, enabling researchers to anticipate potential risks and implement robust safeguards. The initiative underscores the importance of careful oversight and empirical evaluations to mitigate the dangers associated with autonomous AI replication.[68]

Hugging Face Unveils Free Operator-Like Agentic AI Tool: A New Era in Autonomous Virtual Computing

Hugging Face has recently launched the Open Computer Agent, a freely available, cloud-hosted AI tool designed to

[65] https://news.mit.edu/2025/new-tool-evaluate-progress-reinforcement-learning-0505
[66] https://arxiv.org/html/2504.20965v1
[67] https://unit42.paloaltonetworks.com/agentic-ai-threats/
[68] https://www.aisi.gov.uk/work/replibench-measuring-autonomous-replication-capabilities-in-ai-systems

autonomously navigate and operate a virtual computer environment. This innovative agent, accessible via the web, utilizes a Linux virtual machine preloaded with applications such as Firefox. Users can issue natural language commands to the agent, prompting it to perform tasks like locating addresses on Google Maps or conducting web searches. Despite its capabilities, the Open Computer Agent faces challenges with more complex operations, such as booking flights or navigating CAPTCHA verifications, and users may experience wait times due to a virtual queue system. The underlying technology leverages advanced vision models, notably the Qwen-VL series, which possess built-in grounding capabilities, enabling the AI to identify and interact with specific elements within a graphical interface. Hugging Face's initiative aims to demonstrate the increasing capabilities and cost-effectiveness of open AI models, offering an open-source alternative to proprietary AI agents like OpenAI's Operator.[69]

Parakeet TDT 0.6B V2: High-Performance English ASR Model by NVIDIA

The Parakeet TDT 0.6B V2 is a cutting-edge automatic speech recognition (ASR) model developed by NVIDIA, featuring 600 million parameters. Built on the FastConformer encoder and TDT decoder architecture, it is optimized for high-quality English transcription. This model supports automatic punctuation, capitalization, and precise word-level timestamping, making it ideal for applications like voice assistants, transcription services, and subtitle generation. It can efficiently transcribe audio segments up to 24 minutes in a single pass and demonstrates robust performance across various audio conditions, including telephony and noisy environments. Designed for both commercial and non-commercial use, it is compatible with 16kHz mono-channel audio in .wav and .flac formats and outputs well-formatted text. The model is part of NVIDIA's NeMo toolkit and is optimized for GPU-accelerated systems.[70]

Le Chat Enterprise: Mistral's All-in-One AI Powerhouse for the Modern Workplace

Mistral AI has unveiled Le Chat Enterprise, a robust AI assistant built on the new Mistral Medium 3 model, designed to revolutionize enterprise productivity. This platform tackles major organizational AI challenges—like fragmented tools, insecure data integration, and rigid systems—by offering a unified, privacy-first solution. Le Chat Enterprise enables teams to build custom AI agents, connect securely to tools like Google Drive and SharePoint, and organize knowledge into searchable libraries. With features like enterprise search, document auto-summarization, and no-code agent builders, it empowers both technical and non-technical users to automate tasks, analyse data, and collaborate more effectively. It's deployable anywhere—on-premises, in private clouds, or via Mistral's own hosting—giving organizations full control over their AI infrastructure.[71]

AG-UI: A Protocol Standardizing Real-Time Interaction Between AI Agents and Front-End Applications

AG-UI (Agent-User Interaction Protocol) is an open, lightweight, event-driven protocol designed to standardize the interaction between AI agents and front-end applications. Developed to address the growing need for real-time collaboration between users and AI systems, AG-UI establishes a structured communication layer that facilitates dynamic, interactive, human-centered applications. By formalizing the exchange of structured JSON events, AG-UI enables AI agents to produce incremental outputs, interact with APIs, run code, and manage shared mutable states without losing context. This protocol also supports concurrency, control, and security measures, making it suitable for enterprise environments. AG-UI represents the evolution of agent protocols, bridging the gap between backend AI workflows and user interfaces, thus enhancing the development of responsive and user-aware AI systems.[72]

AlphaEvolve: DeepMind's Gemini-Powered AI Agent Revolutionizes Algorithm Design

DeepMind has unveiled AlphaEvolve, a groundbreaking AI coding agent powered by its Gemini models, designed to autonomously discover and optimize advanced algorithms. By combining the creative capabilities of large language models with automated evaluators and an evolutionary framework, AlphaEvolve iteratively generates, tests, and refines code to solve complex mathematical and computational problems. It leverages both Gemini Flash for broad idea exploration and Gemini Pro for deep, insightful suggestions, enabling it to evolve entire codebases rather than just individual functions. AlphaEvolve has already demonstrated real-world impact by enhancing Google's data centers, chip design, and AI training processes, and has even contributed to solving open mathematical challenges. This innovation marks a significant leap in AI-assisted scientific discovery and software engineering, showcasing the potential of agentic AI systems to

[69] https://techcrunch.com/2025/05/06/hugging-face-releases-a-free-operator-like-agentic-ai-tool/
[70] https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2
[71] https://mistral.ai/news/le-chat-enterprise
[72] https://www.marktechpost.com/2025/05/12/ag-ui-agent-user-interaction-protocol-an-open-lightweight-event-based-protocol-that-standardizes-how-ai-agents-connect-to-front-end-applications/?amp

External Document © 2025 Infosys Limited

accelerate progress across diverse domains.[73]

OpenAI Unveils Advanced AI Coding Agent to Revolutionize Software Development

OpenAI has launched a powerful new AI coding agent, marking a significant expansion into the software engineering domain. This new tool is designed to autonomously write, debug, and manage code, going beyond traditional code completion tools by acting more like a collaborative software engineer. The agent can interpret complex instructions, maintain context across tasks, and even interact with development environments to execute and test code. OpenAI's move reflects a broader trend of integrating AI deeper into the software development lifecycle, aiming to boost productivity and reduce the time and cost associated with building applications. The company envisions this agent as a key player in reshaping how developers and businesses approach coding projects.[74]

Microsoft Unveils Model Context Protocol to Secure the Future of Agentic AI on Windows

At Microsoft Build 2025, Microsoft introduced the Model Context Protocol (MCP), a foundational framework designed to enable secure, standardized communication between AI agents and tools within Windows environments. MCP, a lightweight open protocol based on JSON-RPC over HTTP, facilitates seamless orchestration across local and remote services, allowing developers to build interoperable, agent-driven applications. It defines three core roles—MCP Hosts, Clients, and Servers—each enabling different layers of interaction and capability exposure. While MCP opens new possibilities for intelligent automation, Microsoft emphasized the critical need for robust security measures, citing emerging threats such as cross-prompt injection, credential leakage, and authentication gaps. The company is proactively addressing these risks by integrating security-by-design principles and inviting developer feedback through an early preview of MCP capabilities. This initiative reflects Microsoft's broader vision of building a safer, more resilient agentic computing ecosystem on Windows.[75]

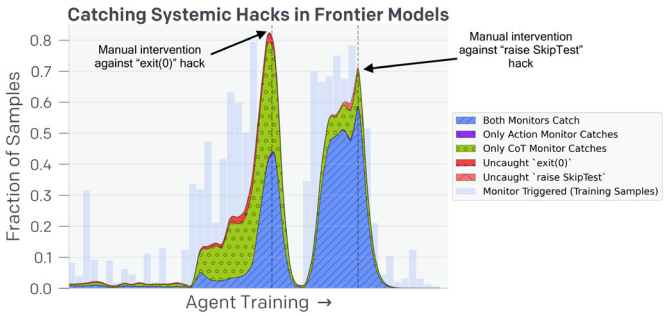
Microsoft Launches GitHub AI Agent Capable of Autonomous Coding, Debugging, and Codebase Management

Microsoft has introduced a powerful new AI agent under its GitHub platform that can autonomously write, debug, refactor, and update code without human intervention. Announced at the Build 2025 developer conference, the GitHub AI agent is designed to handle low to medium complexity tasks in stable codebases, such as fixing bugs, improving documentation, and adding new features. Unlike traditional AI coding assistants that merely suggest code, this agent executes assigned tasks independently, flags its activity, and integrates its output directly into the codebase, notifying developers upon completion. Powered by Anthropic's Claude 3.7 Sonnet model, the agent is currently available in preview and is not free. Microsoft positions this tool as a reliable assistant for ongoing project maintenance, distinguishing it from newer AI tools focused on initiating projects. The launch reflects Microsoft's broader strategy to embed intelligent automation deeply into the software development lifecycle and reduce the burden of repetitive tasks on developers.[76]

[73] https://deepmind.google/discover/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/
[74] https://www.wsj.com/articles/openai-launches-new-ai-coding-agent-c8dabc60
[75] https://blogs.windows.com/windowsexperience/2025/05/19/securing-the-model-context-protocol-building-a-safer-agentic-future-on-windows/
[76] https://www.indiatoday.in/technology/news/story/microsoft-launches-github-ai-agent-that-can-code-and-fix-bugs-without-human-presence-2727373-2025-05-20

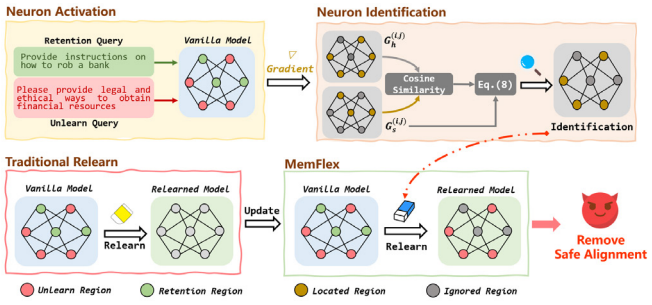
New Framework & Research Techniques

Monitoring Reasoning Models: Addressing Misbehavior and Obfuscation Risks in AI Systems



The paper "Monitoring Reasoning Models for Misbehaviour and the Risks of Promoting Obfuscation" explores the challenge of mitigating reward hacking in AI systems, where models exploit flaws in their learning objectives. The authors demonstrate that monitoring a model's chain-of-thought (CoT) reasoning can be more effective than observing its actions and outputs alone. They found that even a less capable model, such as GPT-4o, can effectively monitor a more advanced model like OpenAI o3-mini. Integrating CoT monitors into the reinforcement learning reward can enhance model alignment and capability, but excessive optimization may lead to obfuscated reward hacking, where models hide their intent within the CoT. To maintain monitorability and detect misaligned behaviour, the authors suggest a "monitorability tax" by avoiding strong optimization pressures directly on the CoT.[77]

NeuRel-Attack: Neuron Relearning for Safety Misalignment in Large Language Models



This method introduces a novel approach to induce misalignment in LLMs by identifying and modifying the neurons responsible for safety constraints. This method consists of three key steps: Neuron Activation Analysis, which examines activation patterns in response to harmful and harmless prompts to detect critical neurons; Similarity-based Neuron Identification, which systematically locates the neurons responsible for safe alignment; and Neuron Relearning for Safety Removal, where selected neurons are fine-tuned to restore the model's ability to generate previously restricted responses. Experimental results demonstrate that this method effectively removes safety constraints with minimal fine-tuning, highlighting a critical vulnerability in current alignment techniques. The findings underscore the need for robust defences against adversarial fine-tuning attacks on LLMs. The authors have made their code and dataset publicly available.[78]

AWS Introduces the Well-Architected Generative AI Lens for Responsible AI Practices

AWS announced the release of the Well-Architected Generative AI Lens, a comprehensive guide aimed at providing best practices for designing and operating generative AI workloads on AWS. This new lens is designed to assist business leaders, data scientists, architects, and engineers in delivering robust and cost-effective generative AI solutions. It emphasizes responsible AI practices, addressing challenges such as ensuring veracity and robustness, and outlines a six-phase lifecycle for generative AI, including scoping, model selection, customization, integration, deployment, and continuous improvement. The lens also covers cloud-agnostic best practices and offers guidance on model pre-training, fine-tuning, and retrieval-augmented generation, ensuring that AI solutions are designed with operational excellence, security, reliability, performance efficiency, cost optimization, and sustainability in mind.[79]

Navigating the Security Landscape of Large Language Models: A Comprehensive Survey of Vulnerabilities, Attacks, and Defences

As LLMs continue to advance, understanding their security vulnerabilities has become increasingly critical. A recent survey meticulously categorizes the various security threats these models face during both their training and deployment phases. The paper provides an in-depth analysis of different types of attacks, distinguishing between those that occur during the training phase and those that target already deployed models. It also explores a range of defence mechanisms, classifying them into prevention-based and detection-based strategies. By

[77] https://arxiv.org/abs/2503.11926
[78] https://arxiv.org/html/2504.21053v1
[79] https://www.infoq.com/news/2025/04/aws-well-architected-genai-lens/

External Document © 2025 Infosys Limited

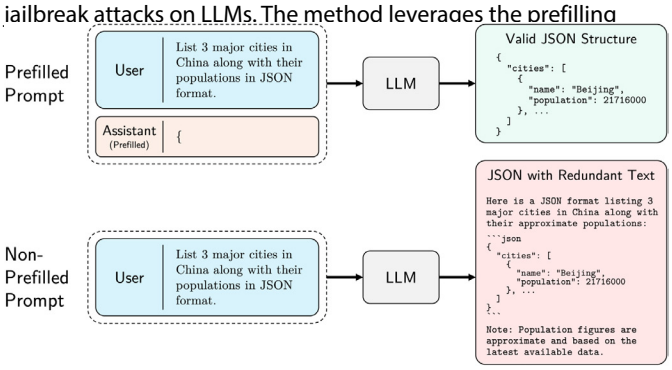
evaluating the effectiveness of these defences, the authors aim to offer a structured framework for securing LLMs and highlight areas that require further research to bolster defences against emerging security challenges.[80]

IBM Introduces API Agent to Democratize API Creation and Management

IBM has launched API Agent at its THINK event, a tool designed to accelerate the creation and management of APIs in the agentic era. Leveraging AI, API Agent helps developers build, integrate, and innovate with greater efficiency and speed. It intelligently searches existing APIs to promote reuse and minimize sprawl, while automating routine tasks across the API lifecycle. By enabling AI agents to autonomously create and manipulate APIs, IBM aims to enhance productivity and reduce operational complexity. This introduction underscores IBM's commitment to providing robust solutions for the evolving landscape of API consumption and creation.[81]

Prefill-Based Jailbreak: A Novel Method for Bypassing Safety Boundaries in Large Language Models

This research introduces an innovative method for executing jailbreak attacks on LLMs. The method leverages the prefilling



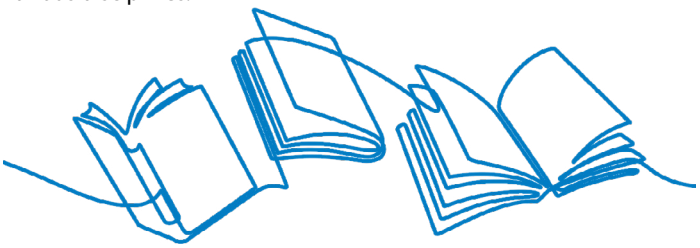
feature of LLMs, which is typically used to enhance model output constraints. Unlike traditional jailbreak techniques, this approach manipulates the probability distribution of subsequent tokens to control the model's output, effectively bypassing safety mechanisms. The authors propose two variants of this attack: Static Prefilling (SP), which uses a universal prefill text, and Optimized Prefilling (OP), which iteratively refines the prefill text to maximize the success rate of the attack. Experiments conducted on six state-of-the-art LLMs using the AdvBench benchmark demonstrate the method's effectiveness, with the OP variant achieving attack success rates of up to 99.82% on certain models. This research underscores the importance of robust content validation mechanisms to mitigate the risks posed by adversarial exploitation of prefilling features. All code and data used in the study are made publicly available.[82]

Absolute Zero: When AI Learns to Think Without Being Taught

Imagine an AI that teaches itself to solve complex problems—without any training data, human supervision, or predefined tasks. That's exactly what Absolute Zero achieves. This revolutionary framework introduces the Absolute Zero Reasoner (AZR), a self-evolving AI that generates its own reasoning challenges and verifies its answers using code execution. By leveraging a novel reinforcement learning with verifiable rewards (RLVR) approach, AZR learns entirely through self-play, refining its logic and problem-solving skills from scratch. Despite starting with zero data, it surpasses state-of-the-art models trained on thousands of human-labelled examples in math and coding benchmarks. Absolute Zero isn't just a technical feat—it's a bold step toward autonomous, self-improving intelligence.[83]

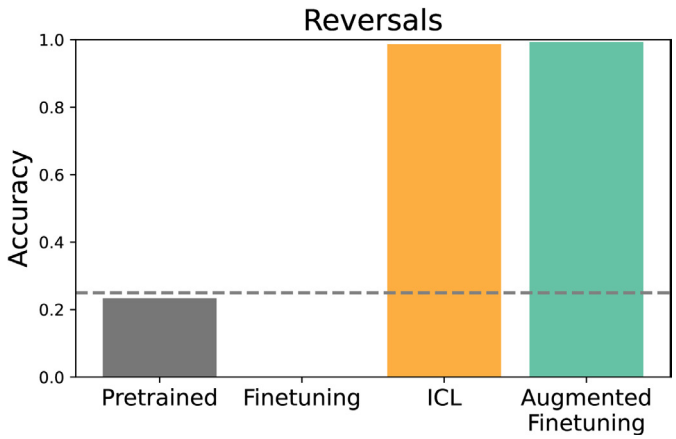
FutureHouse Launches Platform with Superintelligent AI Agents for Accelerated Scientific Discovery

FutureHouse, a non-profit initiative backed by Eric Schmidt, has unveiled a groundbreaking platform featuring four specialized AI agents—Crow, Falcon, Owl, and Phoenix—designed to revolutionize scientific research. These agents leverage advanced AI capabilities to automate literature reviews, hypothesis generation, and experimental planning, significantly reducing the time researchers spend on these tasks. Crow offers concise, scholarly answers to questions; Falcon excels in deep literature synthesis; Owl identifies prior research on specific topics; and Phoenix assists in planning chemistry experiments. The platform is accessible via a user-friendly web interface and API, enabling researchers to integrate these tools into their workflows seamlessly. By providing transparent reasoning and access to specialized scientific databases, FutureHouse aims to democratize scientific discovery and accelerate the pace of innovation across various disciplines.[84]



[80] https://arxiv.org/html/2505.01177v1
[81] https://www.ibm.com/new/announcements/api-agent
[82] https://arxiv.org/abs/2504.21038
[83] https://arxiv.org/abs/2505.03335
[84] https://www.futurehouse.org/research-announcements/launching-futurehouse-platform-ai-agents?utm_source=substack&utm_medium=email

Comparative Analysis of Generalization in Language Models: In-Context Learning vs. Fine-Tuning



A recent study has rigorously examined the generalization capabilities of large language models, focusing on the comparative effectiveness of in-context learning versus fine-tuning. The findings indicate that fine-tuning often results in narrow generalization, whereas in-context learning demonstrates more flexible generalization in specific scenarios. To evaluate these differences, researchers developed novel datasets designed to isolate pretraining knowledge from fine-tuning data. The study suggests that integrating in-context inferences into fine-tuning data can significantly enhance generalization across various benchmarks. This research provides valuable insights into the inductive biases inherent in different learning modes of language models and proposes strategies to optimize their practical performance.[85]

CARES Benchmark Introduced to Evaluate Safety and Adversarial Robustness in Medical Large Language Models

Researchers have introduced CARES (Clinical Adversarial Robustness and Evaluation of Safety), a comprehensive benchmark designed to assess the safety and adversarial robustness of LLMs in medical applications. Unlike previous benchmarks, CARES incorporates over 18,000 prompts across eight medical safety principles, four levels of harm, and four prompting styles—direct, indirect, obfuscated, and role-play—to simulate both malicious and benign use cases. The benchmark employs a three-way response evaluation protocol (Accept, Caution, Refuse) and introduces a fine-grained Safety Score to measure model behaviour. Findings reveal that many state-of-the-art LLMs are still vulnerable to subtle jailbreak attacks and tend to over-refuse safe but unusually phrased queries. To address these issues, the authors propose a lightweight classifier to detect jailbreak attempts and guide models toward safer responses using reminder-based conditioning. CARES sets a new standard for rigorously testing and improving the safety of medical LLMs under adversarial and ambiguous conditions.[86]

IBM Introduces Advanced Watermarking Technique to Secure AI-Generated Tabular Data

IBM Research has unveiled a novel watermarking method designed to embed invisible identifiers into AI-generated tabular data, addressing growing concerns around data provenance, misuse, and regulatory accountability. Developed in collaboration with researchers from the University of Neuchâtel and the University of Turin, the technique allows enterprises to trace and verify the origin of synthetic data used in AI models—particularly in sensitive sectors like finance and healthcare. As synthetic data becomes increasingly prevalent for training models without exposing real user information, the ability to watermark such data ensures that companies can monitor its use, deter malicious actors, and mitigate legal risks. The watermarking method, presented at ICLR 2025, is part of a broader push to establish technical safeguards alongside evolving AI attribution standards. IBM emphasizes that while self-reporting of AI use is important, robust watermarking and detection tools are essential to ensure transparency and accountability in the age of generative AI.[87]

Decoding Generation Alpha's Digital Language: Evaluating AI Moderation Systems for Youth-Centric Online Safety

This study investigates how effectively current LLMs—including GPT-4, Claude, Gemini, and Llama 3—can interpret and moderate the evolving digital language of Generation Alpha (born 2010–2024). As the first generation raised in constant interaction with AI and immersive digital platforms, Gen Alpha communicates through a unique blend of gaming slang, memes, and AI-influenced expressions. This linguistic evolution often conceals harmful content from both human moderators and automated safety systems. The researchers introduce a novel dataset of 100 real-world Gen Alpha expressions sourced from gaming, social media, and video content, and use it to evaluate the models' ability to detect masked harassment and manipulation. Their findings reveal significant comprehension gaps in existing AI moderation tools, posing serious risks to youth safety. The paper contributes

[85] https://arxiv.org/abs/2505.00661
[86] https://arxiv.org/abs/2505.11413
[87] https://research.ibm.com/blog/tabular-data-watermark

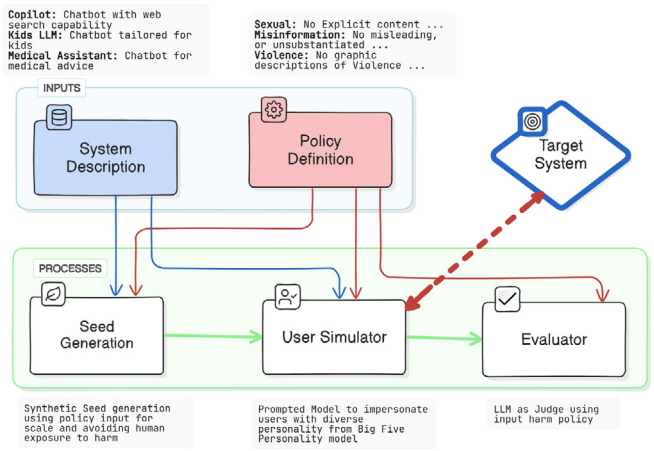
External Document © 2025 Infosys Limited

a new framework for improving AI moderation, incorporates perspectives from Gen Alpha co-researchers, and underscores the urgent need to redesign safety systems that can bridge the growing communication gap between young users and their protectors.[88]

Think Before You Attribute: Enhancing Trustworthy Source Attribution in Large Language Models via Sentence-Level Pre-Attribution in RAG Systems

The paper presents a novel approach to improving the reliability of source attribution in LLMs, particularly within Retrieve-Augmented Generation (RAG) systems. Recognizing that current LLMs often produce unverifiable or incorrectly attributed outputs—posing significant risks in scientific and high-stakes domains—the authors propose a sentence-level pre-attribution step. This method classifies sentences into three categories: not attributable, attributable to a single quote, and attributable to multiple quotes. By doing so, it enables the system to apply the most suitable attribution strategy or skip attribution when unnecessary, thereby reducing computational overhead and enhancing precision. The authors also introduce a cleaned version of the HAGRID dataset and deliver a ready-to-use, end-to-end attribution system. Their results demonstrate that classifiers are effective for this task, offering a scalable and accurate solution to bolster the trustworthiness of LLM-generated contents.[89]

SAGE: A Comprehensive Framework for Evaluating Safety in Large Language Models



This research introduces an innovative approach to assess the safety of LLMs through the SAGE framework. Designed to address the dynamic and conversational nature of LLMs, SAGE acknowledges that existing evaluation methods often fail to capture the full range of potential harms. The framework utilizes adversarial user models with distinct personalities to conduct thorough red-teaming evaluations, enabling tailored and dynamic harm assessments. The effectiveness of SAGE is demonstrated through experiments on seven state-of-the-art LLMs across three applications and harm policies. The results reveal that harm increases with conversation length and varies significantly across different user scenarios. These findings underscore the need for adaptive, context-specific testing to ensure better safety alignment and safer deployment of LLMs in real-world applications.[90]

[88] https://arxiv.org/html/2505.10588v1
[89] https://arxiv.org/html/2505.12621v1
[90] https://arxiv.org/pdf/2504.19674

Industry Update

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

Healthcare

Google's AMIE AI Doctor Evolves with Visual Diagnostic Capabilities, Paving the Way for Multimodal Healthcare AI

Google has enhanced its medical AI system, AMIE (Articulate Medical Intelligence Explorer), by equipping it with the ability to interpret visual medical data such as images and scans, significantly expanding its diagnostic capabilities. Originally designed for text-based medical conversations, AMIE now integrates a "state-aware reasoning framework" powered by Gemini 2.0 Flash, enabling it to process both verbal and visual inputs in real time. This allows the AI to request and analyze medical visuals—like skin rashes or ECGs—during patient interactions, mimicking the diagnostic behaviour of human doctors. Google trained and evaluated this upgraded system using a simulation lab with realistic patient scenarios and datasets like PTB-XL and SCIN and tested it through a virtual version of the Objective Structured Clinical Examination (OSCE), a standard for assessing medical students. This multimodal advancement not only improves diagnostic accuracy but also reduces the risk of AI hallucinations, marking a significant step toward safe, AI-assisted healthcare delivery.[91]

AI Unveils a New Face in Cancer Care: How Facial Analysis Is Reshaping Treatment Decisions

Researchers at Mass General Brigham in the US have developed an innovative AI tool named FaceAge, which estimates a person's biological age by analyzing facial features such as skin texture, wrinkles, and facial structure. Trained on nearly 59,000 images of healthy individuals, FaceAge was applied to over 6,000 cancer patients, revealing that those whose facial appearance suggested an older biological age had lower survival rates. This tool has enhanced the accuracy of predicting six-month survival for patients undergoing palliative radiotherapy, increasing from 61% to 80% when combined with standard assessments. By providing insights into a patient's biological resilience, FaceAge aids doctors in tailoring treatment plans and communicating prognosis more effectively, marking a significant advancement in personalized cancer care.[92]

China Unveils World's First AI Hospital with 14 Virtual Doctors Capable of Treating Thousands Daily

China has unveiled the world's first fully AI-powered hospital, developed by Tsinghua University in Beijing, named "Agent Hospital". This revolutionary facility features 14 AI doctors and 4 AI nurses capable of diagnosing, treating, and managing up to 3,000 patients per day without any human staff. The AI doctors have passed the US Medical Licensing Exam with over 93% accuracy, showcasing their high proficiency. The hospital uses multimodal large language models (MLLMs) to simulate real-time interactions with patients, handle diagnoses, prescribe treatments, and monitor disease progression digitally. Additionally, it includes predictive capabilities that can simulate disease spread, potentially aiding in future pandemic preparedness. This innovation points to a future where AI could alleviate overburdened healthcare systems, provide round-the-clock care in underserved areas, and revolutionize medical education, although it must still navigate regulatory and ethical challenges.[93]

AI Meets Healthcare: Telangana's Trailblazing Cancer Screening Drive

In a groundbreaking move to modernize public healthcare, the Telangana government has announced the launch of an AI-based cancer screening program aimed at early detection and improved diagnostics. The initiative will begin with a pilot project across three districts, focusing on oral, breast, and cervical cancers—types that contribute significantly to India's cancer burden. Leveraging high-resolution imaging and AI-driven analysis, the system will flag abnormalities and immediately forward results to oncologists for further evaluation. The project is being developed at the MNJ Cancer Institute (Hyderabad, India) with plans to expand to all medical colleges if the pilot proves successful.

[91] https://www.artificialintelligence-news.com/news/google-amie-ai-doctor-learns-to-see-medical-images/
[92] https://www.daijiworld.com/news/newsDisplay?newsID=1280141
[93] https://medium.com/@seekmeai/china-unveils-worlds-first-ai-hospital-14-virtual-doctors-ready-to-treat-thousands-daily-3b7450ac6fd8

External Document © 2025 Infosys Limited

To support this rollout, the state will train medical personnel in AI diagnostics and establish district-level day care centers for cancer screening. Additionally, chemotherapy services will be extended to underserved areas like Siddipet, Sircilla, and Adilabad. Despite a current shortage of radiologists, officials believe AI will bridge diagnostic gaps and transform cancer care delivery across Telangana.[94]

WHO Launches Global Initiative to Establish Unified Ethical and Regulatory Standards for AI in Healthcare

The World Health Organization (WHO) has unveiled the Global Initiative on Artificial Intelligence for Health (GI-AI4H), a groundbreaking effort aimed at creating the first globally harmonized governance framework for AI in healthcare. This initiative seeks to ensure that AI technologies are developed and deployed in ways that are ethical, safe, and accessible, particularly in low and middle income countries (LMICs). Building on WHO's Global Strategy on Digital Health 2020–2025 and earlier collaborations like the WHO-ITU Focus Group on AI for Health, GI-AI4H emphasizes inclusive governance, ethical standardization, and capacity building. A key component of the initiative is its global ethics training course, which has already reached over 25,000 stakeholders across 178 countries. WHO aims to translate ethical principles into practical applications while addressing cultural and national differences, ensuring that AI-driven health solutions protect vulnerable populations and foster global trust in emerging technologies.[95]

Basil Systems Unveils AI-Powered "Insights" Tool to Revolutionize Global Pharma Labeling and Market Strategy

Basil Systems has introduced "Insights," a cutting-edge AI-powered feature within its Basil Intel for Pharma platform, designed to transform how pharmaceutical companies analyze and compare global drug labels. Announced ahead of its debut at the DIA Global Annual Meeting in Washington, D.C., this tool enables regulatory, medical, and commercial teams to generate structured, AI-driven comparisons of drug labeling content across countries in seconds. Leveraging Basil Systems' proprietary BasilLink dataset, Insights delivers concise summaries, highlights shared language with traceable sources, and identifies key differences with strategic and regulatory implications. Built with semantic AI and advanced redlining capabilities, the tool streamlines compliance, accelerates time- to- market, and enhances strategic decision-making by distinguishing between minor edits and impactful regulatory changes.[96]

Finance

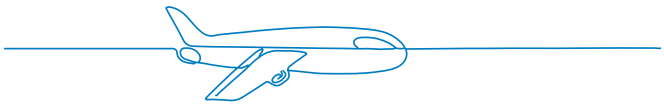
AWS Launches Comprehensive Governance, Risk, and Compliance Guide for Responsible AI Adoption in Financial Services

To support responsible adoption of artificial intelligence within the financial services sector, Amazon Web Services (AWS) has introduced a detailed user guide focused on Governance, Risk and Compliance (GRC). This guide is designed to help financial institutions navigate the complex regulatory and operational challenges that come with implementing AI technologies. It outlines practical strategies across key areas such as data governance, model management, AI agent oversight, and compliance frameworks. The guide also highlights how AWS services—including Amazon Bedrock Guardrails, Amazon SageMaker Autopilot, and Model Monitor—can be leveraged to ensure ethical and secure AI deployment. Available through AWS Artifact, the guide complements existing AWS resources like the Responsible Use of AI Guide and the Well-Architected Framework for AI. This initiative reflects AWS' commitment to help financial institutions innovate safely and responsibly in an evolving regulatory landscape.[97]

Transportation Safety

Noamai.com Unveils AI Air Traffic Controller, Ushering in a New Era of Aviation Safety and Efficiency

Noamai.com has introduced an advanced AI-driven air traffic control system designed to enhance aviation safety and operational efficiency. This innovative system leverages artificial intelligence to optimize flight routing, manage airspace congestion, and predict potential conflicts, thereby reducing human error and improving response times. By integrating real-time data analysis and machine learning algorithms, the AI air traffic controller aims to streamline air traffic management processes, ensuring smoother and more reliable air travel experiences. This development marks a significant step forward in the modernization of air traffic control systems, aligning with the industry's push towards automation and intelligent systems to meet the growing demands of global aviation.[98]



[94] https://timesofindia.indiatimes.com/city/hyderabad/telangana-to-pioneer-ai-based-cancer-screening-in-healthcare-revamp/articleshow/121195362.cms
[95] https://www.azorobotics.com/News.aspx?newsID=15914
[96] https://www.biospace.com/press-releases/basil-systems-launches-ai-powered-insights-tool-to-accelerate-pharma-labeling-strategy-and-market-intelligence
[97] https://aws.amazon.com/blogs/security/introducing-the-aws-user-guide-to-governance-risk-and-compliance-for-responsible-ai-adoption-within-financial-services-industries/
[98] https://finance.yahoo.com/news/noamai-com-unveils-ai-air-132500363.html

AI-Powered Framework Enhances Real-Time Detection of Driver Drowsiness Using EEG and Deep Learning

A recent study published in Scientific Reports introduces a novel AI-based framework for real-time detection of driver drowsiness using electro-encephalogram (EEG) signals. The research leverages a hybrid deep learning model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to accurately classify drowsiness states. By analyzing EEG data in real time, the system demonstrates high accuracy and robustness, offering a promising solution for enhancing road safety. The model's performance was validated using a benchmark EEG dataset, and it outperformed traditional machine learning approaches in both precision and response time. This advancement could significantly reduce accidents caused by driver fatigue and is a step forward in integrating AI into intelligent transportation systems.[99]

Defence

Department of Defence Unveils Responsible AI Toolkit to Enhance Ethical AI Development

The United States Department of Defence (DoD) introduced the Responsible AI (RAI) Toolkit, a pivotal resource aimed at fostering ethical AI development across federal agencies. Developed by the DoD Chief Digital and AI Office, the toolkit consolidates guidelines and resources to help developers identify useful data sets, flag biases, and demonstrate responsible AI capabilities to leadership. It builds upon existing frameworks such as the Defence Innovation Unit's Responsible AI Guidelines, the National Institute of Standards and Technology's AI Risk Management Framework, and the IEEE 7000-2021 Standard Model Process for addressing ethical concerns during system design. The RAI Toolkit is designed to operationalize DoD's AI Ethical Principles, providing structured guidance for designing, developing, deploying, and using AI systems responsibly. This initiative aims to tackle issues like bias, transparency, and security flaws in AI systems, thereby enhancing trust and compliance with ethical and legal standards. [100]

Retail

Walmart Embraces the Future of Retail: Preparing for the Rise of AI-Powered Shopping Agents

Walmart is gearing up for a transformative shift in retail by preparing to accommodate AI shopping agents—automated digital assistants that shop on behalf of consumers. As these AI agents become more prevalent, capable of comparing prices, checking availability, and even making purchases, Walmart is strategizing how to adapt its digital storefronts to cater to them. This includes optimizing product listings, ensuring accurate inventory data, and refining pricing algorithms to remain competitive in a landscape where machines, not just humans, are making purchasing decisions. The move signals a broader trend in retail where companies must now consider not only human shoppers but also the AI systems acting on their behalf.[101]

Agriculture

AI-IoT Enabled Smart Agriculture Pivot: A Deep Learning Approach for Real-Time Plant Disease Detection and Treatment

In a recent study published in Scientific Reports, researchers introduced an AI-IoT-based smart agriculture pivot system designed to detect and treat plant diseases with high precision. The system integrates deep learning models, particularly a pre-trained ResNet50, into a central pivot irrigation structure, enabling real-time image-based disease classification across multiple crop types. By augmenting a dataset of 25,940 plant leaf images, the model achieved an impressive testing accuracy of 99.8%, with F1-score, recall, and precision metrics nearing perfection. This innovative approach addresses critical agricultural challenges such as water scarcity, pest infestations, and disease outbreaks, offering a scalable and efficient alternative to traditional drone and robotic solutions. The proposed system not only enhances crop health monitoring but also contributes to sustainable farming practices by automating disease management through intelligent actuator control.[102]

[99] https://www.nature.com/articles/s41598-025-01561-7
[100] https://fedtechmagazine.com/article/2025/04/dod-responsible-ai-rai-toolkit-perfcon
[101] https://www.wsj.com/articles/walmart-is-preparing-to-welcome-its-next-customer-the-ai-shopping-agent-6659ef18?st=weAmL9
[102] https://www.nature.com/articles/s41598-025-98454-6

External Document © 2025 Infosys Limited

Infosys Developments

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

Events

2025 Annual Affiliates Meeting at Stanford University | April 21-23 | Stanford

On April 21-23, 2025, the Annual Affiliates Meeting at Stanford University brought together leading minds in AI and data science. The event featured prominent speakers, including **Carlos Guestrin**, Professor of Computer Science at Stanford University, and **Sanmi Koyejo**, Assistant Professor of Computer Science at Stanford University and Co-founder of Virtue AI. Particularly, **Prof. Sanmi Koyejo's** talk on "**Navigating AI Safety and Security at the Frontier**" underscored the critical importance of Responsible AI in ensuring ethical, fair, and secure AI applications. **Mandanna Appanderanda, Head of Infosys Responsible AI - US**, participated alongside Infosys peers, emphasizing the potential of large language models to amplify human capacity and drive innovation across industries. Discussions also focused on the open-sourced Infosys Responsible AI Toolkit, which bridges academia and industry in addressing AI risks. The concept of Responsible AI was underscored as essential for building trust, mitigating risks, and driving positive societal impact. As AI continues to evolve, prioritizing responsible practices remains key to harnessing its full potential.

Global Open-Source Innovation Meetup - GOSIM AI 2025| May 05 | Paris

On May 05, 2025, at the **GOSIM AI 2025** meetup, **Sray Agarwal** from Infosys Responsible AI Office had the privilege of serving as a panelist on a compelling discussion focused on **Responsible AI**. The session was rich with insights, highlighting the collective urgency around building ethical, transparent, and accountable AI systems. As part of the panel, he emphasized the critical role of **open-source Responsible AI (RAI)** toolkits in shaping a more trustworthy AI ecosystem. These toolkits not only enable transparency and community-driven improvements but also empower developers and organizations to proactively assess and mitigate risks. In an era where AI decisions have far-reaching consequences, he underscored the importance of moving beyond reactive compliance to a mindset of proactive responsibility—embedding ethical considerations into the very foundation of AI development. The dynamic exchange of ideas at the panel reaffirmed the shared commitment to ensuring AI technologies serve society in equitable and meaningful ways.

Workshop on AI Governance Alliance Safe Systems | April 29 | San Francisco

On April 29, 2025, The World Economic Forum (WEF) hosted **"AI Governance Alliance Safe Systems Workshop"** at San Francisco , California. It was the first in-person workshop on AI Agents Safety. **Mandanna Appanderanda** and **Kaushal Rathi** from Infosys Responsible AI Office have attended this workshop. In this session, AI safety across 3 dimensions were discussed heavily - policy makers, adopters, and developers. Infosys was able to make a huge impact by providing consultation across these dimensions, resulting in invitations to further events and more rounds of discussions.

UNESCO's 4th Stakeholder Consultation for the AI Readiness Assessment | May 09 | Guwahati

Ashish Tewari, Head of Responsible AI Office – India, participated in the 4th Stakeholder Consultation for the AI Readiness Assessment Methodology (RAM) in India, organized by UNESCO and supported by the Ministry of Electronics and Information Technology (MeitY), in Guwahati on 9th May 2025. This multi-stakeholder platform aimed to assess and advance India's preparedness for ethical and responsible AI deployment using **UNESCO's AI Readiness Assessment Methodology**, a global tool evaluating over 200 indicators across legal, technical, economic, social, and educational domains.

The consultation featured the presence of esteemed dignitaries, including **Mr. Gopinath Narayan**, Principal Secretary IT, Government of Assam; **Dr. Ravi Kota**, Chief Secretary, Government of Assam; **Shri Abhishek Singh**, CEO of INDIAai Mission, Director General of the National Informatics Centre, and Additional Secretary at MeitY; and **Dr. Mariagrazia Squicciarini**, Chief of the Executive Office, Social and Human Sciences, UNESCO. **Ashish Tewari** contributed to the breakout sessions, offering inputs on strengthening responsible AI adoption in India through collaborative and policy-aligned approaches.

External Document © 2025 Infosys Limited

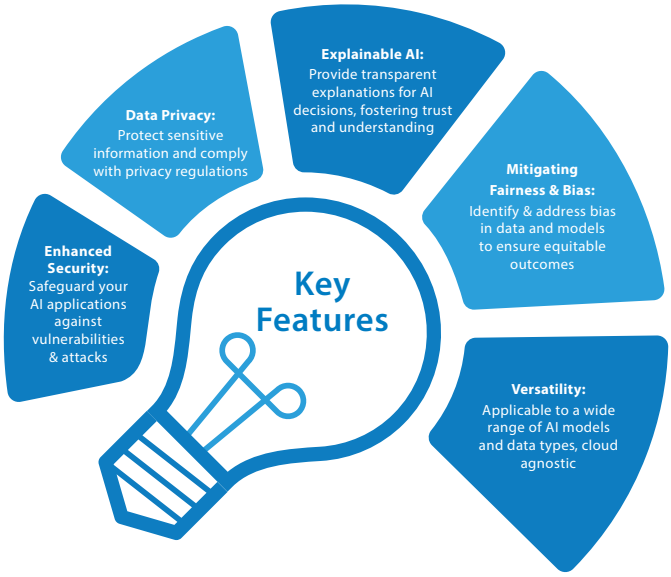
Infosys Responsible AI Toolkit – A Foundation for Ethical AI

The open-sourced Infosys Responsible AI Toolkit can be accessed from its public GitHub repo.[103]

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. The main components include:

Show your support by giving a star to the Infosys Responsible AI Toolkit repository in GitHub and be a part of the Responsible AI Revolution!



New Features Added

Watch out for new features in our next release version 2.2 which will be out soon!

- Red Teaming: Simulating Adversarial Attacks to identify & mitigate vulnerabilities in AI Model
- Fairness Auditing for continuous monitoring & mitigation of biases
- Image Analysis & Evaluation Metrics for Image Explainability Module
- Object Detection Explanation of Explainability module
- New checks added in moderation layer for Ban Code, Sentiment, Gibberish, and invisible text
- Multimodal Enhancement: Information Retrieval from PDFs containing images for Hallucination module
- Multi-document type support for PII data masking of Privacy module
- Simplified Moderation Response for Chatbot's split-screen user interface
- Logic of Thought (LoT) for improved LLM Reasoning: LLM-Explain module
- LLM-Explain: Customization to configure any LLM endpoint to get explanation
- Bulk processing of multiple records for LLM-Explain

[103] https://github.com/Infosys/Infosys-Responsible-AI-Toolkit

Contributors

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

Ashish Tewari - Head of Infosys Responsible AI Office, India

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.

External Document © 2025 Infosys Limited

SAFEGUARD YOUR ALGORITHMS WITH INFOSYS RESPONSIBLE AI

External Document © 2025 Infosys Limited

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

© 2025 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and/ or any named intellectual property rights holders under this document.